ARO

The Scientific Journal of Koya University

KOYA UNIVERSITY

ARO

The Scientific Journal of Koya University

Vol X, No 2(2022)

Contents

ARO Editorial Words

Dear esteemed readers,

We are delighted to present the 19th issue of ARO, the esteemed Scientific Journal of Koya University (KOU), marking another exciting and dynamic season. As an internationally renowned journal, Aro continues to make its mark in the Kurdistan Region of Iraq (KRI). It is a great achievement that Aro has been accepted for indexing in the Emerging Sources Citation Index (ESCI) since February 2016, and we eagerly anticipate the new WoS Impact Report for all ESCI listings in June 2023. This index is under consideration by WoS (Clarivate Analytics) for potential inclusion in the prestigious Science Citation Index Expanded™ (SCIE) following the upcoming evaluation. The listing of ARO's individual articles using unique DOI numbers by WoS (Clarivate Analytics) is a historic milestone for our academic community.

As we embark on our ninth year of leading regional scientific publications with global impact, the dedicated editorial team has worked tirelessly to uphold our mission and ensure that ARO's future publications make an even greater impact and receive more citations. We are proud to have been awarded the DOAJ Seal listing, which signifies our commitment to trusted high-quality open access scientific work. Excitingly, ARO's journey will be further enhanced as WoS (Clarivate Analytics) evaluates our journal for full permanent listing.

ARO's primary mission remains providing valuable resources, support, and guidance to researchers in their scientific paper publication journey, while also ensuring free public access to scientific research through online open access. Although a challenging task, we strive to accomplish this in the years to come. In the following sections, we will delve into the core elements that define ARO's essence. While we have successfully established a committed and well-organized editorial board, attracting high-quality research and scholarly work remains a significant challenge. To address this, KOU is launching the Adjunct Scholar Programs, enabling non-resident scholars to affiliate with KOU and publish in ARO.

Aro was founded with the long-term vision of making science accessible to researchers in Kurdistan and beyond, encompassing diverse scholarly disciplines. It is an open access, peer-reviewed journal that publishes original scientific research, global news, letters and commentary, as well as review articles in the natural sciences and technology. This issue features original research papers spanning various fields.

The overwhelming response from researchers, academics, and professionals over the past nine years has led us to expand our Editorial Board to accommodate the increasing number of submitted scientific manuscripts. However, having a dedicated and well-organized board is just one aspect of the equation. Attracting high-quality research submissions is equally vital. We extend our gratitude to all those who entrusted Aro with their original research work for publication in Vol. X, No. 2 (2022) of the journal, as well as the 42 peer-reviewers from universities worldwide who diligently reviewed and enabled this issue of ARO.

We value and welcome your support and feedback as we continue our mission to advance scientific excellence and knowledge dissemination.

Your support and feedback are invited and appreciated.

Dilan M. Rostam
Editor-in-Chief

Mohammed H. Zangana
Executive Publisher

Salah I. Yahya, Basim M. Fadhil, Fahmi F. Muhammad, Wali M. Hamad, Jorge Correia, Fouad Mohammed, Jacek Binda, Nadhir Al-Ansari, Howri Mansurbeg, Tara F. Tahir and Yazen A. Khaleel.
Executive Editorial Board

iv

A New Design Approach for a Compact Microstrip Diplexer with Good Passband Characteristics

Abbas Rezaei[1] and Salah I. Yahya[2,3]

[1]Department of Electrical Engineering, Kermanshah University of Technology,
Kermanshah, Iran

[2]Department of Communication and Computer Engineering, Cihan University-Erbil,
Erbil, Kurdistan region – F.R. Iraq

[3]Department of Software Engineering, Faculty of Engineering, Koya University,
Koya KOY45, Kurdistan region – F.R. Iraq

Abstract—**This paper presents an efficient theoretical design approach of a very compact microstrip diplexer for modern wireless communication system applications. The proposed basic resonator is made of coupled lines, simple transmission line and a shunt stub. The coupled lines and transmission line make a U-shape resonator while the shunt stub is loaded inside the U-shape cell to save the size significantly, where the overall size of the presented diplexer is only 0.008 λg^2. The configuration of this resonator is analyzed to increase intuitive understanding of the structure and easier optimization. The first and second resonance frequencies are f_{o1} = 895 MHz and f_{o2} = 2.2 GHz. Both channels have good properties so that the best simulated insertion loss at the first channel (0.075 dB) and the best simulated common port return losses at both channels (40.3 dB and 31.77 dB) are achieved. The presented diplexer can suppress the harmonics acceptably up to 3 GHz (3.3 f_{o1}). Another feature is having 31% fractional bandwidth at the first channel.**

Index Terms—**Compact, Diplexer, Insertion loss, Microstrip, Resonator, Return loss**.

I. Introduction

Recently, microstrip devices such as diplexers have been used widely in modern wireless communication systems. Two-channel bandpass-bandpass diplexers can transmit signals through two passbands and eliminate undesired harmonics. Each channel is used for receiving or sending signals from an antenna (Majdi and Mezaal, 2022). Several types of microstrip diplexer are introduced in (Hussein, Mezaal and Alameri, 2021; Chen, et al., 2021; Yahya, Rezaei and Nouri, 2020; Lu, et al., 2020; Su, et al., 2020; Tahmasbi, Razaghian and Roshani, 2021; Rezaei, et al., 2019; Shirkhar and Roshani, 2021; Zhanga,

Zhu and Li, 2018; Rezaei, Yahya, Noori and Jamaluddin, 2019; Dembele, et al., 2019; Yousif and Ezzulddin, 2020; Fernandez-Prieto, et al., 2018; Guan, et al., 2019; Guan, et al., 2014; Noori and Rezaei, 2017). However, all of them occupy large area. Two bandpass filters (BPFs) consisting spiral cells and coupled lines are integrated for obtaining a microstrip diplexer in (Hussein, Mezaal and Alameri, 2021). It has some disadvantages such as undesired harmonics and high losses at both channels. Three coupled lines structures are employed in the layout configuration of presented microstrip diplexer in (Chen, et al., 2021). It could better suppress harmonics but the problem of high losses at both channels is remained. The proposed diplexer in (Yahya, Rezaei and Nouri, 2020) could improve the simulated insertion losses at both channels while it could suppress the harmonics very well. However, it has two high common port return losses at both channels. In (Lu, et al., 2020), a microstrip diplexer with multiple transmission zeroes (TZs) has been designed to operate at 0.755 GHz and 1.056 GHz which is suitable for the Global System for Mobile Communications (GSM). The introduced diplexer in (Su et al., 2020) has large losses at both channels, but in (Tahmasbi, Razaghian and Roshani, 2021) the simulated insertion losses at both channels are very low. However, the proposed structure in (Tahmasbi, Razaghian and Roshani, 2021) could not obtain low return losses at its upper and lower channels. In (Rezaei, et al., 2019), sharpness is not good while it could not attenuate the harmonics. In (Shirkhar and Roshani, 2021), coupled meandrous structure has been employed to improve simulated insertion losses. However, it has undesired harmonics. A balanced-to-balanced microstrip diplexer using a large dual-mode resonator with undesired harmonics and high losses has been presented in (Zhanga, Zhu and Li, 2018). Two similar BPFs with flat channels are integrated in (Dembele, et al., 2019) for obtaining a microstrip bandpass-bandpass diplexer. In (Yousif and Ezzulddin, 2020), the microstrip meandrous closed loops have been utilized to design of a diplexer with good measured return losses. The problem of high losses has been remained in the presented diplexers in

(Fernandez-Prieto, et al., 2018; Guan, et al., 2019; Guan, et al., 2014; Noori and Rezaei, 2017).

While efforts are being made to design a compact microstrip bandpass-bandpass diplexer with good performance, it is hard to compromise the compact size on account of the design features, e.g., low losses, higher harmonics suppression and higher fractional bandwidth. The approach presented here is an efficient one which first, presents and analyzes a resonator. Second, this resonator is used to design two BPFs. Third, the BPFs are optimized so as to be integrated to achieve an efficient diplexer design. Finally, the proposed diplexer design is compared with the previous reported designs in terms of losses, size and harmonics, to prove the efficiency and better performance of the presented compact design. It has the best passbands totally.

II. Resonator Structure

Coupled lines are important parts of bandpass resonators. Because they create many capacitors between the lines, where both lines show inductance features. However, if we use only coupled lines, the dimensions will be very large, while we will not get a good frequency response. Hence, usually a combination of the other resonators is used alongside the coupled lines. Accordingly, we propose the general structure shown in Fig. 1 to design the BPFs. As depicted in Fig. 1, it consists of coupled lines connected to a simple line loaded by a shunt internal resonator. To save the size, we placed the simple transmission line in such a way that it forms a U-shaped structure next to the coupled lines. The stub is loaded inside the U-shape structure to save the overall size again.

The impedance of the coupled lines and internal shunt stub are assumed to be Z_C and Z_S, respectively. As presented in Fig. 1, the transmission line is divided into two parts with the impedances Z_1 and Z_2. The input impedance that displayed from the input port is:

$$Z_{in} = \frac{Z_S Z_1}{Z_S + Z_1} + Z_C + Z_2 = \frac{Z_S Z_1 + Z_S Z_C + Z_S Z_2 + Z_1 Z_2 + Z_C Z_1}{Z_S + Z_1}$$

(1)

One way to create a resonance is to have $Z_S = -Z_1$ According to the proposed resonator structure, Z_1 is an impedance of an inductor. Therefore, the shunt stub must be capacitive. The fill square or rectangular cells are good choices for the internal stub. A larger internal stub shifts the resonance frequency to the lower frequencies. Another way to get the resonance frequency is:

$$Z_S(Z_1 + Z_C + Z_2) + Z_1(Z_2 + Z_C) = 0 \quad (2)$$

A method to solve Eq. (2) is by setting reasonable custom values for some parameters to find the other values easily. To calculate Z_1, we can use the following equations (Hong and Lancaster, 2001):

$$Z_1 = jZ_{C1} \tan(\beta l_1) \quad (3.a)$$

where:

$$Z_{C1} = \frac{120\pi}{\sqrt{\varepsilon_{re}}} \{\frac{w}{h} + 1.393 + 0.677\ln(\frac{w}{h} + 1.444)\}^{-1} \quad (3.b)$$
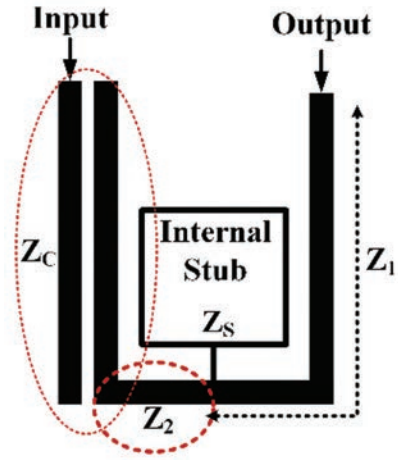
Fig. 1. Proposed resonator consisting of coupled lines connected to a simple line loaded by a shunt internal resonator.

$$\beta = \frac{\pi f_O(GHz)\sqrt{\varepsilon_{re}}}{150} \quad (3.c)$$

In Eq. (3), Z_{C1} and β are the characteristic impedance and propagation constant. If we assume that the length of the transmission line with the impedance Z_1 is $l_1 = 20$ mm with a width of $W = 0.6$ mm on a substrate with the effective dielectric constant of $\varepsilon_{re} = 1.75$ ($\varepsilon_r = 2.22$), Zc_1 become near 105.97Ω. On the other hand, at a target operational frequency of $f_o = 2$ GHz, β become near 17.63π. Finally, $Z_1 = 211.94j$ Ω which is the equivalent of a 16.86 nH inductance. If the internal shunt stub is a large square fill cell with 10 mm × 10 mm dimension, the characteristic impedance of internal stub will be $Z_{CS} = 17.9$ Ω while $\beta = 17.63\pi$. In this case, Z_S must be calculated from:

$$Z_S = \frac{Z_{CS}}{j\tan(0.01\beta)} \quad (3.d)$$

So $Z_S = -j28.9$ Ω, which is equivalent of a capacitor. If we increase the size of internal stub to 20 mm × 20 mm, then $Z_S = -14j\Omega$. Therefore, by increasing the dimension of the shunt stub it will be equal to a large capacitor which can shift the resonance frequency to the left. We can set $Z_2 = 0.5$ and $Z_1 = 105.97j\Omega$ to find the impedance of coupled lines easily. This was a method to solve Eq. (2). We can optimize the dimensions to obtain more compact size at our desired resonance frequency. Using this analyzed resonator, two BPFs can be designed as it is explained in the next section.

III. Design of BPFs

Based on the proposed resonator, two BPFs are designed for wireless applications. The first filter (BPF1) is shown in Fig. 2 which operates at the higher frequency. This filter is composed of coupled lines, simple transmission line and an internal shut stub.

The basic structure of this filter is the same as the analyzed resonator in the previous section. The internal stub is a high impedance section while two step-impedance feed structures are utilized to decrease the losses. The approximated
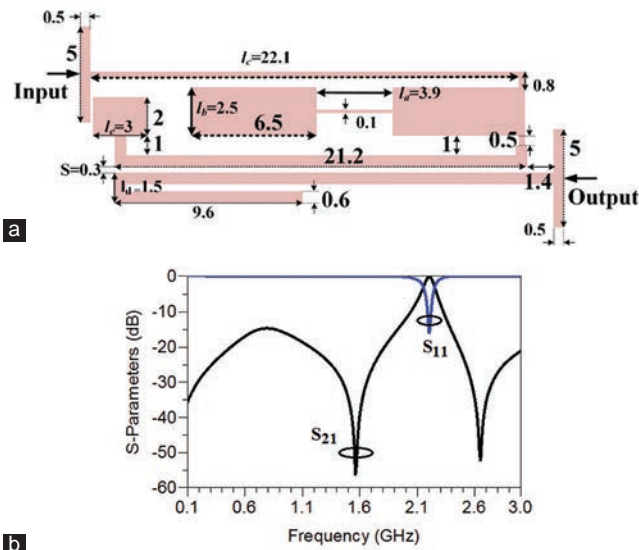
Fig. 2. (a) Layout of BPF1 (all dimensions are in mm), (b) Frequency response of BPF1.

dimensions are obtained according to the method explained in the Resonator Structure section. Then, we optimized the dimensions to get a better frequency response. Tapped line feeds are used to control the insertion loss and return loss, where the loss control is much easier to adjust.

It is simulated on a RT/Duroid 5880 substrate with 2.22 dielectric constant, h = 0.7874 mm and 0.0009 loss tangent. It is simulated by ADS software.

The resonance frequency of BPF1 is located at 2.2 GHz, with 0.23 dB insertion loss while the return loss is 16.1 dB. This filter creates two TZs at 1.56 GHz and 2.65 GHz which improve the selectivity. The harmonic level is lower than 14.6 dB above and after the passband up to 3 GHz. The dimensions of BPF1 are optimized as presented in Fig. 3a-f. Fig. 3a shows S_{21} and S_{11} as a function of the physical length l_a. By increasing this length, the operational frequency shifts to the lower frequency. As depicted in Fig. 3a, increasing the length l_a will also lead to improve selectivity. However, better return loss will be obtained by decreasing l_a. Fig. 3b
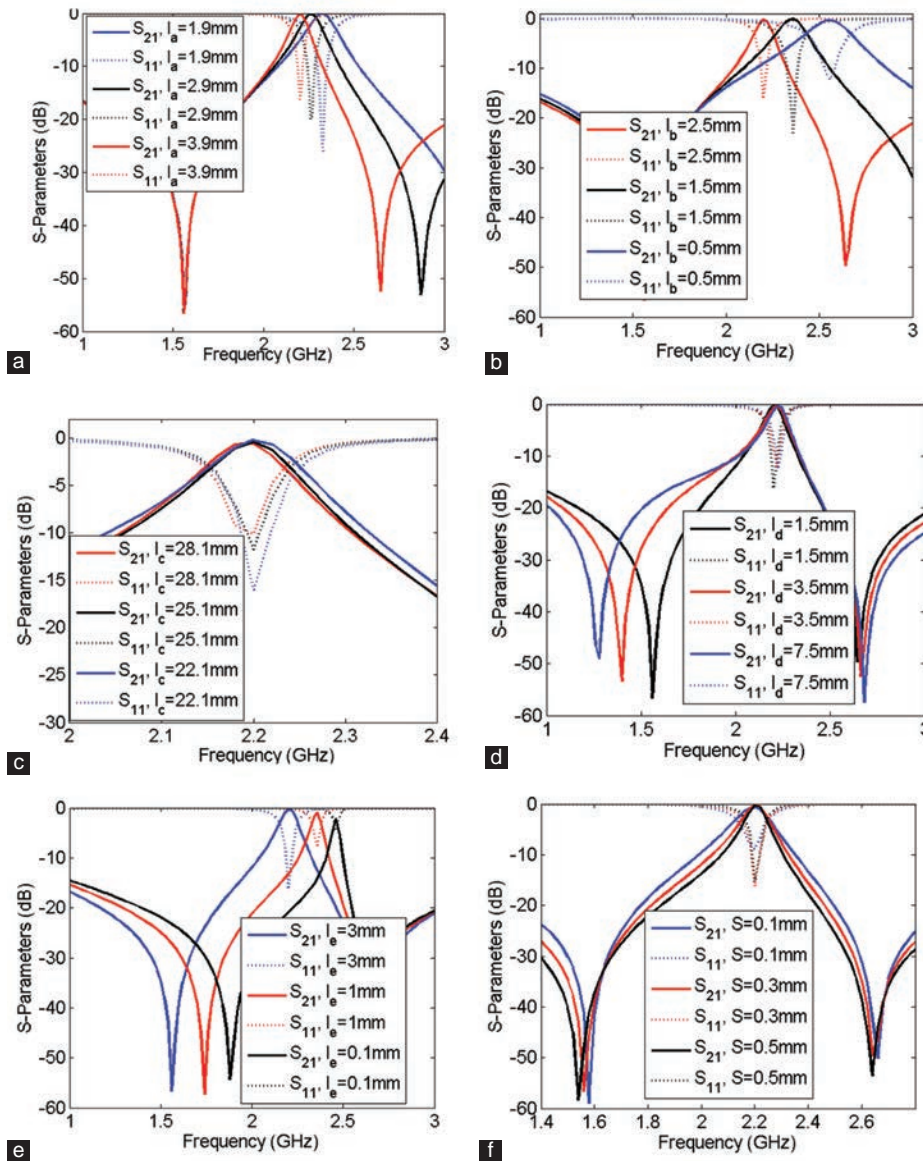
Fig. 3. Frequency response as a function of (a) l_a, (b) l_b, (c) l_c, (d) l_d, (e) l_e and (f) S.

illustrates the frequency response of BPF1 as a function of length l_b. Similar to the effect of l_a, increasing the size of l_b shifts the resonant frequency to the left. On the other hand, decreasing l_b will increase the bandwidth. Therefore, it is wise to choose an intermediate value of l_b. As shown in Fig. 3c, we can decrease the physical length l_b to improve the return loss. However, if we reduce it too much, the frequency response will be broken or at least the resonance frequency will shift to the right. As presented in Fig. 3d, to improve the sharpness of the frequency response, the length l_d can be reduced. Decreasing the physical length l_e will corrupt the frequency response. This is presented in Fig. 3e. As shown in Fig. 3e, by increasing l_e the losses are decreased significantly. The distance between the coupled lines has a significant effect on the return loss. This is shown in Fig. 3f, where by reducing S, the return loss will be increased.

From the above optimization method, it can be concluded that increasing the dimensions will lead to a resonance frequency shifted to the left. However, we must be careful not to destroy the other features by increasing the dimensions. Using the above method, another BPF named as BPF2 is designed to operate at lower frequencies. The layout configuration of BPF2 and its corresponding frequency response are depicted in Fig. 4a and b. As shown in Fig. 4a, the coupled lines and shunt stub are used similar to our proposed resonator. Here, the shunt stub is a rectangular cell.
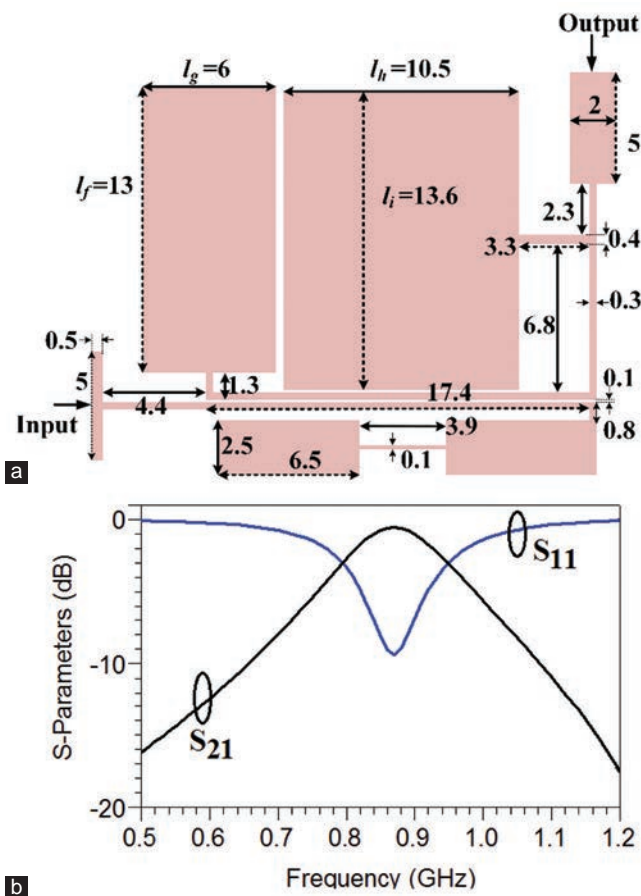
Similar to the proposed resonator, BPF2 is composed of a pair of coupled lines and a shunt stub. Another rectangular stub is connected to the upper line to control the bandwidth. The main shunt stub with physical length l_i is important to form the bandwidth. Similar to the BPF1, tapped line feeds are used to control the insertion and return losses easily. The approximated dimensions can calculate according to the explained method in the resonator design.

To optimize the frequency response, the other stubs are added. Fig. 4b depicts that BPF2 operates at 870 MHz which is suitable for GSM applications.

The passband is from 790 MHz to 950 MHz with 0.57 dB insertion loss. The effect of changing BPFs parameters is presented in Fig. 5a-d. Fig. 5a and b show that two factors can eliminate passband. One of them is decreasing the physical length l_f and the other one is decreasing the length l_g. As depicted in Fig. 5a, choosing a neither low nor high value of l_f is suitable. The length l_g has a similar effect so that we must choose a middle value of it. Increasing the physical lengths l_h and l_i can improve the frequency selectivity. The effect of these lengths on the frequency response is presented in Fig. 5c and d. Moreover, decreasing these lengths increases the losses. Therefore, we can enlarge l_h and l_i insofar as the overall size remains compact. By connecting both BPF1 and BPF2, a microstrip diplexer is designed, as shown in the next section.

IV. Diplexer Structure and Its Simulation Results

A microstrip diplexer usually is composed of two BPFs. In some cases, a matching network is used to obtain a good performance. However, we connected our both filters without any extra matching network that leads to save the size. Our proposed diplexer is presented in Fig. 6.

This diplexer is made by direct connection of BPF1 and BPF2. As shown in Fig. 6, the BPF1 and BPF2 are connected to Port3 and Port2, respectively. According

Fig. 4. (a) Layout configuration of BPF2 where all dimensions are in mm, (b) frequency response of BPF2.

Fig. 5. Frequency response of BPF2 as functions of (a) l_f, (b) l_g, (c) l_h, (d) l_i.

to the optimization method in the previous section, the dimensions are selected so that we have the best frequency response. The high impedance part is repeated in both filters. However, it is seen once in the proposed diplexer structure which leads to save size significantly. The overall size of our designed diplexer is 22.9 mm × 25 mm ($0.09\lambda_g \times 0.098\lambda_g$) which is very compact, where λ_g is the guided wavelength on the substrate calculated at the lower resonance frequency. The overall size according to the second resonance frequency is $0.22\lambda_{g-s} \times 0.24 \lambda_{g-s}$, where λ_{g-s} is the guided wavelength on the substrate calculated at the upper resonance frequency.

The frequency response of the proposed diplexer is depicted in Fig. 7a and b. Both filters and the proposed diplexer are simulated on a RT/Duroid 5880 substrate with $\varepsilon_r = 2.22$, h = 0.7874 mm and 0.0009 loss tangent.

The simulation results are extracted by Agilent Advanced Design System (ADS) software using the EM simulator. As illustrated in Fig. 7a, the first passband is from 735 MHz to 1005 MHz. The fractional bandwidth (FBW) at this channel is 31%. Therefore, this channel is wide which is an advantage. The resonance frequency of this channel is $f_{o1} = 895$ MHz,

which is suitable for GSM applications. The losses in this channel are very low. The simulated insertion and return losses are 0.075 dB and 40.3 dB respectively. Fig. 7a shows that there is an 80.2 dB TZ at 1.494 GHz which improves the sharpness of the first channel. Similar to the lower channel, our upper channel has low losses with 0.14 dB insertion loss and 31.7 dB return loss. The upper operational frequency is $f_{o1} = 2.2$ GHz. This passband is from 2.14 GHz to 2.26 GHz with 5.45% FBW. S_{31} has two TZs of 60.6 dB and 55.954 dB at 1.564 GHz and 2.683 GHz respectively. As shown in Fig. 7, the harmonics are suppressed up to 3 GHz where it has the maximum harmonic level 14.6 dB. Fig. 7b depicts the isolation between channels which is better than 20.71 dB from DC to 3 GHz.

V. Comparison and Discussion

To show the advantages of this work, we compared it with the previous diplexers. The comparison results are listed in Table I, where IL1, IL2, RL1, and RL2 are insertion loss at the first channel, insertion loss at the second channel, return loss at the first channel and return loss at the second channel, respectively. As shown in Table I, the lowest insertion loss at the first channel and the lowest common port return loss at both channels are the achievements of this work. Another significant advantage of our diplexer is to have the most compact size in comparison with the previous diplexers. As illustrated in Table I, we could get an acceptable harmonic attenuation. Only the introduced diplexer in (Yahya, Rezaei and Nouri, 2020) could attenuate better than us. However, it has a very large size of 0.037 λg^2 with high common port return loss at its both channels.

In the comparison table, the data of these works: (Hussein, Mezaal and Alameri, 2021), (Yahya, Rezaei and Nouri, 2020), (Tahmasbi, Razaghian and Roshani, 2021), (Rezaei, et al., 2019), (Shirkhar and Roshani, 2021), (Rezaei, et al., 2019), (Fernandez-Prieto, et al., 2018), and (Noori and Rezaei, 2017) are based on simulation results and the others are based on the measured results. As shown in Table I, our diplexer has low loss, compact size, good harmonic attenuation, and reasonable isolation.

Fig. 6. Layout of the proposed diplexer where all dimensions are in mm.

Fig. 7. Frequency response of the proposed diplexer (a) transfer parameters and common port return loss, (b) isolation between channels.

TABLE I
COMPARISON BETWEEN THIS WORK AND THE PREVIOUSLY REPORTED DIPLEXERS

References	Common port insertion losses: IL_1, IL_2 (dB)	Common port return losses: RL_1, RL_2 (dB)	f_{o1}, f_{o2} (GHz)	Size (λg^2)	Last suppressed harmonic	Isolation (dB)
Our Diplexer	0.075, 0.14	40.3, 31.77	0.895, 2.2	0.008	$3.3 f_{o1}$	20.7
(Hussein, Mezaal and Alameri, 2021)	4.2, 3.3	32, 31	1.85, 2.5	0.096	$1.6 f_{o1}$	31
(Chen et al., 2021)	4.4, 4.6	Better than 15	1.41, 2.02	0.098	$3 f_{o1}$	29
(Yahya, Rezaei and Nouri, 2020)	0.12, 0.18	21, 17	3.5, 5	0.037	$12 f_{o1}$	20
(Lu, et al., 2020)	0.8, 1.1	---	0.755, 1.056	0.040	$2.1 f_{o1}$	24
(Su, et al., 2020)	2.62, 2.34	---	9.55, 10.61	3.56	$2.1 f_{o1}$	45
(Tahmasbi, Razaghian and Roshani, 2021)	0.05, 0.08	26.9, 23.5	1.5, 2.1	0.030	$3.3 f_{o1}$	19
(Rezaei, et al., 2019)	0.36, 0.44	23.7, 23.7	2.88, 3.29	0.028	$1.2 f_{o1}$	23
(Shirkhar and Roshani, 2021)	Better than 0.5	20, 25	1.7, 2.1	0.061	$1.4 f_{o1}$	40
(Zhanga, Zhu and Li, 2018)	1.06, 1.09	Better than 17	2.2, 2.63	0.207	$1.8 f_{o1}$	38
(Rezaei, et al., 2019)	0.10, 0.16	33, 22	1.6, 2.1	0.054	$2.1 f_{o1}$	22
(Dembele, et al., 2019)	1.8, 1.9	10, 10	8.3, 10	0.071	$2.4 f_{o1}$	26.6
(Yousif and Ezzulddin, 2020)	0.5, 2	16, 25	2.6, 6	0.22*	$3 f_{o1}$	23
(Fernandez-Prieto, et al., 2018)	1.15, 1.54	Near 30	2.49, 2.97	0.218	$4 f_{o1}$	30
(Guan, et al., 2019)	1.15, 1.5	Better than 17.7	0.99, 1.46	0.097	$2.5 f_{o1}$	40
(Guan, et al., 2014)	1.2, 1.5	Near 20	1.95, 2.14	0.136	$1.2 f_{o1}$	35
(Noori and Rezaei, 2017)	0.2, 0.4	15, 11	2.36, 4	0.089	$3.3 f_{o1}$	19.8

*Approximate

VI. CONCLUSION

A microstrip diplexer, composed of two BPFs, is introduced in this work. The resonance frequency of the filters and diplexer make them suitable for modern wireless communication networks. First, a resonator, and then, two BPFs were proposed and optimized. Finally, by connecting the BPFs without any extra matching circuit a high performance diplexer was introduced. Both passbands presented good features. The comparison results of the proposed new design approach with the previously reported designs showed that the lowest insertion loss at the lower channel and the lowest return losses at both channels are obtained while our diplexer has the most compact size. Meanwhile, the other parameters such as FBW at the first channel (31%) and harmonic attenuation were good. The isolation and FBW at the upper band were acceptable.

REFERENCES

Chen, C.F., Zhou, K.W., Chen, R.Y., Tseng, H.Y., He, Y.H., Li, W.J. and Weng, J.H., 2021. Design of microstrip multifunction integrated diplexers with frequency division, frequency selection, and power division functions. *IEEE Access*, 9, pp.53232-53242.

Dembele, S.N., Bao, J., Zhang, T. and Bukuru, D., 2019. Compact microstrip diplexer based on dual closed loop stepped impedance resonator. *Progress in Electromagnetics Research C*, 89, pp.233-241.

Fernandez-Prieto, A., Lujambio, A., Martel, J., Medina, F., Martin, F. and Boix, R.R., 2018. Balanced-to-balanced microstrip diplexer based on magnetically coupled resonators. *IEEE Access*, 6, pp.18536-18547.

Guan, X., Liu, W., Ren, B., Liu, H. and Wen, P., 2019. A novel design method for high isolated microstrip diplexers without extra matching circuit. IEEE Access, 7, pp.119681-119688.

Guan, X., Member, F., Yang, H., Liu, H. and Zhu, L., 2014. Compact and high-isolation diplexer using dual-mode stub-loaded resonators. *IEEE Microwave and Wireless Components Letters*, 24(6), pp.385-387.

Hong, J.S. and Lancaster, M.J., 2001. *Microstrip Filters for Rf/Microwave Applications*. John Wiley and Sons, Hoboken.

Hussein, H.A., Mezaal, Y.S. and Alameri, B.M., 2021. Miniaturized microstrip diplexer based on fr4 substrate for wireless communications. *Elektronika Ir Elektrotechnika*, 27(5), pp.34-40.

Lu, Q.Y., Zhang, Y.J., Cai, J., Qin, W. and Chen, J.X., 2020. Microstrip tunable diplexer with separately-designable channels. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(12), pp.2983-2987.

Majdi, K.A. and Mezaal, Y.S., 2022. Microstrip diplexer for recent wireless communities. *Periodicals of Engineering and Natural Sciences*, 10(1), pp.387-396.

Noori, L., Rezaei, A., 2017. Design of a microstrip dual-frequency diplexer using microstrip cells analysis and coupled lines components. International Journal of Microwave and Wireless Technologies, 9(7), pp.1467-1471.

Rezaei, A., Noori, L. and Mohammadi, H., 2019. Design of a miniaturized microstrip diplexer using coupled lines and spiral structures for wireless and WiMAX applications. Analog Integrated Circuits and Signal Processing, 98, pp.409-415.

Rezaei, A., Yahya, S.I., Noori, L. and Jamaluddin, M.H., 2019. Design of a novel wideband microstrip diplexer using artificial neural network. *Analog Integrated Circuits and Signal Processing*, 101(1), pp.57-66.

Shirkhar, M.M. and Roshani, S., 2021. Design and implementation of a Bandpass-bandpass diplexer using coupled structures. *Wireless Personal Communications*, 122(3), pp.2463-2477.

Su, Z.L., Xu, B.W., Zheng, S.Y., Liu, H.W. and Long, Y.L., 2020. High-isolation and Wide-stopband SIW diplexer using mixed electric and magnetic coupling. IEEE Transactions on Circuits and Systems II: Express Briefs, 67(1), pp.32-36.

Tahmasbi, M., Razaghian, F. and Roshani, S., 2021. Design of bandpass-bandpass diplexers using rectangular-, T-, and Lshaped resonators for hybrid power amplifier and 5G applications. *Analog Integrated Circuits and Signal Processing*, 109(3), pp.585-597.

Yahya, S.I., Rezaei, A. and Nouri, L., 2020. Compact wide stopband microstrip diplexer with flat channels for WiMAX and wireless applications. IET Circuits, Devices *and* Systems, 14(6), pp.846-852.

Yousif, A.B. and Ezzulddin, A.S., 2020. A Dual-band coupled line based microstrip diplexer for wireless applications. *Journal of Global Scientific Research*, 10, pp.845-853.

Zhanga, C., Zhu, L. and Li, Y., 2018. Compact microstrip balanced-to-balanced diplexer using stub-loaded dual-mode resonators. *IEICE Electronics Express*, 15(5), pp.1-6.

In Silico Domain Structural Model Analysis of Coronavirus ORF1ab Polyprotein

Mohammed I. Jameel, Rabar J. Noori and Soma F. Rasul

Department of Medical Microbiology, Faculty of Science and Health, Koya University,
Koya KOY45, Kurdistan Region – F.R. Iraq

Abstract—The world today is battling with a coronavirus infection that is considered a global pandemic. Coronavirus infection is mainly attribute to the varying technique of the replication and release of different genomic components of the virus. The present study aims to establish the physical and chemical features, as well as the basic structural and functional properties of *Coronavirus ORF1ab* domain. A molecular approach was adopt in this study using the Swiss Model and Phyre2 server whereas the prediction of the active ligand binding sites was done using Phyre2. The analysis of the structure of the protein showed that it has good structural and heat stability, as well as better hydrophilic features and acidic in nature. Based on the Homology modeling, only two binding active sites were noted with catalytic function being mediated by Zn^{2+} as the metallic heterogeneous ligand for binding sites prediction. The proteins mostly exhibited helical secondary configurations. This study can help in predicting and understanding the role of this domain protein in active coronavirus infection.

Index Terms—Coronavirus, Ligand, Modeling, ORF1ab.

I. Introduction

Coronavirus disease 2019 (COVID-19) is a public health problem facing the world today; it is an infectious disease caused by the coronavirus family which has wreaked havoc to both human lives and the global economy since its emergence in December 2019. It is a global pandemic that has infected over 50 million and killed nearly 1,300,000 persons (World Health Organization, 2020). The disease is more severe among elderly people who have accounted for the largest population of the affected cases (Dowd, et al., 2020; Wu, 2020; Onder, 2020). Coronavirus 2019 has been confirmed in more than 200 countries globally and new cases are being reported in other regions. With the pattern of its transmission, it is believed that a vaccine is urgently needed to develop strong immunity that is needed to prevent a relapse of the disease in the future (Kisslear, et al., 2020).

Statement deleted. Coronavirus (due to COVID-19 occur in 2019 [Schaecher and Pekosz, 2010; McBride and Fielding, 2012]) is an RNA virus (positive stranded) with structural proteins that include envelope protein (E), spike protein (S), membrane protein (M), and nucleocapsid phosphoprotein (N). In addition, the structural genes encode nine accessory proteins, encoded by ORF3a, ORF3d, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF14, and ORF10 genes (Nelson, et al.). Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic (Schaecher and Pekosz, 2010; McBride and Fielding, 2012). The genomic attributes of the novel coronavirus are believed to contribute to the ongoing pandemic (Wu, et al., 2020; Paraskevis, et al., 2020); the envelope protein ion channel (CoV EIC) is responsible for tweaking discharge of virion (Li, et al., 2020). Other severe acute respiratory syndrome SARS 1 and 2 like coronaviruses also have spike proteins and ORF8 and ORF3a proteins but their pathogenicity and pattern of transmission differ from those of SARS-CoV (To, et al., 2020).

Studies have found that the spike proteins of the novel coronavirus facilitate its penetration of epithelial cells through surface interaction with angiotensin-converting enzyme 2 (ACE2), thereby causing human infection. The role of ORF8 and ORF3 in viral infection remains unknown due to the limitations of the existing experimental techniques. However, the mechanism of infection of the severe acute respiratory syndrome-2 is still enigmatic (Rothe, et al., 2020). Following the identification of the structures of the viral proteins of SARS-CoV-2 in early February 2020, many studies have been dedicated to the discovery of novel medicine that may be effective in treating SARS-CoV-2 infection (Muhammed, et al., 2020). The aim of this study to establish the physical and chemical features of coronavirus using different tools.

II. Methods

A. Recovery of the Amino Acid Sequence of the Polyprotein

The polyprotein A.A. sequence was use in the present study was retrieve from the official NCBI sequence databank accessible at https://www.ncbi.nlm.nih.gov. The sequence was sourced with the Accession No: ORF-1; Sequence number: MN908947.3, Gene – ORF1; Polyprotein QHD43415.1; Polyprotein - ORF1; Organism – Pneumonia virus; identifier – Wuhan-Hu-1.

Physiochemical Characterization

Different web servers were used to determine the physiochemical characteristics of the protein sequence. For instance, the structure, aliphatic indices, elimination (coefficient), and instability of the amino acids were predicted using the ProtParam tool (Expasy), coupled with the prediction of the isoelectrical point (pI) and grand average of hydropathicity (GRAVY) (Gasteiger, et al., 2005).

B. The Prediction of Secondary Structure of Protein

The secondary structure of the polyproteins was predicted using the PROFsec, PSIPRED, and SOPMA programming tools (Geourjon, et al., 1995) whereas the DISOPRED tool was used for disorder prediction (Rost, et al., 2004).

C. Prediction of Polyprotein Binding Sites and Gene Ontology

The profISIS server was used or the prediction of the protein-protein binding sites (Ofran and Rost, 2007); this was done through identification of the interacting moieties from the sequence alone through the consolidation of the predicted structural features with the developmental information. The gene ontology (GO) prediction approach was also employed for the prediction of the molecular, cellular, and biological characteristics using the homology to known annotated proteins (Hamp, et al., 2013).

D. Homology Modeling and Polyprotein Validation

Being that the + (3-D) structure in the protein data bank cannot be accessed, this unexplored space domain was modeled using 700 a. a. long-protein sequence. Hence, two programs (Swiss-Model) and (Phyre-2) were used to perform the protein homology modeling (Waterhouse, et al., 2018). The prediction of the secondary protein structure was done using Phyre2 (Kelly, et al., 2015) whereas the Swiss-Model was used to produce the 3D protein model. After the analysis, only the most reasonable (3D) model was selected for the validation step. The Ramachandran plot analysis was used for the approval of the last modeled structure whereas PROCHECK was used for the analysis of the stereochemical property. Finally, the modeled structure was uploaded to the 3D Ligand Site web server for potential binding site prediction (Wass, et al., 2010).

III. Results and Discussion

A. Physiological and Chemical Characterization of the Polyprotein

Coronavirus amino acid sequence was obtained as FASTA file with an instability index (ii) of 27.52 and used as a query sequence for the physiochemical characterization. The protein was acidic in nature (+ve = 76, -ve = 77) whereas the Pi was 7; the MW of the protein was 79,822.08. The protein exhibited a high elimination coefficient value of 102,245 which suggests the existence of +ve Cys residues. However, the aliphatic index value of 77.690 of the search protein implies a good level of stability over a temperature range.

As shown in Table I, the protein exhibited a low GRAVY indices value of –0.204 which suggests a hydrophilic nature that creates a better room for reaction with water (Kyte and Doolottle, 1982).

B. Prediction of Secondary Structure of Polyprotein

The employed default boundaries (window width – 17; similarity threshold – 8; division factor – 3, and number of states – 4) were used for the prediction of the secondary protein structure with the SOPMA tool. With the aid of 511 proteins (used as sub-database) and seven aligned proteins, the following predictions were made using the SOPMA tool: Irregular coils = 35.29% of residues, alpha-helix =38.43% of residues, beta-turn = 9% of residue, and extended strand = 17.29% of residue (Table II). This suggests that the protein has a higher chance of having a helix, strand, and coil, as shown in Fig. 1. The prediction of the secondary protein structure by (PROFsec) (Predict-Protein) using neural organization tool showed the protein to have a prediction precision of >28.71% for helix configuration (α; π; 3_10- helix), 21.43% beta-strand+, and 49.86% loop (L). DISOPRED-based intrinsic disorder profile processing showed that most of the amino acid (>95%) did not meet the recommended +0.5 certainty confidence score for disordered condition which is the least chance for bending, thereby suggesting that the predicted protein is of high strength (Fig. 2).

C. Prediction of Gene Ontology and Binding Sites of Protein

The profISIS software was used for protein binding sites prediction; 14 different binding sites were identified at positions 2; 114–115; 160; 211–212; 249; 316; 417; 482–483;

TABLE I
PHYSIOCHEMICAL PROPERTIES OF THE SARS-CoV-2 POLYPROTEIN USING THE PROTPARAM TOOL

S/n.	Attribute	Value
1	Number of amino acids	700.0
2	Molecular weight	79,822.080
3	Isoelectric point	7.0
4	Instability index	27.520
5	Aliphatic index	77.690
6	Extinction coefficient	102,245.0
7	Total number of –vely charged amino acids	77.0
8	Total no. of +vely charged amino acids	76.0
9	GRAVY (Grand average of hydropathicity)	–0.2040
10	Chemical formula	$C_{3586}H_{5457}N_{947}O_{1038}S_{43}$

TABLE II
SECONDARY PROTEIN STRUCTURE PREDICTION USING SOPMA TOOL

Secondary protein structure element	Value (%)
Alpha helix (Hh)	38.430
310 helix (Gg)	0.000
Pi helix (Ii)	0.000
Beta bridge (Bb)	0.000
Extended strand (Ee)	17.290
β-turn (Tt)	9
Different coil (Cc)	35.290

512; 560; 584; 647; 696; and 700 whereas gene ontology (GO) predicted and classified the utilitarian viewpoints as cellular, molecular, and biological. Catalytic activity includes nucleotide transferase and transferase activity whereas molecular function (MF) includes restriction which involves heterocyclic, organic, and cyclic compound binding, as well as small molecule and nucleic acid restricting activity. From GO analysis, we found that the high outcome in catalytic activity and RNA binding with more than 50. On the other hand, we found the last outcomes in RNA polymerase and transferase activity equal to 41. Whereas the same functions were also observed in the case of biological process (BP) but all functions have sane score equal to (61) and in case of cellular component (CC), all functions have score equal to 63 and in host cell membrane, the score was 54.

D. Modeling of Protein Homology and Validation of X-domain Structure

The sequence of the target protein was keyed in as the information file in the workspace of the employed Swiss-Model. A search in the Swiss-MODEL layout library (SMTL) using HHBlits yielded around complete 50 formats. Out of the 50 templates, only 6 nur.A1 was the suitable template. Then, the objective sequence was selected based on the outcome of the Qualitative Model Energy Analysis (QMEAN) (-0.77), identity sequence proportion (95.83), global model quality estimate (GMQE) 0.74, and coverage (93%). The model relied on track format alignment using ProMod3 with remodeling of the Indel and revamping of the side chains. The achieved model in this study was similar to 6nur.1. The generated model was saved in the PDB format, as shown in Fig. 3. The assessment of the stereochemical nature of the predicted protein structure was done using the Ramachandran plotting map, as shown in Fig. 4. The assessment showed that 534 of the all-out residues (91.1%) domiciled in favored area (A; B; and L) whereas 9% domiciled in the permitted area (a; b; and l). No residue was found in the prohibited region. The quality of the X domain model was predicted through analysis of 118 good resolution structures (2.0 A°) with R-factor of <20%. Deviation of 5.8 in maximum residue properties, as well as bond angle of 5.3 and 93.1% planar, was observed within the cutoff point through PROCHECK analysis. PHYRE2-based modeling of protein homology showed a protein model that modeled 77% of the protein residues (538) at a confidence level of >90% (Fig. 5) which

Fig. 3. SWISS-Model-predicted protein structure with helix strands and coil.

Fig. 1. PSIPRED-based 2° protein model prediction.

Fig. 2. Observed randomness condition from the DISOPRED Web.

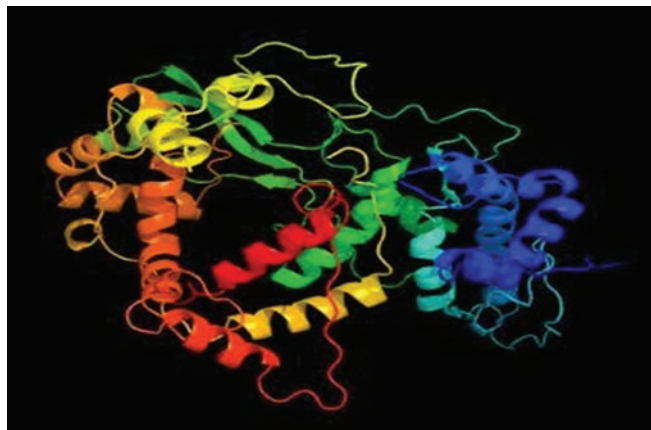Fig. 4. Ramachandran plot of the structure of the predicted protein.

Fig. 5. PHYRE2-predicted protein structure.
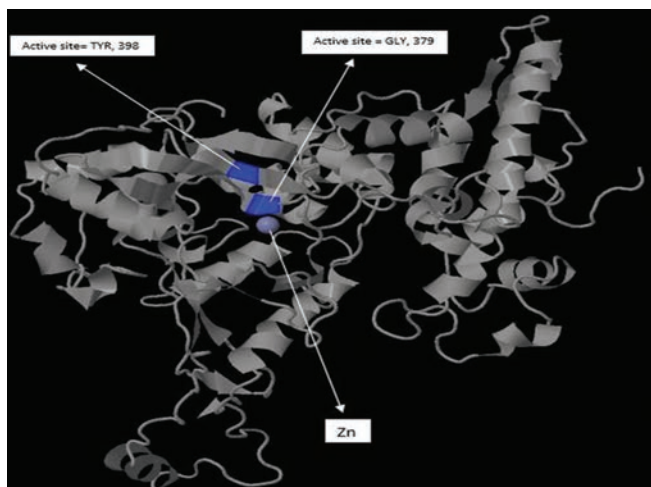

Fig. 6. 3D protein model plus Zn^{2+} ligand on the binding site.

implies the following coordinates X = 72.1690; Y = 75.6500; and Z = 86.0880. The expected secondary structure by Phyre2 was given as 53% alpha-helix, 13% disordered, and 11% beta strand. In the final selected model, the prediction of the 3D ligand site was reliant on Cluster 1 which showed 1 ligand and 1 structure. Only two binding sites were successfully predicted as follows Gly: 1 contact (at residue no. 379) and Tyr: 1 contact (at residue no. 398). The predicted metallic heterogeneous ligand was Zn (1) which binds to the binding site of the predicted ligand (Fig. 6).

Conclusion

We determine the domain structural model of coronavirus orf1ab polyprotein with predicted active site for ligand binding site. This supplies premeditations into the functional role of polyprotein domain in viral pathogenesis.

References

Gasteiger, J. and Walker, M., 2005. *The Proteomics Protocols Handbook*. Springer, Berlin, p.571.

Geourjon, C. and Deleage, G.C., 1995. SOPMA: Significant improvements in

protein secondary structure prediction by consensus prediction from multiple alignments. *Computer Applied Bioscience*, 681(11), pp.681-684.

Gil, S.C. and Von Hippel, P.H., 1989. Calculation of protein extinction coefficients from amino acid sequence data. *Analytical Biochemistry*, 182(2), pp.319-326.

Hamp, T., Kassner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron, M., Hönigschmid, P., Hopf, T.A., Kaufmann, S.A., Kiening, M., Krompass, D., Landerer, C., Mahlich, Y., Roos, M. and Rost, B., 2013. Homology-based inference sets the bar high for protein function prediction. *BMC Journal Bioinformatics*, 14(7), pp.502-511.

Kelly, L.A, Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10(6), p.845.

Kyte, J. and Doolottle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1), pp.105-132.

Li, S., Yuan, L., Dai, G., Chen, R.A., Liu, D.X. and Fung, T.S., 2020. Regulation of the ER stress response by the Ion channel activity of the infectious bronchitis coronavirus envelope protein modulates virion release, apoptosis, viral fitness, and pathogenesis. *Frontiers in Microbiology*, 10, p.3022.

McBride, R. and Fielding, B.C., 2012. The role of severe acute respiratory syndrome (SARS) coronavirus accessory proteins in virus pathogenesis. *Viruses*, 4(11), pp.2902-2923.

Ofran, Y. and Rost, B., 2007. ISIS: Interaction sites identified from sequence. *Bioinformatics Journal*, 23(2), pp.e13-e16.

Paraskevis, D., Kostaki, E.G., Magiorkinis, G., Panayiotakopoulos, G., Sourvinos, G. and Tsiodras, S., 2020. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infection Genetics and Evolution*, 79, pp.104-212.

Raik, G. and Luis, S., 2010. Strategies for protein synthetic biology. *Nucleic Acids Research Journal*, 38(8), pp.2663-2675.

Remmert, M., Andreas, B., Andreas, H. and Johannes, S., 2012. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2), pp.173-178.

Rost, B., Guy, Y. and Liu, J., 2004. The predicProtein server. *Nucleic Acid Research Journal*, 32(2), pp.321-326.

Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., Zimmer, T., Thiel, V., Janke, C., Guggemos, W., Seilmaier, M., Drosten, C., Vollmar, P., Zwirglmaier, K., Zange, S., Wölfel, R. and Hoelscher, M., 2020. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *New England Journal of Medicine*, 382(10), pp.970-971.

Schaecher, S.R. and Pekosz, A., 2010. *Molecular Biology of the SARS-Coronavirus*. Springer, Berlin, pp.153-166.

To, K.K.W., Tsang, O.T., Yip, C.C., Chan, K.H., Wu, T.C., Chan, J.M., Leung, W.S., Chik, T.S., Choi, C.Y., Kandamby, D.H., Lung, D.C., Tam, A.R., Poon, R.W., Fung, A.Y., Hung, I.F., Cheng, V.C., Chan, J.F. and Yuen, K.Y., 2020. Consistent detection of 2019 novel coronavirus in saliva. *Clinical Infectious Diseases*, 71(15), pp.841-843.

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J., Sheng, J., Quan, L., Xia, Z., Tan, W., Cheng, G. and Jiang, T., 2020. Genome composition and divergence of the novel coronavirus (2019- nCoV) originating in China. *Cell Host and Microbe*, 27(3), pp.325-328.

Wu, F., Zhao, S., Yu, B., Chen, Y., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C. and Zhang, Y.Z., 2020. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798), pp.265-269.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F. and Tan, W., 2020. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*, 382(8), pp.727-733.

Machine Learning Algorithms for Detecting and Analyzing Social Bots Using a Novel Dataset

Niyaz Jalal[1] and Kayhan Z. Ghafoor[1,2]

[1]Department of Software and Informatics Engineering, College of Engineering, Salahaddin University-Erbil,
Erbil 44001, Iraq

[2]Department of Computer Science, Knowledge University,
Erbil 44001, Iraq

Abstract—**Social media is internet-based technology and an electronic form of communication that facilitates sharing of ideas, documents, and personal information. Twitter is a microblogging platform and is the most effective social service for posting microblogs and likings, commenting, sharing, and communicating with others. The problem we are shedding light on in this paper is the misuse of bots on Twitter. The purpose of bots is to automate specific repetitive tasks instead of human interaction. However, bots are misused to influence people's minds by spreading rumors and conspiracy related to controversial topics. In this paper, we initiate a new benchmark created on a 1.5M Twitter profile. We train different supervised machine learning on our benchmark to detect bots on Twitter. In addition to increasing benchmark scalability, various autofeature selections are utilized to identify the most influential features and remove the less influential ones. Furthermore, over-/ under-sampling is applied to reduce the imbalance effect on the benchmark. Finally, our benchmark compared with other state-of-the-art benchmarks and achieved a 6% higher area under the curve than other datasets in the case of generalization, improving the model performance by at least 2% by applying over-/under-sampling.**

Index Terms—**Machine learning, Misinformation detection, Twitter bot detection, Twitter profile metadata.**

I. Introduction

In the past decade, the influence of social media has increased rapidly; this growth will be more tangible in the upcoming years. One of the famous social platforms is Twitter; millions of people, including influential figures, have Twitter accounts to interact with their audiences. Besides the advantages, Twitter is used to exploit and delude others in many events, such as the COVID-19 outbreak, the Russian invasion of Ukraine, and the USA presidential elections in 2016 and 2020, by spreading falsified conspiracies and manipulating public opinion (Khanday, Khan and Rabani, 2021; Shevtsov, et al., 2021).

These actions are usually performed by automated programs, so-called bots. Broadly speaking, bots aim to ease automation processes such as sending a friendly message or giving some instruction on social media. Yet, the automation capability of bots is unfavorably used for spreading spam, fake news, and hate speech (Davis, et al., 2016).

A botmaster manages social bot accounts, which controls many social bots to influence public opinion toward a specific ideology or purpose by spreading low credible information (Ferrara, et al., 2016). The social bot impact is so significant that some reports indicate that 9–15% of the active accounts on Twitter are social bots (Varol, et al., 2017). In some cases, the fake news reached 100K users, and false information had 70% more retweets than trustworthy news (Hanouna, et al., 2019; Orabi, et al., 2020a).

In this paper, the machine learning framework is proposed based on a novel dataset. We collected more than 1.5M Twitter accounts during the US presidential election in 3 months. Our dataset only includes the metadata (profile) features that contain a small number of features. The benefits of using profile features are increasing the model's scalability and decreasing training time. The main goal of this paper is to create a benchmark suitable for real-time bot detection by increasing the number of samples and reducing the number of features. At the same time, dataset generalization has increased by 6% average AUC compared to other datasets in this research area. Finally, we can conclude our main distribution as follows:

1. We collected 1.5 million users and created up-to-date 100K datasets; to the best of our knowledge, it is the biggest only metadata dataset.
2. We increase generalization, and our dataset achieves a 6% higher average accuracy among all datasets in the research area.
3. We improve the model performance by at least 2% by introducing over-/under-sampling algorithms with our dataset.
4. We interpret the prediction of our machine learning models with the help of Shapley Additive Explanations (SHAP)

II. Related Work

Different types of bots target various audiences; for example, spam bots for spreading spam, fake follower for increasing followers of a particular account, COVID, or political bots for spreading conspiracy. Alom, Carminati and Ferrari (2018) and Shukla, Jagtap and Patil (2021) tried to detect spam bots on Twitter platform. Alom, Carminati and Ferrari (2018) have used metadata and graph-based features from the 42K dataset collected by a Social Honeypot. They used different ML models; the random forest (RF) achieved the best result compared to other models. Furthermore, Shukla, Jagtap and Patil (2021) have used a public dataset with 38K users and 19 features with different Ensemble models, feature engineering, feature encoding, and feature selection. The Ensemble model with ANN, RF, and AB achieved the best result and also best ensemble model outperformed other individual models by attaining 93% AUC.

Feng, et al. (2021); Khanday, Khan and Rabani, (2021); and Shevtsov, et al. (2021) focused on detecting bots that spread rumors. Khanday, Khan and Rabani (2021) have used Tweeter API to collect tweets related to the COVID19 outbreak. Different ML applied to detect bots that spread COVID rumors. The decision tree (DT) gave the best result with a 99% F1 score; the result showed that tweets generated by bots have a greater length than regular tweets. Shevtsov, et al. (2021) have collected tweets related to the 2020 US presidential election to detect political bots. The data were collected over 2 months with a total of 15.6M tweets and 3.2M users. Furthermore, Bot Sentinel and Botometer were used to label the dataset. Different ML, feature engineering, feature selection, and under-/over-sampling applied to the dataset; as a result, XGBoost achieved 92% as the best F1 score. Feng, et al. (2021) have collected tweets related to different hashtags with 34M tweets and 8M users to create a dataset with a variety of bots in it (called TwiBot-20). The dataset trained with the previous works models, the result shows the model accuracy declined with TwiBot-20. They blamed the absence of user diversity, limited user information, and data scarcity for this decline.

On the other hand, Yang, et al. (2019) and Hayawi, et al. (2022) tried to increase model scalability and reduce training time by only using metadata to train a model. Yang, et al. (2019) have focused on enhancing the generalization of detection models. They used 14 public datasets for training the models; in some cases, two or more datasets have combined. They manually selected 20 features from datasets, then used the RF for training. The RF trained with the entire dataset and then tested with other datasets. As a result of this approach, the generalization between the datasets is very low. In the second experiment, they selected seven datasets for training, and all possible combinations between them were trained (247 combinations) and tested the model with the remaining datasets using the RF classifier. This approach improved the generalization between datasets. Hayawi, et al. (2022) have used public only metadata datasets to train deep neural networks. Furthermore, text features have transformed with long short-term memory (LSTM) and GLoVE. The

same practice as Yang, et al., 2019, was followed: Training the model on one dataset and evaluating it with other datasets. The final results show improvement in the model generalization.

Kudugunta and Ferrara (2018), Rodríguez-Ruiz, et al. (2020), and Martin-Gutierrez, et al. (2021) introduced new detection models for detecting bots on Twitter. Kudugunta and Ferrara (2018) have used a neural network with LSTM and GloVE to detect bots. A total of 16 features from metadata and tweet content from the public dataset had used to train the model. Furthermore, the over-/under-sampling technique used to balance the dataset. The result is 96% AUC for a single tweet account and 99% AUC for account level. Martin-Gutierrez, et al. (2021) have focused on detecting multilingual bots using a deep learning model. The BERT, Flair, and RoBERT models were used separately to transform text-based features into a numeric vector. The model was trained with a 60K public dataset and achieved a 77% F1 score. Rodríguez-Ruiz, et al. (2020) have used a one-class classification. This classification algorithm determines whether the data belong to this class or not, which usually uses imbalanced data. They trained public datasets with binary and one-class classification models with different algorithms. The results indicate that one-class classification achieved a better result than binary classification.

Finally, this paper aims to create a benchmark suitable for real-time bot detection by increasing the number of samples using labeling API to make the labeling process faster and having more labeled data, decreasing the number of features by selecting the most influential features, and ignoring the less influential ones. Furthermore, oversampling and undersampling are used to overcome the over=/under-fitting problem.

III. Models

A. Machine Learning Models

Different machine learning algorithms have used to train the proposal dataset: Adaboost Classifier (AB), Bagging Classifier (BC), DT, Extra Tree Classifier (ET), Gaussian Naive Bayes (GN), K-nearest neighbor (KNN), logistic regression (LR), random forest (RF), support vector machines (SVM), and eXtreme Gradient Boosting (XGBoost). However, RF and ET have tested separately with 100 and 500 trees, and KNN with three and five neighbors. Except for XGBoost, the sklearn Python package was used for all models. Furthermore, 5-fold cross-validation with random shuffling splits the dataset into five parts; the training and validation occur in four portions, and the last portion tests the model performance. Accuracy, precision, recall, F1 score, and AUC (Huang and Ling, 2005) were used to evaluate model performance. (Fig. 1 shows the steps for the experiment).

B. Feature Selection Models

Feature selection is an essential part of any machine learning algorithm. Feature selection has many advantages, for example, removing irrelevant data and reducing data
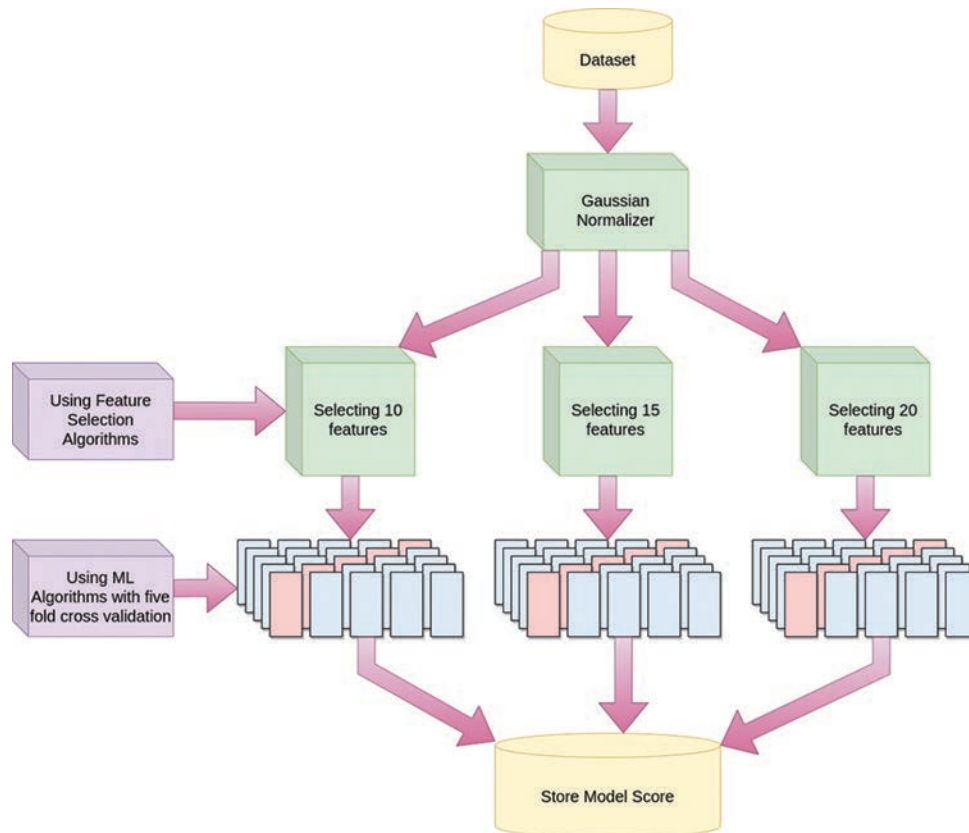
Fig. 1. The proposed experiments to detect the best model approaches.

dimensions, increasing model performances, reducing overfitting and underfitting, and speeding up training time (Shevtsov, et al., 2021).

There are two types of feature selection, manual selection and automatic selection. In our work, we perform both types of feature selections. Two types of features were excluded from original user metadata by manual selection. The first group features have different values for each Twitter account. The second group is the features with almost the same values for all Twitter accounts.

On the other hand, autoselection models use statistics correlation, ML algorithm, permutation importance, and coefficient scores to find the relationship between features; when features are more correlated, a higher score will assign to the candidate feature. In our work, we used recursive feature elimination (RFE) (Granitto, et al., 2006), univariate feature selection (UFS) (Jović, Brkić and Bogunović, 2015), and feature importance (FSI) (Altmann, et al., 2010). All three techniques are implemented in the sklearn Python package.

Our work used: RFE-DT, FSI-RF, FSI-XGboost, ANOVA F-value (Shaw and Mitchell-Olds, 1993), and information gain (Peng, Long and Ding, 2005); we tried to extract 10, 15, and 20 separately using each of the mentioned algorithms.

C. Over-/Under-Sampling Models

When one class has majority samples and others have a minority, oversampling and undersampling are usually used for balancing a dataset. There are a variety of techniques for balancing a dataset; in case of oversampling, the number of minority classes will increase, but under sampling reduces the sample numbers of majority classes (Shevtsov, et al., 2021).

Moreover, some techniques combine both over-/under-sampling. The synthetic minority oversampling technique (SMOTE) (Wang, et al., 2006) is the most popular technique to increase the number of minority classes. SMOTE uses KNN to generate N number of synthetic samples for each sample in the minority class. Furthermore, there are other modified versions of SMOTE, such as ADASYN, Borderline-SMOTE, and SVM-SMOTE.

On the other hand, edited nearest neighbor (ENN) (Wilson, 1972) is an undersampling algorithm that utilizes KNN to eliminate samples close to the boundary or misclassified. Tomek Links (Elhassan and Aljurf, 2016) is another undersampling model created based on condensed nearest neighbors (CNNs). The CNNs randomly chose samples to eliminate from the majority class, but in the case of Tomek Links, the KNN was used to find those samples with the lowest distance from the minority class, then remove them.

In our experiments, mentioned algorithms are applied individually to the dataset, and the results show little or no improvement in the model's performance. As a second experiment, both under-/over-sampling algorithms are combined; first, SMOTE and Tomek links are combined (SMOTETomek), then SMOTE and edited nearest neighbor are combined (SMOTEENN). Applying both under-/over-sampling algorithms to our datasets accelerate the model

performance. We discuss their results in the performance evaluation section.

IV. Datasets

A. Data Collection

The lack of datasets is a barrier to social bot detection (Adewole, et al., 2017). Collecting data from social platforms are prohibited by most influential social media platforms; Twitter is one of a few platforms allowing collecting public data by providing API (Orabi, et al., 2020b). We collected data from Twitter using Twitter API, retrieving tweets related to the USA 2020 presidential election topics. The data collection process started 3 months before the election day (in July 28, 2020) and ended 16 days (November 19, 2020) after the election day. This topic is selected because it is a controversial topic on Twitter and makes all types of bot involvement very high. As shown in Table I, 19 hashtags used to collect data, and a total of 1.5M unique Twitter accounts were collected.

B. Profile Features

The retrieved data from Twitter API contain more than 1000 features, mainly divided into four categories, profile feature (user metadata), context feature, time-based feature, and interaction features (Yang, et al., 2019). All categories have been used to detect bots on Twitter platforms (Orabi, et al., 2020b). Still, in our work, we only focus on the profile feature, a small object containing all critical information related to an account. The advantage of using profile features is reducing training time without affecting model performance. The user profile object contains 44 features in total, such as name, description, picture's URL, colors, statistical, Boolean flags, language related, and other parameters. We divided the profile features into three groups: Newly derived, dropped, and direct used features. In the next sections, we describe each of them. It is also worth mentioning that the latest version of Twitter API removes some features from user profile object, but they are not essential features; the feature selection process eliminated them.

C. Dropped Features

The values of properties from profile features are different; some properties have different (unique) values for each account, and others are totally similar (include none values). In those cases, properties do not give any meaningful

information to distinguish between humans and bots. After analyzing each property, the result shows that 23 features represent ineffective data, meaning either their values are different for each account or they have one or two unique values for all accounts regardless of account type (bot or human). id_str, translator_type, time_zone, contributors_ enabled, following, profile_text_color, profile_background_ color, profile_sidebar_border_color, profile_link_color, profile_sidebar_fill_color, lang, withheld_in_countries, is_ translator, utc_offset, notifications, id, follow_request_sent, is_translation_enabled, protected, profile_background_image_ url_https, profile_background_image_url, profile_image_url, and profile_image_url_https features are eliminated from our datasets.

D. Direct used Features

The majority of the previous works have used a set of features and they have proven their effectiveness (Yang, et al., 2019; Martin-Gutierrez, et al., 2021). Those features are either numeric or binary features. The numeric features obtain essential statistical information such as the number of followers. The binary features give the basic account information, for example, is account verified or not. Followers_count, friends_count, listed_count, statuses_count, favourites_count, has_extended_profile, default_profile, geo_enabled, and verified are some features that exist in a majority of previous works. Furthermore, we add default_ profile_image, profile_use_background_image, and profile_ background_tile features to our dataset by converting the Boolean values (true/false) into numerical representation (1/0) (Equation 6). Table II shows features: Name, types, and description. In our work, all features mentioned in this section are used directly without performing any feature engineering; simultaneously, we use the numeric feature to create new features.

E. New Derived Features

Derived features are created by performing a mathematical or logical operation on one or more original features. We made 21 new features based on calculating other features. The previous works proved that the extracted features include valuable information for the detection process (Hayawi, et al., 2022). We create three new feature sets for statistical, text, and binary features. In the case of statistical features, the features are divided by an account age (the account creation date minus the last tweet date in our dataset). Furthermore, we find the ratio between followers and friends. The second set is text features; the metadata object contains name, screen_name, and description as a text feature. We calculate the text length for name, screen_ name, and description (Equation 1); the number of digits (Equation 2), frequency (Equation 5 when numerator is the n-g (Brown, et al., n.d.), and the denominator is the total unique characters.), and entropy (shows in Equation 3 when numerator is Shannon entropy (Shannon and Weaver, 1949) and denominator is feature length) for name and screen_name; similarity between name and screen_name

TABLE I
LIST OF HASHTAGS THAT USED FOR COLLECTING DATA

Hashtag		
#democrat	#politics	#america
#mikepence	#vote	#guncontrol
#election	#bluewave	#getoutthevote
#republican	#gop	#president
#racistinchief	#joebiden	#trump
#biden	#trumpvoters	#donaldtrump
	#racistpresident	

TABLE II
LIST OF FEATURES USED IN THIS PAPER WITH THEIR DATA TYPE AND SHORT DESCRIPTION

Feature	type	Description
listed_count	Numerical	Public lists that use members of
statuses_count	Numerical	Total number of tweets and retweet
friends_count	Numerical	Number of users the account following
favorites_count	Numerical	Total number of tweets liked by an account
followers_count	Numerical	Number of users following this account
default_profile	Binary	If profile is set
has_extended_profile	Binary	If profile is extend is true otherwise false
geo_enabled	Binary	Twitter has access to profile location
verified	Binary	If profile is verified
profile_background_tile	Binary	If profile is tiled is true otherwise false
profile_use_background_image	Binary	If profile has background image
default_profile_image	Binary	If profile image is default
profile_has_banner_url (DB)	Binary	If profile has banner URL
has_entities (DB)	Binary	If entities is set is true otherwise false
has_location (DB)	Binary	If location is set is true otherwise false
has_url (DB)	Binary	If URL is set is true otherwise false
account_age (DB)	Numerical	last_tweet_date - account_created_date
name_length (DT)	Numerical	Length of name feature
screen_name_length (DT)	Numerical	Length of screen_name feature
description_length (DT)	Numerical	Length of description feature
num_digits_in_name (DT)	Numerical	Number of digit in name feature
num_digits_in_screen_name (DT)	Numerical	Number of digit in screen_name feature
name_entropy (DT)	Numerical	Entropy of name feature
screen_name_entropy (DT)	Numerical	Entropy of screen_name feature
screen_name_freq (DT)	Numerical	Mean bigram frequency for screen name
name_freq (DT)	Numerical	Mean bigram frequency for name
name_similarity (DT)	Numerical	Similarity between screen_name and name
tweet_freq (DS)	Numerical	Statuses_count/account_age
followers_growth_rate (DS)	Numerical	followers_count/account_age
favourites_growth_rate (DS)	Numerical	favourites_count/account_age
listed_growth_rate (DS)	Numerical	listed_count/account_age
friends_growth_rate (DS)	Numerical	friends_count/account_age
followers_friends_ratio (DS)	Numerical	followers_count/friends_count

DB: Binary derived, DS: Statistically derived, DT: Text derived

(shows in Equation 4 when numerator is the number of matches, and the denominator is the length of both texts).

The last set includes features that either contains a value or not (Equation 6); for example, the location feature is either set or not (empty). In our work, we call those features binary-derived features; we convert them to one if it has a value; otherwise, we convert them to zero. Table II shows statistically derived features (their names end with DS), text-derived features (their names end with DT), and binary-derived features (their names end with DB.)

F. Data Labeling

Data labeling is a process of giving actual labels to the data; labeling is a mandatory step for supervised machine learning (Derhab, et al., 2021). We use the data label for comparing with predicted results by ML models to determine how well the model performed. As mentioned earlier, the total number of users in our dataset is more than 1.5M. Two possible ways to know

$$Feature_{len} = \begin{cases} len\,(feature),\ len\,(feature) \geq 0 \\ 0,\ None \end{cases}$$

$$Feature_{digit\_count} = \begin{cases} digit\,(feature),\ len\,(feature) \geq 0 \\ 0,\ empty \end{cases}$$

$$Feature_{entropy} = \frac{-\sum_{i=1}^{k} p(i)*\log_2 p(i)}{len\,(feature)}$$

$$Name_{similarity} = \frac{2M}{N}$$

$$Feature_{freq} = \frac{\sum_{i=1}^{k} C(b_i)}{K}$$

$$Feature_{convert}(x) = \begin{cases} 1,\ (x = True\ or\ x = Vaue\ exist) \\ 0,\ (x = False\ or\ x = Value\ not\ exist) \end{cases}$$

the actual data labels are manually checking or using previously trained models. In our case, manual checking is impossible due to dataset size; therefore, we choose Botometer API to classify our dataset. The Botometer has two versions: BotometerLite, which only uses metadata to classify; Botometer-V4, which predicts based on more than 1000 features, including tweets, metadata, and other features. In our case, we use BotometerLite because it follows the same strategy as our work by only focusing on metadata features for labeling data. The BotometerLite allows 20K samples to be labeled every day. In our case, it took around 75 days to classify 1.5M samples. The BotometerLite score sample is between 0 and 1. In our case, the API could not classify approximately 16.9% of our data, and 83.1% remaining samples are classified, as shown in Table III. To create our dataset, we classify any account score equal to or greater than 0.6 as bot accounts, so 42K accounts are classified as

bots. Regarding a human account, we randomly selected 58K accounts that scored <0.1. However, all the samples bigger than 0.1 and <0.6 are ignored to have a big gap between bot and human. In conclusion, we created a dataset with 100K samples classified based on BotometerLite. Table III shows the start and end of the labeling process and the total accounts that were classified based on the specific range.

V. Performance Evaluation

A. Performance Evaluation of Machine Learning Models

We tested our dataset with 10 different models (Fig. 1 shows the steps for the experiment); however, we used two versions of KNN, RF, and BC and 5-fold cross-validation to reduce the chance of overfitting in the dataset. However, different algorithms were tested to normalize our dataset, such as scaling, standardizing, transforming, and normalizing. Transforming (Gaussian normalization) usually achieves the best result. Furthermore, it is worth mentioning that the dataset without normalization was tested but achieved a very low result. Therefore, we use Gaussian normalization in all experiments in this paper. Five selection models are utilized; the feature selection extracts 10, 15, and 20 features separately. We trained different models (using 13 ML algorithms). Table IV shows the max, min, and mean result based on an accurate measurement. As shown in Table IV, RF with 20 features and ET with 15 features achieved the best accuracy result. However, surprisingly, increasing the number of features has a very slim effect on performance and an enormous impact on training time. For example, in the case of RF with 100 trees, the difference between the model with 20 features and 10 features is just 0.04.

Nevertheless, in the case of AB, KNN with five neighbors, and RF with 500 trees, the model performed better when trained with 10 features rather than 15 or 20. Moreover, the mean columns show that SVM and BC overall performance achieve the best result. Another interesting finding is that the increasing number of trees or neighbors has little impact on the performance. For instance, in the case of ET, the model achieved precisely the same result, RF performance declined slightly, and KNN performance increased somewhat. Table V shows the result of the same experience as Table IV, but besides average (mean) accuracy, it also reveals the average of recall, precision, F1 score, and AUC.

B. Feature Selection Evaluation

As mentioned earlier, we trained different models based on five feature selection algorithms – the performance of all algorithms was excellent in 15–20 features. REF-DT obtains the best performance, a little bit higher than other algorithms. However, for average best performance, information gain gets a higher result. However, in 10 features, the performance of the ANOVA F-value declined by more than 10%, but other algorithms performed relatively high. Nevertheless, when we consider time, XGBoost is achieving the best; for instance, in the case of 20 features, it is 3 times faster than RFE-DT with a 0.45 decline in performance; the same is true for 10–15

TABLE III
Labeling Process for BotometerLite

Score range	BotometerLite
Start date	January 27, 2022
End date	April 13, 2022
0–0.09	310,026
0.1–0.19	321,790
0.2–0.29	236,290
0.3–0.39	175,183
0.4–0.49	112,001
0.5–0.59	50,313
0.6–0.69	19,837
0.7–0.79	12,367
0.8–0.89	6995
0.9–1	2134
Error	253,064
Total	1,500,000

TABLE IV
Mean (average), Maximum, and Minimum Accuracy Values for Different ML Algorithms

ML models	20			15			10		
	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean
AB	98.00	73.55	94.32	97.98	73.43	94.05	98.01	73.45	92.06
BC-100	98.61	73.52	95.22	98.61	73.52	95.03	98.57	73.46	92.26
DT	97.67	73.60	94.19	97.69	73.43	93.94	97.65	73.52	91.16
ET-100	98.66	73.46	94.67	98.64	73.52	94.48	98.64	73.39	91.93
ET-500	98.66	73.51	94.72	98.67	73.50	94.59	98.64	73.52	92.00
GN	93.09	67.83	80.87	92.83	67.50	80.09	92.34	67.01	78.59
KNN-3	97.94	72.24	89.65	98.25	73.44	91.23	98.29	60.84	89.82
KNN-5	98.05	67.03	89.89	98.32	73.44	91.25	98.37	73.40	90.80
LR	93.88	69.86	85.70	93.52	62.49	84.64	93.38	66.84	83.07
RF-100	98.67	73.62	95.21	98.65	73.43	94.97	98.63	73.50	92.21
RF-500	98.00	73.55	94.32	97.98	73.43	94.05	98.01	73.45	92.06
SVM	98.61	73.52	95.22	98.61	73.52	95.03	98.57	73.46	92.26
XGboost	97.67	73.60	94.19	97.69	73.43	93.94	97.65	73.52	91.16

features with a little difference. Surprisingly, the appropriate time for 15 features is higher than for 10–20 feature models. We thought that the time for the 15 models should be between 10 and 20. We achieved the time results using the sklearn Python package, which is a very respectful package among the Python community; yet, we are not sure about the exact reasons. Table VI shows the mean (average), maximum, and minimum accuracy for different feature selection models; each value is calculated based on 13 ML models (Table IV). It also shows FitTime (the time for fitting the model for each train set split) and ScoreTime (the time for fitting the model on the test set).

However, regarding the most influential features for the 10 features approach, our algorithms selected favorites_count, friends_growth_rate, friends_count, followers_count, num_digits_in_screen_name, tweet_freq, favorites_growth_rate, followers_growth_rate, followers_friends_ratio, and description_length as the most 10 influential features. The same experiment was repeated for the 15-feature approach; besides, features in the 10-feature list, also listed_count, statuses_count, Name_similarity, account_age, and Name_entropy have added to complete the 15-feature list. Finally,

TABLE V
MEAN (AVERAGE) FOR DIFFERENT ML ALGORITHMS WITH VARIOUS PERFORMANCE MEASUREMENTS

ML models	20					15					10				
	A	F1	R	P	AUC	A	F1	R	P	AUC	A	F1	R	P	AUC
AB	94.32	96.98	92.56	90.75	95.37	94.05	96.61	91.99	89.90	95.56	92.06	94.70	89.03	85.88	94.64
BC-100	95.22	97.02	93.67	92.43	95.83	95.03	97.27	93.67	92.75	95.11	92.26	94.45	89.76	88.21	93.05
DT	94.19	94.76	92.55	91.79	94.08	93.94	94.36	92.04	91.17	94.18	91.16	91.15	88.27	86.81	91.75
ET-100	94.67	96.95	92.90	91.18	95.74	94.48	97.18	92.98	91.62	94.87	91.93	93.77	88.86	86.63	93.73
ET-500	94.72	97.07	93.06	91.53	95.56	94.59	96.92	92.91	91.48	95.37	92.00	94.03	89.36	87.62	92.96
GN	80.87	90.20	73.18	68.39	85.66	80.09	90.15	72.10	68.00	85.18	78.59	88.54	67.33	60.02	86.66
KNN-3	89.65	92.34	87.26	85.61	89.51	91.23	93.18	88.30	85.86	92.80	89.82	91.98	87.43	86.15	90.78
KNN-5	89.89	93.42	87.59	85.61	90.26	91.25	93.84	88.42	86.03	92.71	90.80	92.74	87.68	84.99	92.68
LR	85.70	91.58	81.41	78.40	86.09	84.64	91.23	78.57	75.54	86.82	83.07	88.96	76.69	71.94	85.53
RF-100	95.21	97.16	93.80	92.82	95.56	94.97	96.81	93.23	91.95	95.86	92.21	94.37	89.41	87.52	93.51
RF-500	94.74	96.76	93.02	91.65	95.36	95.06	96.99	93.53	92.50	95.47	92.13	94.46	89.55	88.07	92.91
SVM	90.54	94.42	87.93	85.77	91.18	90.02	93.90	87.07	84.66	91.08	87.99	91.78	83.86	80.40	90.28
XGboost	95.16	97.14	93.45	91.83	96.25	95.06	96.91	93.30	91.92	96.07	92.84	95.03	89.97	87.14	95.34

A: Accuracy, F1: F1 score, R: Recall, P: Precision

TABLE VI
MEAN (AVERAGE), MAXIMUM, AND MINIMUM ACCURACY VALUES FOR DIFFERENT FEATURE SELECTION MODELS

Selection models	20					15					10				
	MAX	MIN	MEAN	FT	ST	MEAN	MAX	MIN	FT	ST	MIN	MEAN	MAX	FT	ST
ANOVA F-value	97.04	76.96	91.09	228.58	4.98	96.61	75.30	91.07	308.07	6.93	86.46	75.09	82.26	123.15	7.52
Information gain	98.62	73.52	93.05	340.72	4.13	98.52	73.52	92.98	461.53	5.72	98.48	60.84	91.75	315.85	5.10
RFE-DT	98.67	72.24	92.44	488.23	5.03	98.71	73.47	92.30	731.81	5.90	98.64	69.76	91.87	535.74	4.87
RF	98.46	67.83	91.95	296.79	5.08	98.47	67.50	91.42	415.34	5.46	98.47	67.39	91.87	276.53	5.02
XGboost	98.22	67.03	91.49	123.83	4.35	98.15	62.49	91.11	454.45	5.17	98.21	66.84	91.62	123.26	4.41

FT: Fit Time, ST: Score time

TABLE VII
MEANS FOR DIFFERENT PERFORMANCE MEASUREMENTS WITH OVER-/UNDER-SAMPLING MODELS TESTED WITH VARIOUS ML MODELS

ML model	None					SMOTETomek					SMOTEENN				
	A	F1	R	P	AUC	A	F1	R	P	AUC	A	F1	R	P	AUC
AB	97.89	99.61	97.45	96.85	98.05	98.27	99.77	98.26	97.89	98.63	99.70	99.99	99.70	99.65	99.75
BC-100	98.55	99.63	98.25	97.94	98.57	98.94	99.80	98.94	98.85	99.04	99.81	99.98	99.81	99.83	99.80
DT	97.60	97.53	97.11	97.15	97.08	98.29	98.29	98.29	98.33	98.25	99.71	99.71	99.71	99.72	99.71
ET-100	98.61	99.72	98.32	97.94	98.70	99.03	99.87	99.03	98.90	99.17	99.87	100	99.87	99.86	99.89
ET-500	98.63	99.75	98.34	97.98	98.71	99.03	99.89	99.03	98.91	99.16	99.88	100	99.88	99.89	99.88
GN	89.55	97.67	86.28	79.01	95.02	88.72	97.87	87.72	80.60	96.23	90.03	98.48	89.29	82.87	96.79
KNN-3	98.32	98.91	97.97	97.18	98.77	98.72	99.38	98.72	98.30	99.13	99.78	99.91	99.78	99.71	99.84
KNN-5	98.36	99.13	98.00	97.11	98.92	98.76	99.53	98.75	98.27	99.24	99.74	99.93	99.74	99.60	99.87
LR	91.27	94.43	89.10	85.79	92.68	90.94	94.77	90.72	88.57	92.98	92.94	95.81	92.83	91.18	94.55
RF-100	98.62	99.70	98.34	98.07	98.60	98.95	99.85	98.95	98.87	99.03	99.86	100	99.86	99.91	99.82
RF-500	98.63	99.73	98.34	98.10	98.59	98.98	99.87	98.98	98.90	99.06	99.84	100	99.84	99.89	99.79
SVM	98.45	99.65	98.13	97.64	98.62	98.76	99.77	98.76	98.42	99.10	99.68	99.97	99.68	99.56	99.79
XGboost	98.62	99.74	98.33	98.03	98.63	98.98	99.88	98.98	98.87	99.10	99.89	100	P. 89	99.89	99.89

A: Accuracy, F1: F1 score, R: Recall, P: Precision

TABLE VIII
FIVE BEST AUC-ROC ACHIEVED ML MODELS

Selection model	ROC AUC mean
SVM	75.75
ET-500	75.03
KNN-5	74.73
ET-100	74.66
KNN-3	74.40
SVM	75.75
ET-500	75.03

for the 20-feature approach, geo_enabled, Screen_name_freq, Name_freq, screen_name_length, and name_length features have been chosen with features from the 15-feature set to complete the 20-feature list. In addition, we draw the SHAP plot (Lundberg, Erion, and Lee, 2018) for the dataset using a RF model. The X-axis represents a SHAP value for each feature and the Y-axis shows features name order by most to least influential. In the SHAP plot, a positive value means pushing to one (bot-like), and a negative value means driving to zero (human-like). Color is another indication in the SHAP
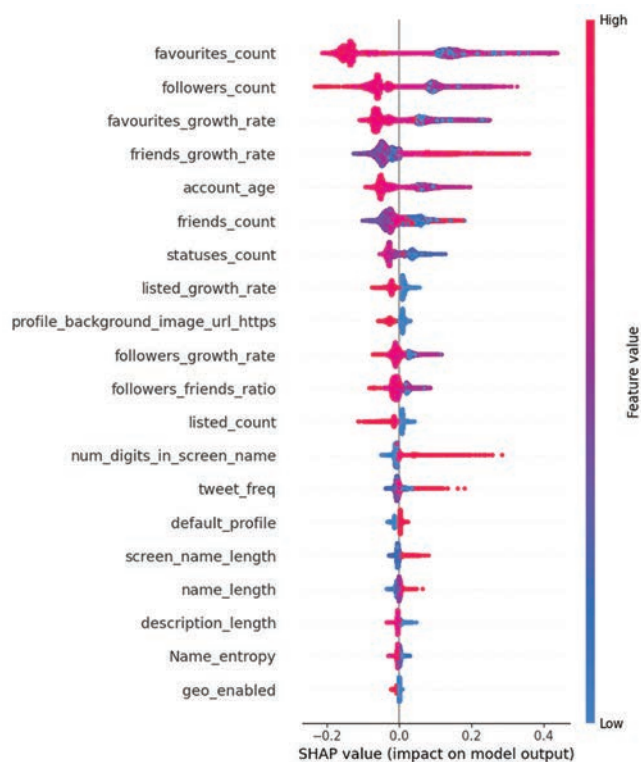
Fig. 2. The SHAP graph for our dataset. The graph shows the 20 features with the highest impact on the dataset. Using RF and Euclidean normalization.

plot, as shown on the right side of the Y-axis; when the value of features is large, the color moves to red, but when the value of the features is small, the color moves to blue. The plot shows that the original statistic features and statistically derived features are highly effective and ranked at the top of the graph. Furthermore, some features drive the model in one direction; for example, in the case of the num_digits_in_ screen_name feature, when the number of digits is increased, the result moves toward one (bot-like), and in the case of the listed_count feature, when the number rises the model drive toward zero (human-like). Fig. 2 shows the result of the SHAP plot, the order of the features may be slightly different than what we mentioned in the text because the SHAP plot only depends on one algorithm, but our experience is based on five algorithms.

C. Oversampling and Undersampling Evaluation

We use the synthetic minority oversampling technique, Tomek Link (SMOTETomek), and Edited Nearest Neighbor (SMOTEENN) to reshape the dataset size. The original dataset includes 58K of humans and 42K of bots; SMOTEENN reshapes the dataset to 54K for each human and bot, and SMOTETomek reshapes the dataset to 57K for each human and bot. Our experiment (Fig. 3) used Gaussian normalization and 10 best features based on feature selection experiment. In addition, an original dataset was used without any changes. As shown in Table VII, reshaping the dataset increases performance measurements positively. When SMOTETomek and SMOTEENN were applied, all ML
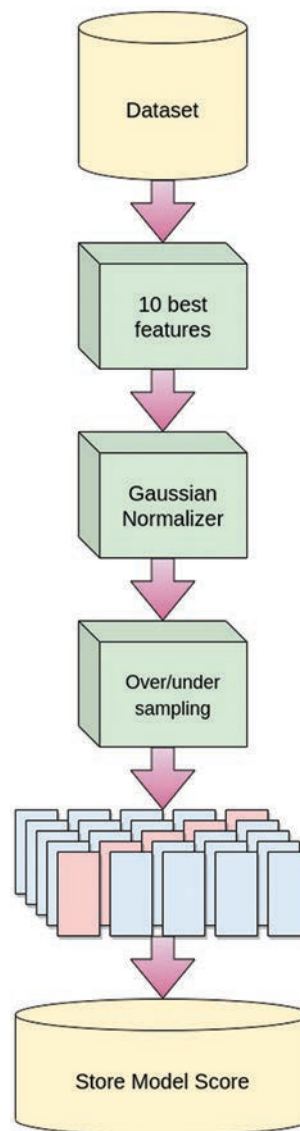

Fig. 3. The proposed experiments to determine the impact of over-/under-sampling algorithms on bot detection performance.

models' measurement performance increased (GN accuracy decreased with SMOTETomek as an exception). However, when we compare SMOTEENN and SMOTETomek, SMOTEENN performed much better than SMOTETomek. Furthermore, we used some oversampling algorithms such as SMOTE, ADASYN, and Borderline-SMOTE to reshape a dataset. Those algorithms increase the samples of lower class (oversampling) without performing any undersampling. The results are inferior, and usually, the original dataset performs better (their results are not mentioned in Table VII).

D. Generalization

Creating new types of bot with a different specialty are easy; because of that, many Botmasters develop new types of a bot that can avoid detection by the previous models. Therefore, a model's ability to distinguish between humans and bots outside their datasets (unseen data) is essential for any machine learning model, and this characteristic is
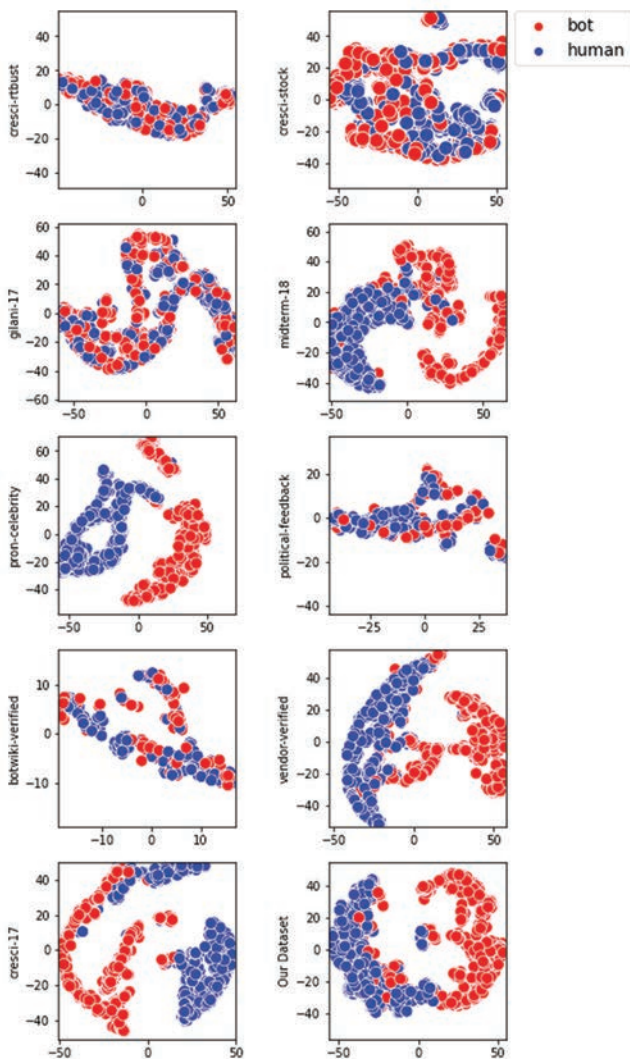
Fig. 4. The distribution of bots and humans for all the datasets using a
TSNE plot, 1000 samples for each class.

Fig. 5. The generalization experiments using one dataset for training and
other datasets for testing.

called generalization. We use nine datasets out of 11 from Yang, et al., 2019, work to test our models capability for generalization. The dataset's names are Cresci-Rtbust, Cresci-stock, Gilani-17, midterm-18, pron-celebrity, political-feedback, Botwiki-verified, vendor-verified, and Cresci-17. (Fig. 4 shows the distribution of each dataset). However, Varol-Icwsm and Caverlee datasets are excluded because their public version only contains id (unique id for an account) and the remaining features are missing.

We trained the model with one of the datasets (including ours) and tested the model with the remaining datasets separately. The 10 best score features are used in feature selection experiment, SMOTEENN for balancing dataset, and Gaussian normalization. (Fig. 5 shows the steps for the experiment). Table VIII shows five ML models achieved the highest ROC AUC when the models were trained on one dataset and tested with the remaining datasets.

The SVM model achieves the best AUC average, but when compared to the result of SVM and ET-500, the generalization of ET-500 is higher than SVM. As a result, Fig. 6 is a complete comparison of all datasets based on the ET-500
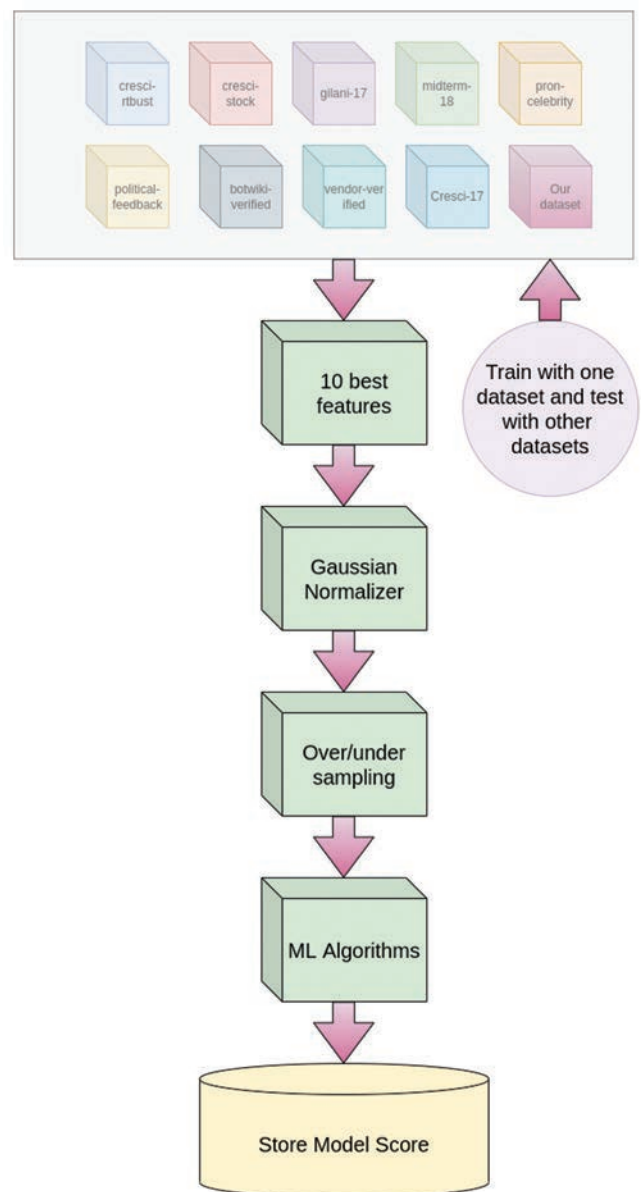
model. As shown in Fig. 6, our dataset achieved the best average accuracy among all datasets, 6% higher than others. Moreover, it achieved the best result for three datasets; the other two achieved good results, but the remaining datasets performed poorly. An average of 75% is not satisfying for a realistic bot detection model.

Nevertheless, two critical points explain why the results are inadequate. First, our models are trained with 10 features from metadata which is insufficient for a real-life situation to increase the performance. Other features such as tweet content need to be considered. Second, there are different types of bots with different parameters; for example, fake followers and spammer bots have quite unlike specialties. Because of that, we thought in those cases, when the model performs poorly, the test dataset contains bots that do not exist in a trained dataset or vice versa.
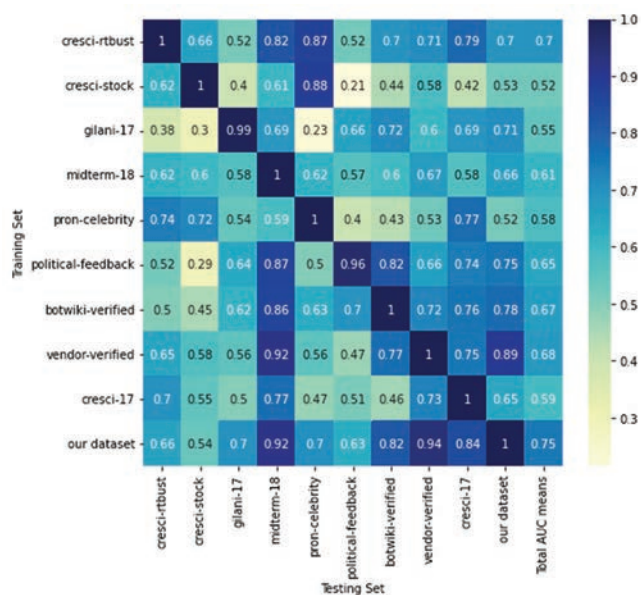
Fig. 6. AUC scores of ET-500 classifiers trained on one dataset and tested on others; the final columns also show each dataset's mean AUC.

As a result, we believe a dataset cannot achieve high performance for all other datasets unless to make the bot detection multi-class classification instead of a binary one. Instead of having one class for bots, we should create different bot classes such as spammer, fake follower, and others. Furthermore, it is worth mentioning that in our experiment, the result of the comparison between other datasets is much lower (besides our dataset) than the ones mentioned in Yang, et al., 2019, paper. We thought that it was because of a decrease in the number of features; the same experiment was repeated with the 20 features mentioned in Yang, et al., 2019, paper, which increased the average result by almost 2%, but still considerably less than the number in the original paper.

VI. Conclusion

This paper proposes novel benchmarks (datasets) with 100K samples based on 1.5M metadata collected from Twitter API. We collected data for more than 100 days to capture various bot types related to controversial topics like the US presidential election. Moreover, we use state-of-the-art online API to obtain the ground truth labels for the benchmark. The dataset includes 100K samples, and to the best of our knowledge, it is the largest only metadata dataset in this research area. Furthermore, this paper applied various autofeature selections and over-/under-sampling to the benchmark to increase the benchmark's generalization and scalability, reduce training time, and prevent over-/under-fitting while achieving very accurate results based on five-cross validation. As a result, our dataset achieved better AUC compared to other datasets by 6% in the case of generalization. Furthermore, applying the SMOTEENN technique achieved 2% higher results than the original dataset. In the future work, we intend to increase dataset

generalization by including additional features rather than metadata features. Since we thought that bot detection on Twitter is a multiclass nature problem, we planned to create a multiclass benchmark instead of a binary one. Dataset and experiment codes are available on this GitHub link to support overcoming the lack of datasets for the future research.

References

Adewole, K.S., Anuar, N.B., Kamsin, A., Varathan, K.D. and Razak, S.A., 2017. Malicious accounts: Dark of the social networks. *Journal of Network and Computer Applications*, 79, pp.41-67.

Alom, Z., Carminati, B. and Ferrari, E., 2018. Detecting spam accounts on Twitter. In: *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*. Institute of Electrical and Electronics Engineers Inc., Piscataway, New Jersey. pp.1191-1198.

Altmann, A., Toloşi, L., Sander, O. and Lengauer, T., 2010. Permutation importance: A corrected feature importance measure. *Bioinformatics*, 26(10), pp.1340-1347.

Brown, P.F., de Souza, P.V., Mercer, R.L., Della Pietra, V.J. and Lai, J.C., n.d. *Class-Based n-gram Models of Natural Language*. Computational linguistics, 18, pp.467–480.

Davis, C.A., Varol, O., Ferrara, E., Flammini, A. and Menczer, F., 2016. BotOrNot: A System to Evaluate Social Bots. In: *Proceedings of the 25th International Conference Companion on World Wide Web*. pp.14-16.

Derhab, A., Alawwad, R., Dehwah, K., Tariq, N., Khan, F.A. and Al-Muhtadi, J., 2021. Tweet-based bot detection using big data analytics. *IEEE Access*, 9, pp.65988-66005.

Elhassan, T. and Aljurf, M., 2016. Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. *Global J Technol Optim*, 1, pp.1-11.

Feng, S., Wan, H., Wang, N., Li, J. and Luo, M., 2021. TwiBot-20: A comprehensive twitter bot detection benchmark. *arXiv*, 2021, p.13088.

Ferrara, E., Varol, O., Davis, C., Menczer, F. and Flammini, A., 2016. The rise of social bots. *Communications of the ACM*, 59, pp.96-104.

Granitto, P.M., Furlanello, C., Biasioli, F. and Gasperi, F., 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83, pp.83-90.

Hanouna, S., Neu, O., Pardo, S., Tsur, O. and Zahavi, H., 2019. Sharp power in social media: Patterns from datasets across electoral campaigns. *Australian and New Zealand Journal of European Studies*, 11, pp.95-111.

Hayawi, K., Mathew, S., Venugopal, N., Masud, M.M. and Ho, P.H., 2022. DeeProBot: A hybrid deep neural network model for social bot detection based on user profile data. *Social Network Analysis and Mining*, 12, p.43.

Huang, J. and Ling, C.X., 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17, pp.299-310.

Jović, A., Brkić, K. and Bogunović, N., 2015. A review of feature selection methods with applications. In: 2015 *38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE. pp.1200-1205.

Khanday, A.M.U., Khan, Q.R. and Rabani, S.T., 2021. Identifying propaganda from online social networks during COVID-19 using machine learning techniques. *International Journal of Information Technology (Singapore)*, 13, pp.115-122.

Kudugunta, S. and Ferrara, E., 2018. Deep neural networks for bot detection. *Information Sciences*, 467, pp.312-322.

Martin-Gutierrez, D., Hernandez-Penaloza, G., Hernandez, A.B., Lozano-Diez, A. and Alvarez, F., 2021. A deep learning approach for robust detection of bots in

twitter using transformers. *IEEE Access*, 9, pp.54591-54601.

Orabi, M., Mouheb, D., Al Aghbari, Z. and Kamel, I., 2020a. Detection of bots in social media: A systematic review. *Information Processing and Management*, 57, p.102250.

Orabi, M., Mouheb, D., Al Aghbari, Z. and Kamel, I., 2020b. Detection of bots in social media: A systematic review. *Information Processing and Management*, 57, p.102250.

Peng, H., Long, F. and Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, pp.1226-1238.

Rodríguez-Ruiz, J., Mata-Sánchez, J.I., Monroy, R., Loyola-González, O. and López-Cuevas, A., 2020. A one-class classification approach for bot detection on Twitter. *Computers and Security*, 91, 101715.

Shannon, C.E. and Weaver, W., 1949. *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana, IL.

Shaw, R.G. and Mitchell-Olds, T., 1993. ANOVA for unbalanced data: An overview. *Ecology*, 74, pp.1638-1645.

Shevtsov, A., Tzagkarakis, C., Antonakaki, D. and Ioannidis, S., 2021. Identification of Twitter Bots Based on an Explainable Machine Learning Framework: The US 2020 Elections Case Study. *Proceedings of the International AAAI Conference on Web and Social Media*.

Shukla, H., Jagtap, N. and Patil, B., 2021. Enhanced twitter bot detection using ensemble machine learning. In: *Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021*. Institute of Electrical and Electronics Engineers Inc., Piscataway, New Jersey. pp.930-936.

Varol, O., Ferrara, E., Davis, C.A., Menczer, F. and Flammini, A., 2017. Online Human-bot Interactions: Detection, Estimation, and Characterization. In: *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, pp.280-289.

Wang, J., Xu, M., Wang, H. and Zhang, J., 2006. Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding. In: *2006 8th International Conference on Signal Processing*. IEEE.

Wilson, D.L., 1972. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, 3, pp.408-421.

Yang, K.C., Varol, O., Hui, P.M. and Menczer, F., 2019. Scalable and generalizable social bot detection through data selection. *arXiv*, 2019, p. 09179.

Train Support Vector Machine Using Fuzzy C-means Without a Priori Knowledge for Hyperspectral Image Content Classification

Akar H. Taher

Department of Software Engineering, Faculty of Engineering, Koya University,
Koya KOY45, Kurdistan region - F.R. Iraq

Abstract—In this paper, a new cooperative classification method called auto-train support vector machine (SVM) is proposed. This new method converts indirectly SVM to an unsupervised classification method. The main disadvantage of conventional SVM is that it needs a priori knowledge about the data to train it. To avoid using this knowledge that is strictly required to train SVM, in this cooperative method, the data, that is, hyperspectral images (HSIs), are first clustered using Fuzzy C-means (FCM); then, the created labels are used to train SVM. At this stage, the image content is classified using the auto-trained SVM. Using FCM, clustering reveals how strongly a pixel is assigned to a class thanks to the fuzzification process. This information leads to gaining two advantages, the first one is that no prior knowledge about the data (known labels) is needed and the second one is that the training data selection is not done randomly (the training data are selected according to their degree of membership to a class). The proposed method gives very promising results. The method is tested on two HSIs, which are Indian Pines and Pavia University. The results obtained have a very high accuracy of the classification and exceed the existing manually trained methods in the literature.

Index Terms—Automatic training, Clustering, Cooperative classification, Fuzzy C-means, Support Vector Machine.

I. Introduction

Nowadays, hyperspectral image (HSI) classification attracts the attention of researchers due to the rich information they contain. Moreover, this type of image can be used in many applications for the same reason. Among the applications of HSI, mining and geology (Goetz, et al., 1985), ecology (Ryan, et al., 2014), civil or military surveillance (Lagueux, et al., 2012), agriculture (Lacar, Lewis and Grierson, 2001), medicine (Akbari, et al., 2010), food safety and quality (Feng and Sun, 2012), and teledetection (Tarabalka, et al., 2010; Cariou, Moan and Chehdi, 2020; Alameddine, Chehdi and

Cariou, 2021; Cariou, Le Moan and Chehdi, 2022; Dong, et al., 2022; Sellami and Tabbone, 2022; Sun, et al., 2022) can be listed. The problem with most of the methods used for HSI classification is that they need *a priori* knowledge to train the classifier.

In the classification process, each pixel vector of the HSI must be given a distinct label. In the past two decades, a variety of pixel-wise (spectral-based) techniques has been used to solve this problem, including k-nearest neighbors (Bruzzone and Cossu, 2002), support vector machine (SVM) (Bruzzone and Cossu, 2002), and sparse representation (Chen, Nasrabadi and Tran, 2013). Among the vast number of classification methods, SVM has relatively demonstrated good performance for identifying high-dimensional data even when a small number of training samples are available (Camps-Valls and Bruzzone, 2005). In HSI classification, SVM can successfully overcome the Hughes phenomena (Hughes, 1968) and the problem of small training sample sizes. As a result, SVM and its enhanced algorithms perform better than other approaches. However, the problem with these approaches is that they strictly need previously labeled data to train the SVM. These labeled data are not available in all cases and not an easy task to obtain.

In Guo, et al., 2019, the SVM is used with a guided filter, in which two fusion methods are used to combine spectral and special features. In Shang, et al., 2022, another SVM-based method is proposed, it also contains a step of filtering. The problem with these methods is that they use a filter that may cause information loss, and it needs some parameters to be fixed for the filtering process. In Pathak and Kalita, 2019, another spectral-spatial SVM-based classification is presented, in this methods, a sliding window of fixed size is used to extract the spatial feature; however, the size of the window may affect the efficiency of the method. In Li, Li and Pan, 2019, SVM is combined with deep learning, and the results obtained using this method are very interesting. The disadvantage of using deep learning is that it dramatically increases the number of features, leading to a very high computation time. Many other SVM-based methods are proposed in the literature to classify the contents of HSI (Tarabalka, et al., 2010; Awad and Khanna, 2015; Wu, et al., 2016; Guo, et al., 2019; Li, Li and Pan, 2019; Pathak and

Kalita, 2019; Zhao, et al., 2020; Mounika, et al., 2021; Ren, et al., 2021; Shang, et al., 2022). The problem of all the above-mentioned methods is: (1) They need labeled data *(a priori* knowledge) to train the SVM and (2) the training data are selected randomly. To this end, the main contribution of this study and the proposed method is to overcome these two disadvantages, without losing the classification rate accuracy.

The remaining of this paper is structured as follows: after this introductory section, section II explains the theory of both SVM and FCM, in section III the proposed method is introduced, then the section gives details about the datasets used in this research to validate the proposed method. The final part of this same section gives the results obtained using this newly proposed method, the method is called: Auto Train Support Vector Machine (ATSVM). Further, the results obtained using this new technique are compared to those obtained using classic SVM only and other SVM-based methods found in the literature, which are all trained with Ground Truth labels. In section IV the conclusions are given.

II. Basic Elements

To better understand this approach, its' basic elements, that is, FCM and SVM, are explained in detail below:

A. Fuzzy C-means (FCM)

FCM (Bezdek, Ehrlich and Full, 1984) is a fuzzified version of K-means clustering (MacQueen, 1967) which adds a fuzzification operation, this fuzzification gives the ability to solve more complex clustering problems (Li, et al., 2009). More clearly, this method assigns a class membership to a data point, which depends on the distance of the data point to a particular class compared to all other classes. FCM tries to minimize the objective function below:

$$J = \sum_{j=1}^{NC}\sum_{i=1}^{N} u_{ij}^{m}\left(F_i - g(c_j)\right)^2 \tag{1}$$

with the constraint:

$$\sum_{j=1}^{NC} u_{ij}^{m} = 1 \quad \forall i \tag{2}$$

Where:
- NC is the number of classes,
- N is the number of pixels in the dataset (image),
- F_i is the vector of Nf features representing the pixel x_i,
- $g(C_j)$ is the center of gravity of class C_j,
- $m \in [1,\infty]$: is the fuzzification factor,
- and u_{ij} represents the entry (i, j) of the partition matrix, with $0 \le u_{ij} \le 1$.

By giving high membership values to data points near the center of their clusters, the objective function is minimized; in contrast, low membership values are provided to data points far from the class center. The following equations update the class centers and membership functions:

$$g(c_j) = \sum_{i=1}^{N} \frac{u_{ij}^{m}}{\sum_{k=1}^{N} u_{ik}^{m}} F_i \tag{3}$$

$$u_{ij} = \frac{\left\|F_i - g(c_j)\right\|^{\frac{1}{m-1}}}{\left\|\sum_{j=1}^{NC} F_i - g(c_j)\right\|^{\frac{1}{m-1}}} \tag{4}$$

The four steps of FCM can be centered on the following phases:

Step 1: Set the membership matrix $U = \left[u_{ij}\right]$, $1 \le j \le NC$, $1 \le i \le N$ randomly between 0 and 1 and satisfy the condition in Equation (2).

Step 2: Find cluster (class) centers $g(C_j)$ using Equation (3).

Step 3: Update the degree of membership u_{ji} using Equation (4).

Step 4: Redo *steps 2* and *3* until the algorithm converges (negligible difference between the current membership matrix and the previous membership matrix or the maximum number of iterations).

B. SVM

SVM is a powerful classification method that was initially developed by Boser, Guyon, and Vapnik (Boser, Guyon and Vapnik, 1992; Vapnik, 1995). On the other hand, SVMs are a collection of connected supervised techniques used for regression and classification (Gove and Faytong, 2012). More clearly, SVM is a classification and regression prediction method that automatically avoids over-fitting while maximizing predictive accuracy using machine learning theory. SVMs can be characterized as systems that use a high-dimensional feature space as the hypothesis space for a linear function that is trained using an optimization theory-based learning technique which incorporates a learning bias.

A classification process is composed of training and testing data that consist of some data samples (Duda and Hart, 1973). Each sample of data in the training group contains one target value and several features. SVM aims to create a model which predicts the target value of data samples in the testing group in which they contain the features only (Cristianini and Shawe-Taylor, 2000). Being a supervised approach, SVM relies on known labels to determine whether the system is operating properly. This information gives the desired response, validating the accuracy and the efficiency of the system. The first step in SVM classification involves determining whether characteristics have a close relationship to the recognized classes. This is referred to as feature selection or extraction. Feature selection and SVM classification together can be deployed to identify important elements that are involved in whatever processes recognize the classes or not (Cristianini and Shawe-Taylor, 2000).

III. PROPOSED METHOD AND DATASETS USED FOR VALIDATION

A. FCM and SVM in Cooperation

In this article, a cooperative approach that combines FCM and SVM is proposed in which the data are first clustered through using FCM, then the obtained class labels are used to train the SVM (see the diagram in Fig. 1).

In this proposed approach, after clustering the datasets using FCM, the obtained labels are used to train SVM instead of using known labels coming with datasets (known labels are not available in all cases). The choice of these two methods to cooperate (i.e., FCM and SVM) is not done arbitrarily. First, the reason behind choosing FCM is that the fuzzy decision gives very important information about the classification of each data point (pixel) in the image, that is, the degree of membership of each pixel to the specific clusters. Thereafter, this information is used to choose the pixels which are used to train the SVM. Second, the reason behind choosing SVM is that this method has shown very promising results in the classification of high-dimensional data and HSIs as we mentioned before in the introductory section.

B. Datasets

To validate the results of the proposed method, the Pavia University and the Indian Pines HSIs are used, as they are very well known and widely used HSIs.

Pavia University dataset

The Reflective Optics System Imaging Spectrometer (ROSIS-03) optical sensor was used to capture this image of an urban area. According to the specs, the ROSIS-03 sensor captures 115 bands with a spectral coverage of 0.43–0.86 μm. Each pixel has a spatial resolution of 1.3 m. The test site was close to the University of Pavia's Engineering School in Pavia, Italy. The pixels are 610 by 340. Due to noise, 12 channels were eliminated. Processing was done on the remaining 103 spectral channels. Nine classes of interest were considered: Tree, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow, and soil (Fig. 2) (Tarabalka, et al., 2010).

Indian Pines dataset

This HSI was captured by the Airborne Visible/Infrared Imaging Spectrometer sensor over the agricultural Indian Pine test site in Northwest Indiana in the USA. The spatial dimension of the image is 145 by 145 pixels and has a 20 m per pixel spatial resolution. The spectral dimension is 224 components. The number of components is reduced to 200 by removing components covering the region of water
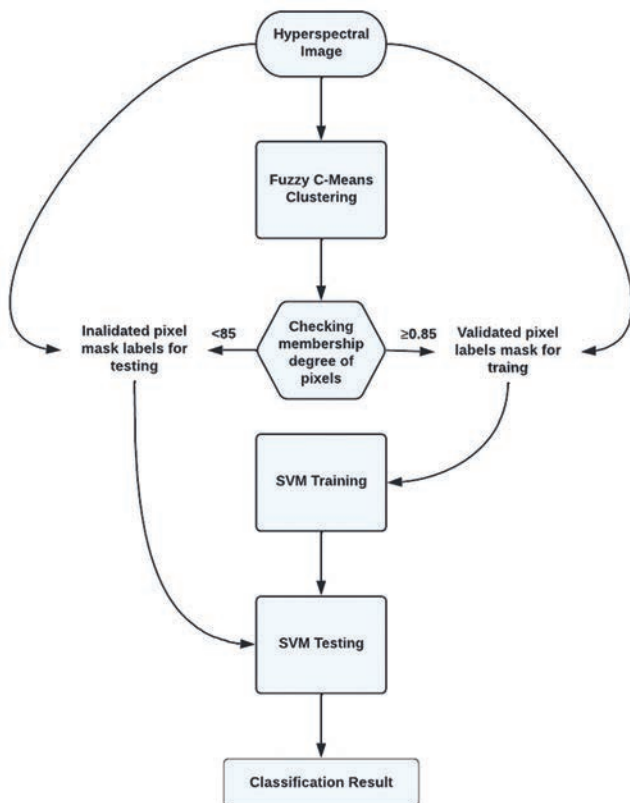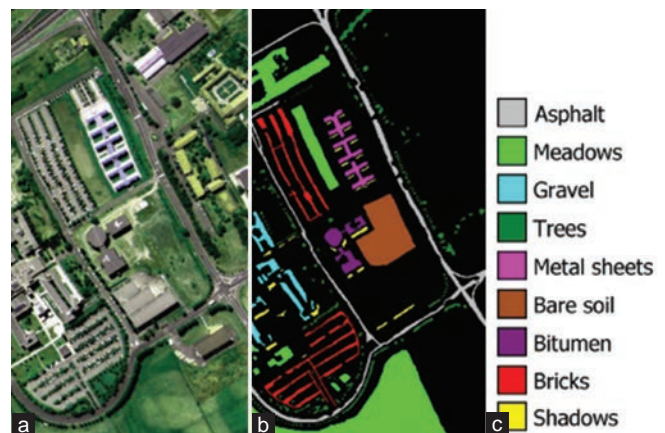

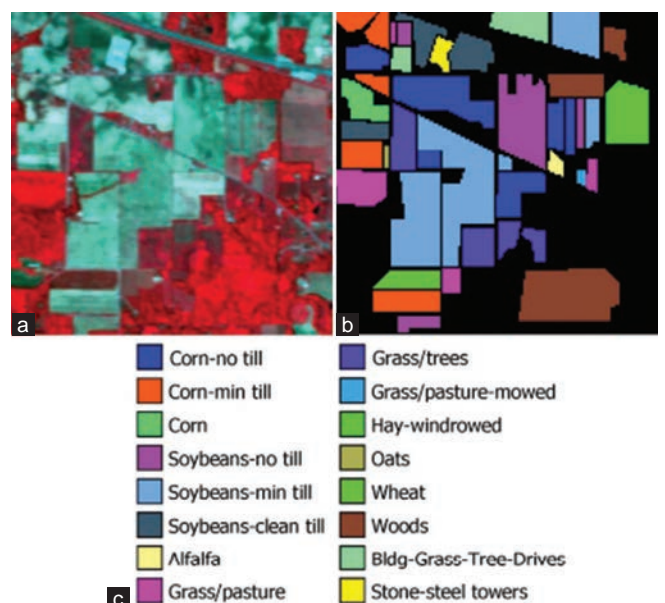Fig. 2. Pavia University image. (a) Three-band color composites. (b) Ground truth. (c) Color code and class names.

Fig. 3. Indian Pines image. (a) Three-band color composites. (b) Ground truth. (c) Color code and class names.

Fig. 1. Proposed method (ATSVM) flow diagram.

absorption: [104–108], [150–163], and 220. Sixteen classes of interest were considered: Alfalfa, Corn-notill, Corn-mintill, Corn, Grass-pasture, Grass-trees, Grass-pasture-mowed, Hay-windrowed, Oats, Soybean-notill, Soybean-mintill, Soybean-clean, Wheat, Woods, Buildings-Grass-Trees-Drives, and Stone-Steel-Towers (Fig. 3) (Tarabalka, et al., 2010).

C. Results

Algorithms used and fixing their parameters

To test the proposed approach, the FCM and SVM algorithms of MATLAB™ version 2021 are used. The fixed parameter for each algorithm is given in Tables I and II.

TABLE I
FCM FIXED PARAMETERS

FCM
Distance: Euclidian
Number of iterations: 100
Fuzzification parameter (m): 2
Tolerance: 10e-5
Number of classes: 16 for Indian Pines and 9 for Pavia University
Validation index: 0.85 (new parameter not exiting in standard FCM, added by the proposed method)

TABLE II
SVM FIXED PARAMETERS

SVM
Model type:
o Preset: Medium Gaussian SVM
o Kernel function: Gaussian
o Box constraint level: 1
o Multiclass method: One-versus-One
o Standardize data: true
Optimizer options:
o Hyperparameter options disabled
Feature selection: Disabled

TABLE III
FCM CLASSIFICATION EXAMPLE (U)

	Class 1	Class 2	Class 3
Pixel 1	0.9	0.03	0.07
Pixel 2	0.8	0.09	0.11
Pixel 3	0.02	0.88	0.1

Validation process

To validate the proposed method (ATSVM), the method is applied to the previously presented datasets. First, the data are organized in a matrix of (SxF) format (samples in rows and features in columns). Then, data are clustered with the FCM algorithm. The output of FCM is a fuzzy decision for each data sample (pixels in the HSI). In the proposed method, a validation parameter index is introduced which is fixed to 0.85. This parameter is used to select the data points used for training the SVM algorithm (the data points with membership degree ≥0.85 are chosen for training). More clearly, this index is compared to the degree of membership of each pixel after the FCM clustering, for example, if the whole data are clustered to three classes using FCM, each pixel will have a degree of membership to the three class and the summation of all the degrees of a pixel is equal to one (Equation 2 and Table III). The reason behind using the pixels with a high degree of membership to a class is that in the case classification, confidence is high, for more clarification, the example in Table III. In this case, Pixels 1 and 3 are chosen for training SVM, but Pixel 2 is not chosen as its greatest degree of membership is smaller than the fixed threshold (0.85).

After choosing the pixels with high confidence of classification, their classification result is defuzzied. This is done by giving the label of where the membership degree is the largest, by this step, each pixel will get a label that indicates the class they belong to. In the example in Table III, Pixel 1 will get Label 1 and Pixels 3 will get Label 2, as the maximum membership degree is in Classes 1 and 2, respectively. These labels are used to train the SVM. Normally, these labels need to be known *a priori*, but in this proposed method, they are created by the FCM algorithm to train the SVM. At this point, the SVM algorithm is trained using these created labels. The pixels that are not used for training (as Pixel 2 in Table III) are used for testing.

The results of ATSVM for Pavia University and Indian Pines are shown in Table IV and it is observed that the proposed methods have the correct classification rate equal to 0% for some classes. This is because these classes contain a little number of samples and the SVM classier is not trained sufficiently by this number of samples. This problem is no unique, and it is repeated in other methods found in the literature (Tables V and VI). To show the efficiency of the

TABLE IV
CONFUSION MATRIX OF THE CLASS-SPECIFIC ACCURACY (CSA) FOR PAVIA UNIVERSITY HIS USING ATSVM

Class	Asphalt	Meadows	Gravel	Trees	Painted metal sheets	Bare Soil	Bitumen	Self-blocking bricks	Shadows
Asphalt	8134	34	0	0	0	13	0	17	259
Meadows	123	8783	0	61	46	0	1	10	95
Gravel	0	0	981	0	0	0	0	0	0
Trees	0	0	202	6790	160	0	0	22	0
Painted metal sheets	0	0	154	0	963	0	0	0	0
Bare soil	0	0	0	0	0	949	0	0	0
Bitumen	0	0	1	0	0	0	3110	0	0
Self-blocking bricks	0	0	0	3	3	1	0	7184	0
Shadows	0	11	0	0	0	0	165	0	4501
CSA%	96.18068	96.31539	100	94.64734	86.21307	100	99.96786	99.90266	96.2369

TABLE V
COMPARISON OF ATSVM APPLIED ON PAVIA UNIVERSITY HSI WITH OTHER METHODS FOUND IN THE LITERATURE

Class names	ATSVM	SVM	SVM+Watershed	SVM+PartClust	SVM-HSEG	GS-SVM
Asphalt	96.18	84.93	93.64	90.72	94.77	93.50
Meadows	96.31	70.79	75.09	92.73	89.32	95.50
Gravel	100	67.16	66.12	82.09	96.14	86.00
Trees	94.64	97.77	98.56	99.21	98.08	97.50
Painted metal sheets	86.21	99.46	99,91	100	99.82	99.50
Bare soil	100	92.83	97.35	96.78	99.76	98.08
Bitumen	99.96	90.42	96.23	92.46	100	99.00
Self-blocking bricks	99.9	92.78	97.92	97.8	99.29	93.50
Shadows	96.23	98.11	96.98	97.74	96.48	98.38
OA%	96.77	81.01	85.42	93.59	93.85	96.06

TABLE VI
COMPARISON OF ATSVM APPLIED ON INDIAN PINES HSI WITH OTHER METHODS FOUND IN THE LITERATURE

Class names	ATSVM	SVM	SVM+ISODATA	SVM+EM	SVM-MSF	SVM-MSF+MV	GA-SVM
Alfalfa	0	74.4	12	0	94.9	94.9	88.96
Corn-notill	100	78.2	79.32	71.65	91	93.2	75.83
Corn-mintill	100	69.6	84.95	84.15	95.7	96.6	68.84
Corn	70.46	91.9	75.83	60.66	95.7	95.7	82.51
Grass-pasture	100	92.2	93.75	93.97	94.6	94.6	87.54
Grass-trees	100	91.7	94.8	99.11	92.4	97.3	90.68
Grass-pasture-mowed	0	100	91.67	93.97	100	100	71.43
Hay-windrowed	100	97.7	97.51	99.09	99.8	99.8	99.58
Oats	100	100	16.67	0	100	100	87.51
Soybean-notill	100	82	83.85	82.02	98	93.9	72.84
Soybean-mintill	0	58	93.16	95.05	82	82	74.52
Soybean-clean	83,14	87.9	85.17	90.05	86	97.2	79.08
Wheat	99.51	98.8	93.19	98.95	99.4	99.4	96.08
Woods	99.52	93	97.17	95.36	97.6	99.7	92.41
Buildings-grass-trees-drives	97.92	61.5	79.53	69.3	68.8	68.8	85.78
Stone-steel-towers	0	97.8	86.05	86.05	95.6	95.6	89.13
OA%	96.62	78.2	88.53	87.25	88.4	91.8	82.83

TABLE VII
CONFUSION MATRIX OF THE CLASS-SPECIFIC ACCURACY (CSA) FOR INDIAN PINES HSI USING ATSVM

Class names	Alfalfa	Corn-notill	Corn-mintill	Corn	Grass-pasture	Grass-trees	Grass-pasture-mowed	Hay-windrowed	Oats	Soybean-notill	Soybean-mintill	Soybean-clean	Wheat	Woods	Buildings-grass-trees-drives	Stone-steel-towers
Alfalfa	0	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corn-notill	0	1428	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corn-mintill	0	0	830	0	0	0	0	0	0	0	0	0	0	0	0	0
Corn	0	0	0	167	70	0	0	0	0	0	0	0	0	0	0	0
Grass-pasture	0	0	0	0	483	0	0	0	0	0	0	0	0	0	0	0
Grass-trees	0	0	0	0	0	730	0	0	0	0	0	0	0	0	0	0
Grass-pasture-mowed	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0
Hay-windrowed	0	0	0	0	0	0	0	478	0	0	0	0	0	0	0	0
Oats	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0
Soybean-notill	0	0	0	0	0	0	0	0	0	972	0	0	0	0	0	0
Soybean-mintill	0	0	0	0	0	0	0	0	0	0	2455	0	0	0	0	0
Soybean-clean	0	0	0	0	0	0	0	0	0	0	94	499	0	0	0	0
Wheat	0	0	0	0	0	0	0	0	0	0	0	1	204	0	0	0
Woods	0	0	0	0	0	0	0	0	0	0	0	0	0	1259	6	0
Buildings-grass-trees-drives	0	0	0	0	0	0	0	0	0	0	6	0	0	2	378	0
Stone-steel-towers	0	0	0	0	0	0	0	0	0	0	93	0	0	0	0	0
CSA %	0	100	100	70.46	100	100	0	100	100	100	100	84.14	99.51	99.52	97.92	0

proposed method, the obtained results are compared with other SVM-based methods proposed in the literature (Tarabalka, et al., 2008; Tarabalka, Benediktsson and Chanussot, 2009; Fauvel, et al., 2013; Zhao, et al., 2020). It is important to mention, all these methods (unlike ATSVM) require previously known labels to train the SVM. The comparing results for the test images are shown in Tables V-VII, respectively, with overall accuracy (OA) of 96.77% for Pavia University and 96.62% for Indian Pines HSIs.

It is observed that the proposed methods have the correct classification rate equal to 0% for some classes. This is because these classes contain a little number of samples and the SVM classier is not trained sufficiently by this number of samples. This problem is no unique and it is repeated in other methods found in the literature (Tables V and VI). To show the efficiency of the proposed method, the obtained results are compared with other SVM-based methods proposed in the literature (Tarabalka, et al., 2008; Tarabalka, Benediktsson and Chanussot, 2009; Fauvel, et al., 2013; Zhao, et al., 2020). It is important to mention, all these methods (unlike ATSVM) require previously known labels to train the SVM. The comparing results for the test images are shown in Tables V and VI.

IV. Conclusion

A new unsupervised SVM-based clustering method is proposed. It can be concluded from the obtained results that the proposed method (ATSVM) is giving excellent results. In addition, the advantage of the proposed method is that it does not need *a priori* knowledge to train the SVM (no previously known labels are needed). Further, using FCM enables the choice of the train data instead of choosing them randomly. The method works well on big classes and it is less efficient on smaller classes. The problem of low accuracy in small classes is not unique to our proposed method (detailed results are shown in the previous section). FCM clustering needs the number of clusters to be defined. The future work will be, first improve the FCM algorithm to detect the correct number of clusters automatically and second calculate the introduced fixed threshold automatically.

References

Akbari, H., Kosugi, Y., Kojima, K. and Tanaka, N., 2010. Detection and analysis of the intestinal ischemia using visible and invisible hyperspectral imaging. *IEEE Transactions on Biomedical Engineering*, 57, pp.2011-2017.

Alameddine, J., Chehdi, K. and Cariou, C., 2021. Hierarchical unsupervised partitioning of large size data and its application to hyperspectral images. *Remote Sensing*, 13, p.4874.

Awad, M. and Khanna, R., 2015. *Support Vector Machines for Classification*. Springer, Berlin. pp.39-66.

Bezdek, J.C., Ehrlich, R. and Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. *Computers and Geosciences*, 10, pp.191-203.

Boser, B.E., Guyon, I.M. and Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: *Proceedings of the 5th Annual Workshop on Computational Learning Theory, COLT' 92*. Association for Computing Machinery, New York, USA. pp.144-152.

Bruzzone, L. and Cossu, R., 2002. A multiple-cascade-classifier system for a robust and partially unsupervised updating of land-cover maps. *IEEE Transactions on Geoscience and Remote Sensing*, 40, pp.1984-1996.

Camps-Valls, G. and Bruzzone, L., 2005. Kernel-based methods for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 43, pp.1351-1362.

Cariou, C., Le Moan, S. and Chehdi, K., 2022. A novel mean-shift algorithm for data clustering. *IEEE Access*, 10, pp.14575-14585.

Cariou, C., Moan, S. and Chehdi, K., 2020. Improving K-nearest neighbor approaches for density-based pixel clustering in hyperspectral remote sensing images. *Remote Sensing*, 12, p.3745.

Chen, Y., Nasrabadi, N.M. and Tran, T.D., 2013. Hyperspectral image classification via kernel sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 51, pp.217-231.

Cristianini, N. and Shawe-Taylor, J., 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge.

Dong, Y., Liu, Q., Du, B. and Zhang, L., 2022. Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification. *IEEE Transactions on Image Processing*, 31, pp.1559-1572.

Duda, R.O. and Hart, P.E., 1973. *Pattern Classification and Scene Analysis*. 1st ed. Wiley, New York.

Fauvel, M., Tarabalka, Y., Benediktsson, J.A., Chanussot, J. and Tilton, J.C., 2013. Advances in spectral-spatial classification of hyperspectral images. *Proceedings of the IEEE*, 101, pp.652-675.

Feng, Y.Z. and Sun, D.W., 2012. Application of hyperspectral imaging in food safety inspection and control: A review. *Critical Reviews in Food Science and Nutrition*, 52, pp.1039-1058.

Goetz, A.F.H., Vane, G., Solomon, J.E. and Rock, B.N., 1985. Imaging spectrometry for earth remote sensing. *Science*, 228, pp.1147-1153.

Gove, R. and Faytong, J., 2012. Chapter 4 machine learning and event-based software testing: Classifiers for identifying infeasible Gui event sequences. In: Hurson, A. and Memon, A. Eds. *Advances in Computers*. Elsevier, Amsterdam, Netherlands. pp.109-135.

Guo, Y., Yin, X., Zhao, X., Yang, D. and Bai, Y., 2019. Hyperspectral image classification with SVM and guided filter. *EURASIP Journal on Wireless Communications and Networking*, 2019, p.56.

Hughes, G., 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14, pp.55-63.

Lacar, F.M., Lewis, M.M. and Grierson, I.T., 2001. Use of hyperspectral imagery for mapping grape varieties in the Barossa Valley, South Australia. In: *Geoscience and Remote Sensing Symposium, 2001. IGARSS '01. IEEE 2001 International*. pp.2875-2877.

Lagueux, P., Puckrin, E., Turcotte, C.S., Gagnon, M.A., Bastedo, J., Farley, V. and Chamberland, M., 2012. Airborne Infrared Hyperspectral Imager for Intelligence, Surveillance and Reconnaissance Applications. *Proceedings of SPIE the International Society for Optical Engineering*.

Li, X., Lim, B.S., Zhou, J.H., Huang, S., Phua, S.J., Shaw, K.C. and Er, M.J., 2009. Fuzzy Neural Network Modelling for Tool Wear Estimation in Dry Milling Operation. *Annual Conference of the PHM Society*. Available from: http://www.papers.phmsociety.org/index.php/phmconf/article/view/1403 [Last accessed 2022 Jul 02].

Li, Y., Li, J. and Pan, J.S., 2019. Hyperspectral image recognition using SVM combined deep learning. *Journal of Internet Technology*, 20, pp.851-859.

MacQueen, J., 1967. Some Methods for Classification and Analysis of Multivariate Observations. Vol. 5. In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*.

Mounika, K., Aravind, K., Yamini, M., Navyasri, P., Dash, S. and

Suryanarayana, V., 2021. Hyperspectral Image Classification Using SVM with PCA. In: *2021 6ᵗʰ International Conference on Signal Processing, Computing and Control*. pp.470-475.

Pathak, D.K. and Kalita, S.K., 2019. Spectral Spatial Feature Based Classification of Hyperspectral Image Using Support Vector Machine. In: *2019 6ᵗʰ International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, Noida, India. pp.430-435.

Ren, J., Wang, R., Liu, G., Wang, Y. and Wu, W., 2021. An SVM-based nested sliding window approach for spectral spatial classification of hyperspectral images. *Remote Sensing*, 13, p.114.

Ryan, J.P., Davis, C.O., Tufillaro, N.B., Kudela, R.M. and Gao, B.C., 2014. Application of the hyperspectral imager for the coastal ocean to phytoplankton ecology studies in Monterey Bay, CA, USA. *Remote Sens*, 6, pp.1007-1025.

Sellami, A. and Tabbone, S., 2022. Deep neural networks-based relevant latent representation learning for hyperspectral image classification. *Pattern Recognition*, 121, p.108224.

Shang, Y., Zheng, X., Li, J., Liu, D. and Wang, P., 2022. A comparative analysis of swarm intelligence and evolutionary algorithms for feature selection in SVM-based hyperspectral image classification. *Remote Sensing*, 14, p.3019.

Sun, L., Zhao, G., Zheng, Y. and Wu, Z., 2022. Spectral spatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60, pp.1-14.

Tarabalka, Y., Fauvel, M., Chanussot, J. and Benediktsson, J.A., 2010. SVM and MRF-based method for accurate classification of hyperspectral images. IEEE Geoscience and Remote Sensing Letters, 7, pp.736-740.

Tarabalka, Y., Benediktsson, J. and Chanussot, J., 2009. Classification of Hyperspectral Data Using Support Vector Machines and Adaptive Neighborhoods. *International Conference on Image Processing*.

Tarabalka, Y., Chanussot, J., Benediktsson, J., Angulo, J. and Fauvel, M., 2008. Segmentation and Classification of Hyperspectral Data Using Watershed. *IGARSS 2020 Proceedings*, pp.652-655.

Vapnik, V.N., 1995. *The Nature of Statistical Learning Theory*. Springer, New York.

Wu, Y., Yang, X., Plaza, A., Qiao, F., Gao, L. and Cui, Y., 2016. Approximate computing of remotely sensed data: SVM hyperspectral image classification as a case study. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9, pp.1-13.

Zhao, C., Zhao, H., Wang, G. and Chen, H., 2020. Improvement SVM classification performance of hyperspectral image using chaotic sequences in artificial bee colony. *IEEE Access*, 8, pp.73947-73956.

An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method

Haval A. Ahmed[1], Peshawa J. Muhammad Ali[1], Abdulbasit K. Faeq[1], and Saman M. Abdullah[1,2]

[1]Department of Software Engineering, Faculty of Engineering, Koya University,
Koya KOY45, Kurdistan Region, F.R. Iraq

[2]Department of Computer Engineering, Faculty of Engineering, Tishk International University,
Erbil, Kurdistan Region, F.R. Iraq

Abstract—Data normalization can be useful in eliminating the effect of inconsistent ranges in some machine learning (ML) techniques and in speeding up the optimization process in others. Many studies apply different methods of data normalization with an aim to reduce or eliminate the impact of data variance on the accuracy rate of ML-based models. However, the significance of this impact aligning with the mathematical concept of the ML algorithms still needs more investigation and tests. To identify that, this work proposes an investigation methodology involving three different ML algorithms, which are support vector machine (SVM), artificial neural network (ANN), and Euclidean-based K-nearest neighbor (E-KNN). Throughout this work, five different datasets have been utilized, and each has been taken from different application fields with different statistical properties. Although there are many data normalization methods available, this work focuses on the min-max method, because it actively eliminates the effect of inconsistent ranges of the datasets. Moreover, other factors that are challenging the process of min-max normalization, such as including or excluding outliers or the least significant feature, have also been considered in this work. The finding of this work shows that each ML technique responds differently to the min-max normalization. The performance of SVM models has been improved, while no significant improvement happened to the performance of ANN models. It is been concluded that the performance of E-KNN models may improve or degrade with the min-max normalization, and it depends on the statistical properties of the dataset.

Index Terms—Min-max normalization, Support vector machine, Artificial neural network, Euclidean-based K-nearest neighbor, Mean squared error.

I. Introduction

Min-max data normalization is one of the data scaling methods that cast data in a specific range of [0,1] or [−1,1]. The main aim of such scaling is improving the performance of machine learning (ML) techniques. Min-max normalization is mainly used to speed up the convergence of some techniques utilizing gradient descent algorithm for convergence, like in artificial neural networks (ANN), and to eliminate the domination of some features over others in the techniques using distance measures like Euclidean K-nearest neighbor (E-KNN). Supposing that a dataset may contain different feature ranges, distance measures like Euclidean may assign more weight to features with larger ranges than those with small ranges. Therefore, min-max data normalization is used to equalize the weight of these features and make them have the same effect on the decision-making process. Nevertheless, there are still arguments among researchers about the exact impact of the min-max data normalization on the performance of supervised ML techniques, some of them look at normalization as a necessary step in machine learning process, while other see it as unnecessary step. The causes of this difference may include the quality of the datasets, the nature of the dataset, the application field, or to the machine learning technique itself and how it deals with the data. On the other hand, there are many different techniques of data normalization, they may respond differently to the different ML techniques. Min-max is one of the most used techniques with different ranged attribute datasets. It's easy to implement and has an approved effects on the performance of the models. This study aims to reveal the ambiguity of the real effect of the min-max data normalization and more specifically to investigate its impacts on the regression performance of ML models. Therefore, the main questions that addressed by this work is why ML techniques show disparity responds to the min-max normalization method.

The rest of this article is structured as follows: Section 2 presents the related works, whereas Section 3 is the methodology used in this research work consisting of six subsequent stages: Selecting and downloading the datasets, implementing min-max feature normalization, implementing ML techniques on the datasets, removing outliers, feature selection, and removing outliers with feature selection. Section 4 summarizes all the observed results, and Section 5 discusses the observed results. Section 6 concludes this research work.

II. Related Works

The majority of the works in the literature that have investigated the effect of the min-max data normalization reported a positive impact of the min-max normalization on the adopted ML techniques in their studies (Dadzie and Kwakye, 2021; Shahriyari, 2017), while some other studies determined that its usefulness varies from good to bad depending on the nature of the datasets and the ML model (Ambarwari, Adrian, and Herdiyeni, 2020) (Ahsan, et al., 2021). On the other hand, very limited studies concluded the degradation of the ML model accuracy with the present of min-max normalization. However, no comprehensive interpretation about their achieved results has been mentioned in those studies (Singh, et al., 2015).

In this work, we used most datasets from the most recognized and benchmarked dataset repository available online which is the University of California Irvine (UCI) Repository Dataset. We found that most of the researcher works in this area using datasets from the same repository (Ahsan, et al., 2021; Bhardwaj, Mishra, and Desikan, 2018; Dadzie and Kwakye, 2021; Jayalakshmi and Santhakumaran, 2011; Pires, et al., 2020; Shorman, et al., 2018). The research direction is progressed toward determining the effects of the min-max scaling regardless of the nature of the dataset application. Therefore, in this work, five different benchmarked datasets from the UCI repository have been adopted as well having different number of records and features. Because this work focuses on three ML techniques, which are SVM, ANN, and E-KNN, reviewing some relevant works on ML techniques, in general, and on these three techniques, more specifically, are needed.

One of the most common ML techniques is SVM. Many research works showed that SVM has more sensitive responds than ANN and E-KNN toward the normalization techniques. Results in many SVM-based works showed the usefulness of using min-max normalization with SVM (Dadzie and Kwakye, 2021; Shahriyari, 2017; Ambarwari, Adrian, and Herdiyeni, 2020). Despite that, there are still some studies proved that normalization has no effect or has very little effect on the accuracy rate of the SVM-based models (Singh, et al., 2015). Moreover, there are studies (Ahsan, et al., 2021) proved the degradation of the performance of the SVM-based models while attached to min-max normalization method.

On the other hand, research works commented differently on the suitability of using min-max method with ANN. Some works showed the negative effect of min-max normalization (Singh, et al., 2015), whereas others concluded that no significant performance improvements were observed (Jayalakshmi and Santhakumaran, 2011). No improvements were also observed for higher-order neural networks (Prasetyo, Setiawan, and Adji, 2020) and deep learning algorithms (Pires et al., 2020), while other research works reported an obvious improvement in the accuracy of the models (Ambarwari, Adrian, and Herdiyeni 2020).

The ambiguity is not only existed with ANN-based works, the uncertainty about the suitability of the min-max method and its impact on the accuracy of ML-based models also found in E-KNN-based works that used for classification or clustering. Some studies observed that attaching min-max with E-KNN technique will improve the accuracy rate within a very small range (Ambarwari, Adrian, and Herdiyeni 2020; Dadzie and Kwakye, 2021), however, other studies showed a significant improvement in the accuracy rate of the E-KNN-based models (Ahsan, et al., 2021; Rajeswari and Thangavel, 2020). There other research works stating the min-max normalization impact on the performance of E-KNN depends on the nature of the dataset, it may enhance or degrade the performance of E-KNN models (Muhammad Ali, 2022).

It becomes clear from reviewing the above relevant works that there is non-clear vision about the situation where the min-max normalization could be utilized with ML techniques for accuracy improvement. In other words, the question about the condition(s) or the circumstance(s) that make the min-max normalization responds a positive or negative effect on the performance of supervised learning models, needs be answered. To the best of our knowledge, we could not find a comprehensive study that tackles this problem.

Whereas the focus of this study is on evaluating the impact of the min-max normalization on the performance of some adopted regression models; it is necessary to review the validation measurements that have been used in the previous works. For regression models, the mean squared error (MSE) is the widely adopted performance measurement (Rajeswari and Thangavel, 2020; Singh, et al., 2015; Jayalakshmi and Santhakumaran, 2011; Shorman, et al., 2018; Bhardwaj, Mishra, and Desikan 2018). Moreover, the root mean square error (RMSE) (Prasetyo, Setiawan, and Adji, 2020), the coefficient of determination (R2) (Aksu, Güzeller, and Eser, 2019), and mean absolute error (MAE) were also used as performance measurements. On the other hand, the fitting time is an important performance measurement which is the time needed to fit the models (Shahriyari, 2017). The number of steps for convergence or number of iterations, epochs, and the complexity level of the model are other measures that could be used as performance measure. In this research work, performance measurement depends on the MSE.

Beside the mathematical process of ML techniques, data characteristics are another problem that challenging the performance of min-max method for accuracy impairment. These challenges are "outliers," "noise amplification," and "out of range data." The min-max normalization preserves the real relations among instances of the same feature, which makes it very sensitive to "outliers." Having anomalies in any feature forces the data to aggregate in a small range between 0 and 1, and leaves a wide range empty, this makes anomalies study a necessary step before implementing min-max normalization (Kappal, 2019). Moreover, the min-max eliminates the real differences among different features, this is called noise amplification, which is the enlargement of the small effect attributes and making them equal to the big effect attributes, which leads to a decrease in the accuracy of the models (Pires, et al., 2020). This challenge could be solved by removing the features with little significance among the input features of the models.

In this work, an outlier study and feature selection study were implemented on the original dataset. There are many methodologies available for outlier removing, we utilized interquartile range method. Literatures utilized different feature selection methods (Sattari, et al., 2021); this work used Pearson correlation method. Therefore, four different pairwise result sets are adopted: The pairwise result achieved using the original dataset, the pairwise result of the outlier clean dataset, the pairwise result after removing the least significant feature, and finally, the pairwise result set after implementing both outlier and features selection on the original dataset.

The scaling process of the min-max normalization depends on the features' minimum and maximum values, therefore, an "out of range" problem arises when the trained model is receiving a data instance outside the feature bounds, which leads to wrong predictions. The "out of range" challenge mostly happens with time series datasets. Studies suggested many solutions in this regard (Ogasawara, et al., 2010), widening the range between minimum and maximum values by a specific ratio of 20% is one of the suggested solutions. This free space above the max and below the min works as spare space in case such data appeared, but the time series data are beyond the scope of this article.

From the above literature review, we can conclude that there is a clear disagreement among researchers about the impacts of min-max data normalization on the regression performance of the ML models. Utilizing data normalization always enhances the performance of the ML model. In our opinion, this question needs more investigating before answering it. This research aims to reveal the ambiguity of the impact of the min-max data normalization on the regression performance of three ML algorithms SVM, ANN, and E-KNN by testing five different datasets and comparing the results with and without data normalization.

III. Methodology

In this research, three different ML techniques have been tested on five benchmark datasets to investigate the impact of min-max data normalization on the regression performance of these ML techniques. The methodology can be summarized as follows, Fig. 1.

Step 1: Adopting suitable datasets from the UCI website,

Step 2: Implementing min-max feature normalization, range [0,1],

Step 3: Applying and validating ML techniques on both original and the normalized datasets,

Step 4: Comparing the MSE results of the both cases to show the impact of min-max on the performance of the models,

Step 5: Implementing anomaly detection and removing outliers to the original datasets and then repeating Steps 2, 3, and 4,

Step 6: Implementing feature selection to the original datasets and then repeating Steps 2, 3, and 4

Step 7: Implementing anomaly detection and feature selection together to the original datasets and then repeating Steps 2, 3, and 4.
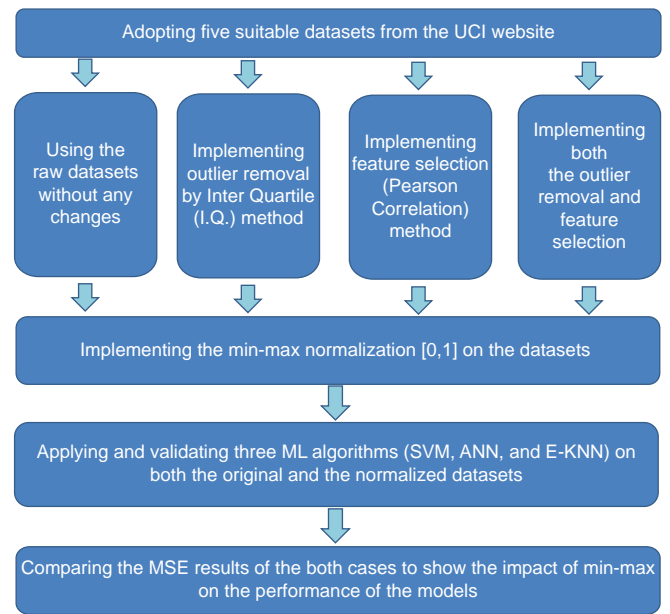

Fig. 1. The methodology of the research work.

A. The UCI datasets

Five different benchmark datasets were downloaded from the machine learning repository website of the University of California Irvine (UCI) (Dheeru and Graff, 2019). The reason behind selecting these datasets is the existing variation in the ranges of records among all attributes. Some of the datasets have big differences in their ranges like in the Airfoil Self-Noise Dataset (Table I) or very similar ranges like in power plant dataset (Table II). Such a variation is expected to have an impact on the regression performance of ML algorithms, and consequently, the impact of the min-max scaling method will appear more, which is the main target that this work aimed to investigate. The datasets belong to the real applications of physics, life sciences, engineering, and business, Table III. The dataset's statistical properties are shown in Tables I,II,IV-VI.

B. Implementing min-max feature normalization, range [0,1].

As shown in Equation (1), a normalized data sample x' could be obtained from the original data sample x. For an attribute, it is mostly dependent on instances with the maximum and minimum values in the same attribute. In this normalization method, the original data sample component values will be transformed to the range of [0,1].

$$x' = [(\frac{x - oldMin}{oldMax - oldMin}) * (newMax - newMin)] + newMin \quad (1)$$

Where:

x' is the normalized data sample,

x is the original data sample,

oldMin is the minimum data among any attribute of the original dataset,

oldMax is the maximum data among any attribute of the original dataset,

newMin is the minimum of the normalized dataset, and

newMax is the maximum of the normalized dataset.

TABLE I
STATISTICAL PROPERTIES OF THE AIRFOIL SELF-NOISE DATASET (BROOKS, POPE, AND MARCOLINI, 1989)

Feature	Maximum value	Minimum value	Range	Mean	Variance	SD
F1	20,000	200	19800	2886.380572	9,938,717.384	3152.573137
F2	22.2	0	22.2	6.782302063	35.0242405	5.918128125
F3	0.3048	0.0254	0.2794	0.136548237	0.008749868	0.093540728
F4	71.3	31.7	39.6	50.86074518	242.5116138	15.5727844
F5	0.0584113	0.000400682	0.058010618	0.01113988	0.000172929	0.013150234
Target	140.987	103.38	37.607	124.8359428	47.59146318	6.898656622

SD: Standard deviation

TABLE II
STATISTICAL PROPERTIES OF THE COMBINED CYCLE POWER PLANT DATASET (TÜFEKCI, 2014)

Feature	Maximum value	Minimum value	Range	Mean	Variance	SD
F1	37.11	1.81	35.3	19.65123119	55.53935724	7.45247323
F2	81.56	25.36	56.2	54.30580372	161.4905445	12.707893
F3	1033.3	992.89	40.41	1013.259078	35.2691519	5.938783706
F4	100.16	25.56	74.6	73.30897784	213.1678478	14.60026876
Target	495.76	420.26	75.5	454.3650094	291.2823183	17.066995

SD: Standard deviation

TABLE III
DATASETS PROPERTIES

Dataset	Number of features	Type of features	Number of instances	Supervised learning	Application
Airfoil Self-Noise Dataset (Brooks, Pope, and Marcolini, 1989)	5	Real	1503	Regression	Physics
Physicochemical Properties of Protein Tertiary Structure Dataset (Rana, 2013)	9	Real	9146 instances included	Regression	Life sciences
Combined Cycle Power Plant Dataset (Tüfekci, 2014)	4	Real	9568	Regression	Energy
Concrete Compressive Strength Dataset (Yeh, 1998)	8	Real	1030	Regression	Civil engineering
Real Estate Valuation Dataset (Yeh and Hsu, 2018)	6	Integer and Real	414	Regression	Business

TABLE IV
STATISTICAL PROPERTIES OF THE PHYSICOCHEMICAL PROPERTIES OF PROTEIN TERTIARY STRUCTURE DATASET (RANA, 2013)

Feature	Maximum value	Minimum value	Range	Mean	Variance	SD
F1	32,240.2	2783.15	29,457.05	9873.68162	16,094,604.52	4011.808135
F2	11,787.1	403.5	11,383.6	3016.435929	2,102,621.452	1450.041879
F3	0.56848	0.09362	0.47486	0.302155567	0.003941913	0.062784658
F4	343.239	10.6891	332.5499	103.4039974	3018.359088	54.9395949
F5	4,467,324.7	374,315.5155	4,093,009.223	1,369,092.965	311,794,123,479.2	558,385.2823
F6	470.897	33.6462	437.2508	145.5447009	4803.146285	69.30473494
F7	83,153.57	1108.9	82,044.67	3987.14593	3,536,332.356	1880.513854
F8	337	0	337	70.04286027	3192.870115	56.50548747
F9	47.4559	15.5049	31.951	34.48790348	35.17094729	5.930509868
Target	20.981	0	20.981	7.833154384	37.46611427	6.120956974

SD: Standard deviation

TABLE V
STATISTICAL PROPERTIES OF THE CONCRETE COMPRESSIVE STRENGTH DATASET (YEH, 1998)

Feature	Maximum value	Minimum value	Range	Mean	Variance	SD
F1	540	102	438	281.1656311	10,921.74265	104.5071416
F2	359.4	0	359.4	73.89548544	7444.083725	86.27910364
F3	200.1	0	200.1	54.18713592	4095.548093	63.99646938
F4	247	121.75	125.25	181.5663592	456.0602447	21.35556707
F5	32.2	0	32.2	6.20311165	35.6826025	5.973491651
F6	1145	801	344	972.9185922	6045.656228	77.75381809
F7	992.6	594	398.6	773.5788835	6428.099159	80.1754274
F8	365	1	364	45.66213592	3990.437729	63.16991158
Target	82.599225	2.331807832	80.26741697	35.81783583	279.0797167	16.70567917

SD: Standard deviation

TABLE VI
STATISTICAL PROPERTIES OF THE REAL ESTATE VALUATION DATASET (YEH AND HSU, 2018)

Feature	Maximum value	Minimum value	Range	Mean	Variance	SD
F1	2013.5833	2012.666667	0.9166666	2013.148953	0.079521365	0.281995327
F2	43.8	0	43.8	17.71256039	129.7887038	11.39248453
F3	6488.021	23.38284	6464.63816	1083.885689	1,592,920.631	1262.109595
F4	10	0	10	4.094202899	8.676334351	2.945561806
F5	25.01459	24.93207	0.08252	24.96903007	0.000154013	0.012410197
F6	121.56627	121.47353	0.09274	121.5333611	0.000235536	0.015347183
Target	117.5	7.6	109.9	37.98019324	185.1365075	13.6064877

SD: Standard deviation

C. Applying and validating ML techniques on the datasets

In this work, SVM, ANN, and E-KNN were adopted as regression methods. The target of applying all tests was to determine the impact of min-max data normalization on the regression performance of these techniques through implementing the technique on the original dataset and the normalized version individually then comparing their results. The reason for adopting SVM, ANN, and E-KNN techniques is the diversity of their nature. SVM utilizes Lagrange optimizer and tries to maximize the margins using different kernel functions whereas ANN utilizes a gradient descent algorithm as an optimization technique to minimize errors and to reach the goals faster. On the other hand, E-KNN utilizes Euclidean distance to determine the distance between the tested samples with the neighbors, where this distance is affected by the range of the values of the features. In all the tests, mean squared error (MSE) is computed as a performance measure and 10-fold as a validation technique.

For testing the SVM on all of the different datasets, four different kernel functions were considered (linear, Gaussian, radial base, and polynomial) on both the original and the normalized datasets. The MSE of the test set has been used for comparison purposes. The tests were designed such that the code runs on both the original datasets and the normalized dataset with the same parameters to determine the impact of the min-max normalization in MSE. The minimum MSE of the four tests was considered for comparison purposes with other techniques.

For each one of the datasets, 50 different models of ANN were tested. These models include either one or two hidden layers, with a different number of nodes in each hidden layer ranging between 3 and 100 nodes. In addition, different transfer functions were used in the hidden layers (sigmoid, tanh, and ReLU). The same model with the same parameters has been applied on both datasets (the original dataset and the normalized dataset), where the minimum MSE of the 50 tests was considered for comparison purposes with other techniques. It is known that ANN model results vary from one run to another, therefore, each one of the 50 different models is tested 10 times and the average is presented as the MSE result of the model within an acceptable standard deviation range. This is despite that each test of the 10 tests was validated by 10-fold validation.

A group of 33 models that used E-KNN (23 weighted-neighbor models and 10 traditional E-KNN models) was tested on the five datasets. The weighted group consisted of different models having five neighbors with different contribution weights ranging from 5% to 100%. The best weighted E-KNN model for all the datasets was 50%, 20%, 15%, 10%, and 5% from the closest neighbor to the farthest neighbor accordingly. The second group consisted of different E-KNN models considering up to 10 neighbors with equal contribution weights. The tests were implemented on both datasets, the original and the normalized. Minimum MSE among all the adopted experiments is observed to be used for comparison purposes.

D. Implementing anomaly detection and removing outliers

An outlier detection is conducted on the adopted five datasets by implementing two different methods, mean-standard deviation, and the interquartile method. The first method is considering the data lay outside the range (mean $\pm$ 3 $\times$ standard deviation) as an anomaly, and the second method is considering the data lay outside the range (Q1 - 3 $\times$ IQ, Q3 + 3 $\times$ IQ) as an anomaly, (Table VII).

In this research work, the wider range is considered for anomaly detection to decrease the number of anomalies, which is the interquartile method. Any data laid outside the wider range had been removed from the datasets. The same previous ML techniques were simply repeated, and the results were observed for comparison purposes.

E. Implementing feature selection

To determine the least significant feature among the features of each of the five datasets, this research adopts Pearson correlation. The correlation values are ranging between −1 and 1, the least significant feature has the correlation value between the feature and the target closer to zero (Table VIII).

F. Implementing anomaly detection and feature selection together

In this step, the combined effect of feature selection and outlier removal is investigated by implementing both procedures in the previous steps and then repeating the same above procedure. This step shows the impact of anomaly removal and the least significant feature removal together.

IV. RESULTS

The minimum MSE results using different machine learning techniques on the five datasets are shown in Table IX and Fig. 2.

After removing the outliers from the datasets, the MSE values changed in some of the datasets as one can observe in Table X and Fig. 3.

The same techniques with the same parameters were repeated after removing the least significant feature among all other features of each dataset. The results are shown in Table XI and Fig. 4.

Again, and after removing both the outliers and the least significant feature, the same techniques with the same parameters were repeated in Table XII and Fig. 5.

V. Discussion

The MSE values in the four (Tables IX-XII) are the minimum MSE observed after implementing a large number of various models and structures validated by 10-fold validation. For SVM, each number is the minimum of four models with different kernels (linear, Gaussian, radial base, and polynomial), which are applied to the original and the normalized data separately resulting in 8 models per case and 32 models for all of the cases for each dataset, 160 models for all the five datasets. Whereas for ANN, the parameters are the number of hidden layers (up to two layers), the number of nodes in each layer (up to 100 nodes), and the type of activation function (sigmoid, tanh-hyperbola, and ReLU). Totally, 50 models were applied to the original and the normalized data separately, resulting in 100 models for each case, and a total of 400 models per each one of the five datasets, 2000 ANN models were tested as a grand total. Finally, regarding the E-KNN pair 66 models were tested, the total is 264 for each pair for each dataset, the grand total is 1320 models which were tested.

For the SVM, it is clear that the adopted min-max normalization shows a positive impact on the results observed from all the datasets. There are differences in the ratio of the improvement, but we can see clearly that all the datasets responded positively to the min-max normalization with SVM (Table IX). This improvement in the results is reported in the literature as well (Ambarwari, Adrian, and Herdiyeni 2020; Dadzie and Kwakye, 2021; Shahriyari, 2017). The reason is due to the fact that the SVM does not have any tool to weigh one dimension versus other(s), but rather it focuses on optimizing the line, plane, or the hyperplane that separates the classes.

The results for ANN are not similar to the SVM, as there is no significant difference observed in comparing the pairwise results of the five datasets. The reason behind this is the existence of transfer functions (activation functions) in each hidden layer node of the neural network, where they may

TABLE VII
NUMBER OF OUTLIERS IN THE ADOPTED DATASETS USING (MEAN ± 3X STANDARD DEVIATION) AND (Q1 − 3 × IQ, Q3 + 3 × IQ) METHODS

Dataset	Number of outliers (mean ± 3 × SD)	Number of outliers (Q1 − 3 × IQ, Q3 + 3 × IQ)
Airfoil self-noise dataset	78	35
Physicochemical properties of protein tertiary structure dataset	806	213
Combined cycle power plant dataset	58	0
Concrete compressive strength dataset	49	33
Real estate valuation dataset	10	8

SD: Standard deviation

TABLE VIII
THE PEARSON CORRELATION VALUES FOR THE LEAST SIGNIFICANT FEATURES OF THE ADOPTED DATASETS WITH THE TARGET FEATURE

| Dataset	Least significant feature	Pearson correlation between the feature and the target
Airfoil self-noise dataset	F4	0.1251
Physicochemical properties of protein tertiary structure dataset	F7	−0.0018
Combined cycle power plant dataset	F4	0.3898
Concrete compressive strength dataset	f3	−0.1058
Real estate valuation dataset	F1	0.0875

TABLE IX
MINIMUM VALUES OF THE MEAN SQUARED ERROR RESULTS OF THE DIFFERENT MACHINE LEARNING MODELS
(ORIGINAL *VS*. MINIMUM-MAXIMUM NORMALIZED DATASETS)

Serial number	Machine learning technique	Airfoil self-noise dataset		Physicochemical dataset		Power plant dataset		Concrete strength dataset		Real estate valuation dataset	
		Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization
1	SVM	48	14	37	25	21	17	118	45	101	60
2	ANN	3	3	21	21	15	15	28	29	63	63
3	E-KNN	35	4	38	18	15	13	60	65	64	70

SVM: Support vector machine, ANN: Artificial neural network, E-KNN: Euclidean-based K-nearest neighbor

TABLE X
MINIMUM VALUES OF THE MEAN SQUARED ERROR RESULTS OF THE DIFFERENT MACHINE LEARNING MODELS (AFTER REMOVING OUTLIERS)

Serial number	Machine learning technique	Airfoil self-noise dataset		Physicochemical dataset		Power plant dataset		Concrete strength dataset		Real estate valuation dataset	
		Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization
1	SVM	45	13	37	25	21	17	89	34	97	60
2	ANN	3	2	20	20	15	15	31	31	63	66
3	E-KNN	34	4	38	18	15	13	61	52	66	70

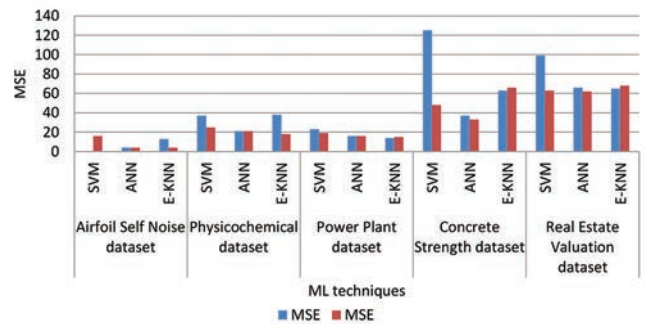SVM: Support vector machine, ANN: Artificial neural network, E-KNN: Euclidean-based K-nearest neighbor

TABLE XI
MINIMUM VALUES OF THE MEAN SQUARED ERROR RESULTS OF THE DIFFERENT MACHINE LEARNING MODELS (AFTER REMOVING THE LEAST SIGNIFICANT FEATURE)
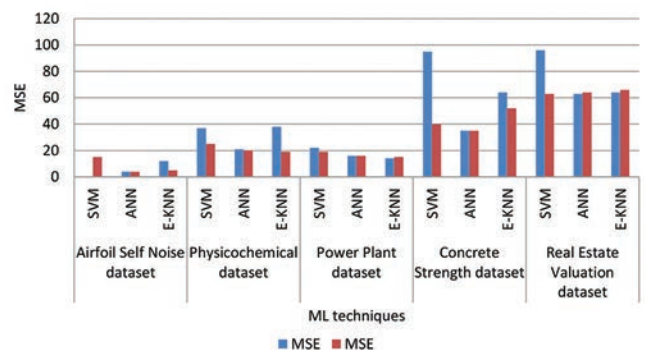
Serial number	Machine learning technique	Airfoil self-noise dataset		Physicochemical dataset		Power plant dataset		Concrete strength dataset		Real estate valuation dataset	
		Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization
1	SVM	28	16	37	25	23	19	125	48	99	63
2	ANN	4	4	21	21	16	16	37	33	66	62
3	E-KNN	13	4	38	18	14	15	63	66	65	68

SVM: Support vector machine, ANN: Artificial neural network, E-KNN: Euclidean-based K-nearest neighbor

TABLE XII
MINIMUM VALUES OF THE MEAN SQUARED ERROR RESULTS OF THE DIFFERENT MACHINE LEARNING MODELS
(AFTER REMOVING OUTLIERS AND LEAST SIGNIFICANT FEATURE)

Serial number	Machine learning technique	Airfoil self-noise dataset		Physicochemical dataset		Power plant dataset		Concrete strength dataset		Real estate valuation dataset	
		Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization
1	SVM	25	15	37	25	22	19	95	40	96	63
2	ANN	4	4	21	20	16	16	35	35	63	64
3	E-KNN	12	5	38	19	14	15	64	52	64	66

SVM: Support vector machine, ANN: Artificial neural network, E-KNN: Euclidean-based K-nearest neighbor
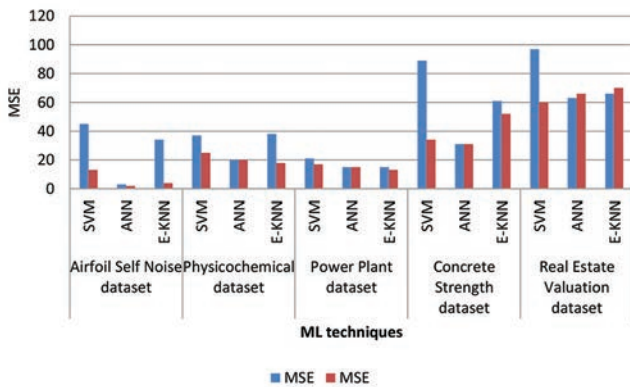
Fig. 2. Minimum values of the mean squared error results of the different machine learning models (original vs. min-max normalized datasets).
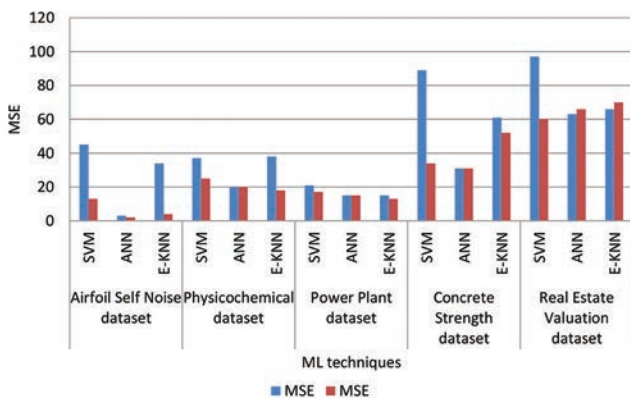
Fig. 3. Minimum values of the mean squared error results of the different machine learning models (after outlier removal).

Fig. 4. Minimum values of the mean squared error results of the different machine learning models (after removing the least significant feature).

Fig. 5. Minimum values of the mean squared error results of the different machine learning models (after removing outliers and least significant feature).

work like a normalization layer as well. The backpropagation procedure to adjust weights performs a denormalization process of the data; this is happening in each training cycle during the training session of the ANN models (Table IX).

For E-KNN, the results are different because the impact of min-max depends on the nature of the dataset itself. The MSE results of two of the datasets (Airfoil Self-Noise and Physicochemical) have been improved significantly, whereas no improvement happened to the power plant dataset after implementing the min-max data normalization, the reason could be because of the close ranges of the various dimensions in this dataset, while the MSE results of the two datasets concrete-strength and house-valuation datasets are degraded significantly. The reason behind the diversity of the effect of min-max data normalization on the MSE results of E-KNN returns to its dependence on the Euclidean distance

TABLE XIII
COMPARING THE RELIABILITY OF THIS RESEARCH WITH OTHER PREVIOUS LITERATURES

Serial number	Research work	Performance of SVM model	Performance of ANN model	Performance of E-KNN model	Number of datasets used	Type of application of the datasets
1	This research work	Improvement	No significant improvement observed	Depends on the nature of the dataset	5	5
2	Dadzie and Kwakye 2021	Improvement	N/A	No significant change	1	1
3	Shahriyari 2017	Improvement	Degradation	N/A	1	1
4	Ambarwari, et al., 2020	Improvement	Improvement	No significant change	1	1
5	Ahsan, et al., 2021	Degradation	N/A	Improvement	1	1
6	Singh, et al., 2015	No significant change	Degradation	N/A	1	1
7	Jayalakshmi and Santhakumaran, 2011	N/A	No significant change	N/A	1	1
8	Rajeswari and Thangavel, 2020	N/A	N/A	Improvement	5	1

SVM: Support vector machine, ANN: Artificial neural network, E-KNN: Euclidean-based K-nearest neighbor, N/A: Not available

which will be affected significantly by the normalization process in a positive or a negative way (Table IX).

After implementing the anomaly detection process to the five datasets and removing all the outliers according to the procedures mentioned in the methodology section of this article (Table VII), generally, and similar to the previous experiments, SVM has been improved by implementing min-max normalization, no significant improvement was observed for ANN, whereas the E-KNN still depends on the nature of the datasets. The normalized data of the concrete-strength dataset showed a little bit more positive response by implementing SVM and E-KNN to the anomaly clean datasets, whereas the ANN model remained stable. This is showing the computational power of ANN against anomalies, (Table X).

Furthermore, feature selection is implemented by removing the least significant feature among all the features of the dataset. A drawback of removing the least significant feature is not always significantly increasing the performance of the regression model, especially in the power plant dataset (Table VIII). In general, the pairwise comparisons of the results in Table XI show the same conclusions observed with the original datasets (with all features), SVM improvement, no significant change to the ANN, and E-KNN depending on the nature of the dataset. The normalized result of implementing E-KNN on the real estate dataset is improved slightly (Table XI).

The combined effect of implementing both the removal of anomalies and removal of the least significant feature is shown in Table XII. The results of the normalized dataset of implementing E-KNN to the concrete-strength and real estate datasets are improved (Table XII).

The results in Tables IX-XII showed that the E-KNN is sensitive to the nature of the data, therefore, it is better to check the performance of the E-KNN models with and without normalization, with and without outliers, with and without feature selection, and then to decide which one is the best, as there is no specific rule could be generalized for E-KNN.

The power plant dataset (Table II) is not impacted by the min-max normalization during this research, and it did not contain outliers. By an intensive look at the feature ranges, we can see that it has very similar feature ranges,

the standard deviation of the ranges is 15, which is a very small value compared to other datasets. This type of dataset is called a homogeneous dataset, it cannot be impacted by min-max data normalization.

Comparing the results observed in this research work with the results collected from reviewed literature showed that this research work implemented a better research methodology and analysis, also, the results are more precise and accurate (Table XIII). Most of the previous literature depended on testing only one dataset or in the best cases depended on using different datasets of one application field. This research work depended on five datasets belonging to five different real applications with different numbers of attributes and different attribute ranges. Therefore, the conclusions of this research are considered more reliable.

VI. CONCLUSION

The importance of implementing min-max data normalization mainly depends on the ML technique and the nature of the dataset. Experiments that are adopted in this work show that the min-max normalization is useful with SVM, whereas it makes no significant effect with ANN, which is because of the ANN's ability during the training stage to perform the normalization implicitly by itself. The powerful computation nature of ANN eliminates the effect of the min-max data normalization because it implicitly includes activation functions that work like the normalization layer. Even it eliminates the effect of outliers and least significant feature as well. Therefore, it is not important if min-max normalization is used with ANN or not. Depending on the nature of the dataset, min-max data normalization with E-KNN may result in performance improvement or degradation.

KNN uses distance measurements for determining the closest neighbors. In this research, Euclidean-based KNN is tested only. It has been concluded that using min-max data normalization may have a bad impact on the performance of the E-KNN, because it may eliminate the natural domination relations among the attributes and the target which leads to performance degradation, this has happened with two datasets concrete-strength and house-valuation. Therefore, it is better to test the E-KNN model with both the original dataset and

the normalized dataset before deciding if the min-max data normalization is useful or not.

E-KNN performance with min-max normalization may be improved by implementing outlier cleaning methods or by removing the least significant features, but this is not doing very well with SVM or ANN because of their powerful computation nature. No need to implement min-max data normalization to the homogeneous datasets whatever is the ML algorithm.

This research work has been compared to the literature in terms of the number of used datasets and their application fields. Unlike other works that adopt the use of a single dataset or datasets from one single application field, our research tried to be more reliable using five different datasets from five various application field.

REFERENCES

Ahsan, M., Mahmud, M.A., Saha, P.K., Gupta, K.D. and Siddique, Z. 2021. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, 9, p.52.

Aksu, G., Güzeller, C.G. and Eser, M.T. 2019. The effect of the normalization method used in different sample sizes on the success of artificial neural network model. *International Journal of Assessment Tools in Education*, 6, pp.170-92.

Ali, P.J.M., 2022. Investigating the Impact of min-max data normalization on the regression performance of K-nearest neighbor with different similarity measurements. *ARO The Scientific Journal of Koya University*, 10, p.10955.

Ambarwari, A., Adrian, Q.J. and Herdiyeni, Y. 2020. Analysis of the effect of data scaling on the performance of the machine learning algorithm for plant identification. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 4, pp.117-122.

Bhardwaj, C.A., Mishra, M. and Desikan, K. 2018. Dynamic Feature Scaling for K-Nearest Neighbor Algorithm.

Brooks, T.F., Pope, D.T. and Marcolini, M.A. 1989. *Airfoil Self-noise and Prediction (NASA Reference Publication)*. In: Technical Report 1218. National Aeronautics and Space Administration, United States.

Dadzie, E. and Kwakye, K. 2021. Developing a Machine Learning Algorithm-Based Classification Models for the Detection of High-Energy Gamma Particles.

Dheeru, D. and Graff, C. 2019. UCI Machine Learning Repository. School of Information and Computer Science. Vol. 25. University of California, Irvine, CA, p27.

Jayalakshmi, T. and Santhakumaran, A. 2011. Statistical normalization and back propagationfor classification. *Journal of Computer Theory and Engineering*, 3

pp.89-93.

Kappal, S. 2019. Data normalization using median median absolute deviation MMAD based Z-Score for robust predictions vs. Min-max normalization. *London Journal of Research in Science Natural and Formal*, 19, pp.39-44.

Ogasawara, E., Martinez, L.V., De Oliveira, D., Zimbrão, G., Pappa, G.L. and Mattoso, M. 2010. Adaptive Normalization: A novel Data Normalization Approach for Non-stationary Time Series. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp.1-8.

Pires, I.M., Hussain, F., Garcia, N.M., Lameski, P. and Zdravevski, E. 2020. Homogeneous data normalization and deep learning: A case study in human activity classification. *Future Internet*, 12, pp.194.

Prasetyo, J., Setiawan, N.A. and Adji, T.B. 2020. Improving normalization method of higher-order neural network in the forecasting of oil production. In: EDP Sciences.

Rajeswari, D. and Thangavel, K., 2020. The performance of data normalization techniques on heart disease datasets. *International Journal of Advanced Research in Engineering and Technology*, 11, pp.2350-2357.

Rana, P.S. 2013. Physicochemical properties of protein tertiary structure data set. UCI Machine Learning Repository, pp. Available from: https://www.archive. ics. uci. edu/ml/datasets/Physicochemical+Properties+ of+Protein+ Tertiary+ Structure. [Last accessed 2022 Apr 01].

Sattari, M.A., Roshani, G.H., Hanus, R., Nazemi, E., 2021. Applicability of time-domain feature extraction methods and artificial intelligence in two-phase flow meters based on gamma-ray absorption technique. *Measurement*, 168, p.108474.

Shahriyari, L. 2017. Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. *Brief Bioinformation*, 20, pp.985-94.

Shorman, A.R., Faris, H., Castillo, P.A., Merelo, J.J. and Al-Madi, N. 2018. The Influence of Input Data Standardization Methods on the Prediction Accuracy of Genetic Programming Generated Classifiers. *IJCCI 2018-Proceedings of the 10th International Joint Conference on Computational Intelligence,* pp.79-85.

Singh, B.K., Raipur, N.I.T., Verma, K. and Thoke, A.S. 2015. Investigations on impact of feature normalization techniques on classifier's performance in breast tumor classification. *International Journal of Computer Applications*, 116, pp.11-15.

Tüfekci, P. 2014. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power and Energy Systems*, 60, pp.126-40.

Yeh, I.C. 1998. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28, pp.1797-1808.

Yeh, I.C. and Hsu, T.K. 2018. Building real estate valuation models with comparative approach through case-based reasoning. *Applied Soft Computing*, 65, pp.260-271.

A Computational Model for Temperature Monitoring During Human Liver Treatment by Nd: YaG Laser Interstitial Thermal Therapy (LITT)

Bazhdar N. Mohammed and Dilshad S. Ismael

Department of Physics, College of Science, Salahaddin University-Erbil,
Erbil 44001, Kurdistan Region - F.R. Iraq

Abstract—Describing heat transfer in biological organs is absolutely challenging because it is involved with many complex phenomena. Therefore, understanding the optical and thermal properties of living system during external irradiation sources such as laser interstitial thermal therapy (LITT) are too important for therapeutic purposes, especially for hyperthermia treatments. The purpose of this study was to determine a proper laser power and irradiation time for LITT applicator to irradiate liver tissue during hyperthermia treatment. For this aim, bioheat equation in one-dimensional spherical coordinate is solved by Green function method to simulate temperature distribution and rate of damage around irradiated target and how thermal and optical properties such as laser power, laser exposure time, and blood perfusion rate affect the rate of temperature distribution. Guiding equations according to the suggested boundary conditions are written and solved by MATLAB software. The outcomes show that increasing laser exposure time and power increase the temperature, especially at the nearest distance from the center of diffusion. Accordingly, a decrease in blood perfusion rate leads to decrease temperature distribution. The findings show that the model is useful to help the physicians to monitor the amount of heat diffusion by laser power during the treatment to protect healthy cells.

Index Terms—Bioheat equation, Computational simulation, Green function, Hyperthermia, Laser interstitial thermal therapy.

I. Introduction

Thermal therapies are considered minimally invasive techniques that are used to kill cancer cells with few or no harm to healthy tissue when part of the living body, for example, brain, prostate, or liver is exposed by external sources such as laser interstitial thermal therapy (LITT) (Chen, et al., 2021), radiofrequency (Wang, et al., 2015), or microwave ablation (Chu and Dupuy, 2014) to rise temperatures up to normal value. One of the approaches that can be used in non-invasive treatments is hyperthermia. It is a treatment mechanism among thermal therapies for treating normal and cancerous tissues by elevation the core temperature to nearly 50°C and below 40°C no measurable effects are observed (Niemz, 2019, Adeleh, et al., 2021). However, such treatments are done when conventional surgery either has less probability to success or too risky for the patient (Blauth, et al., 2020, Dutta and Kundu, 2018).

Among this clinical thermal procedure, the application of LITT is a promising technique that carries out under MRI or CT control to treat unresectable liver hepatic tumors and abnormalities. LITT is minimally invasive technique that performed by implanting a laser catheter in to the tumor cells due to hyperthermia and coagulative effects, Fig. 1. A specific beam of applied laser based on photon transport theory interacts with electric and magnetic properties of cell to increase the temperature at a sufficient level to destroy them (O'neal, Hirsch, and Halas, 2004, Andres, et al., 2020, Soares, et al., 2012). From the diffusion viewpoint, the transfer of thermal energy through living tissue depends on fundamental thermodynamic principles represented by bioheat transfer equations, which is very important to understand the process of temperature profiles (Ash, et al., 2017, Khaleel, Yahya, and Ibrahim, 2019).

As a matter of fact, therapy planning still is challenging problem and difficult issue because it is involving with many complicated thermal mechanisms that should be taken into account, such as metabolic heat generation, convection and blood perfusion rate, vascular structure, thermal conductivity, specific heat capacity of tissue, and biological conditions (Yassene and Verhey, 2005, Bhowmik, Singh, and Repaka, 2013) that are why there is not exit a unique model to give best accuracy. Moreover, the laser properties such as power and irradiation time are too significant which must be properly controlled to avoid unnecessary thermal damage of healthy cells and protect tissue from injuries by vaporization and carbonization. To handle these reasons, computational simulation can provide lots of potential benefits in clinical context to predict the degree of damage and temperature distribution around the catheter applicator that is absolutely necessary to understand the behavior
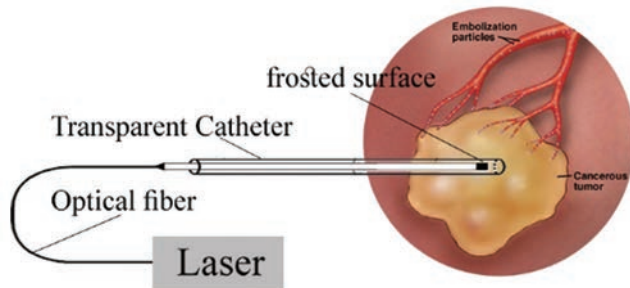
Fig. 1. Practical setup of laser interstitial thermal therapy for specific target. The optical fiber is sent to the liver tissue by means of specially designed (Niemz, 2019).

of increasing temperature in biological organs during the treatment to protect healthy cells.

Over the years, many governing mathematical equations have been proposed and modified since the middle of the last century when Penne's introduced his pioneering work about the distribution of temperature in the forearm (Pennes, 1948). A rate diffusion process model of laser-irradiated tissue has been proposed by Zhu, et al. (2002) to predict dynamic thermal response and the degree of injuries around the surface of applicator (Zhu, et al., 2002). In another work, the LITT model has been proposed to accurate modeling bioheat transfer mechanism and tissue response to thermal energy induced by different external thermal sources (Feng and Fuentes, 2011). Shibib, Munshid, and Lateef (2017) performed finite element method (FEM) to predict the physical properties of human liver tissue through the temperature distribution and measure the contagion damage volume by LITT (Shibib, Munshid, and Lateef, 2017). Andres, et al., have presented a numerical simulation to measure overestimated temperature until water vaporization in *ex vivo* porcine liver tissue. In Blauth, et al, 2020, a model simulated for LITT based on two approaches. In the first approach, radiation transport equation coupled with Pennes bioheat equation to estimate the temperature around coagulated region. In the second approach, LITT used for ablation purpose to transfer heat to *ex vivo* pig liver. A particular solution has been compared with nine test cases that have been measure using an optical fiber thermal diffuser applicator for delivering the rate of laser energy of neodymium-doped yttrium aluminum garnet (Nd: YaG) (Andres, et al., 2020). Furthermore, in Abbas, Hobiny, and Alzahrani, 2020, Giordano, Gutierrez, and Rinaldi, 2010, Li, Qin, and Tian, 2019, Özen, Helhel, and Cerezci, 2008, Skandalakis, et al., 2020, Vogl, et al., 2004, Zhou, Chen, and Zhang, 2007, various techniques have been used to predict the temperature distribution in biological tissue during the treatments.

In this study, a new method is approached to predict the effect of thermal and optical properties of LITT during liver treatment by hyperthermia. The parameters are conducted to Pennes bioheat equation and solved in one-dimensional spherical coordinate based on Green function to predict how liver tissue affected by laser power, exposure time, and blood perfusion rate as well as estimate the degree of damage produced by radiative LITT technique. For this purpose, a

Nd: YaG laser at a wavelength 1064 nm is used because of high penetration depth. In addition, the results compared with an existence computational model. The findings show that the current computational model is effective tool for predicting the degree of liver tissue damage through LITT power and helpful to control the amount of heat diffusion during the treatment.

II. Mathematical Formulation and Solution of the Problem

For the modeling, surface diffusion type of LITT applicator is considered to transfer thermal energy to the tissue (Niemz, 2019). Furthermore, to keep the variance of the temperature profiles by LITT, vascular structures of liver tissue are considered to be uniformly distributed and spherically diffused. The model summarizing in to two boundary parts: The first part located at the center of diffusing that corresponds to the interface between the laser applicator and tissue and the second part refers to ambient body temperature, Fig. 1. The transient problem is demonstrated based on bioheat equation which is given by Zhu, et al., 2002. For simplicity, it is being assuming tissue has constant thermal properties and the medium of heat diffusion is non-isotropic:

$$\rho c \frac{\partial T(r,t)}{\partial t} = \nabla \cdot \left(k \nabla T(r,t) \right) + \rho_b c_b \omega_b (T_c - T(r,t)) + Q_m + Q_{ext} ,$$

(1)

Where, t is t time and T is the temperature distribution, ρ and c are tissue density and specific heat capacity, respectively. k is tissue thermal conductivity, T_c is tissue surface temperature at the boundary of sphere. ρ_b is density of blood, c_b is blood-specific heat capacity, ω_b is blood perfusion rate term represents the effect of blood flow in vessels or microcapillaries, Q_m is a metabolic rate due to the basal metabolism, and Q_{ext} represents external heat generation.

To solve this, inhomogeneous partial differential operator a Green function is performed which belongs to the instantaneous point source (Faryad and Lakhtakia, 2018). Green's function operator is defined corresponding to the boundary conditions of the system to the action of laser beam pulse as an external heat source acting at an any distance (r) from the center of diffusion and at any time (t) (Hahn and Özisik, 2012). Now, Equation (1) can be rearranged and rewritten by introducing a new variable $\Theta = T-T_c$ and the linear differential operator L to give more suitable mathematical form (Hobiny et al., 2021):

$$\left(\frac{\partial}{\partial t} + \beta \nabla^2 + \alpha^2 \right) \theta_{(r,t)} = L_1 \left(\theta_{(r,t)} \right) = \frac{Q}{\rho c} ,$$

(2)

Where, $\alpha^2 = \frac{\omega_b c_b \rho_b}{\rho c}$, $\beta = \frac{k}{\rho c}$, and $Q = Q_m + Q_{ext}$. To find basic solution for this differential operator, we start from Green's characteristic equation:

$$L_1 \left(G_{(r,t|r',\tau)} \right) = \delta \left(r,r' \right) \delta \left(t,\tau \right) ,$$

(3)

Where, G is Green's function and δ is the Dirac delta function. If the domain unbounded the general representation of Green's function, the equation denoted by $V_{(r,t|r',\tau)}$. In this case, Equation (2) in a radial coordinate can be written as:

$$L_1(V) = \left(\frac{\partial}{\partial t} + \beta \frac{\partial^2}{\partial r^2} - \frac{2\beta}{r} \frac{\partial}{\partial r} + \alpha^2 \right) V = \frac{\delta(r-r')\delta(t-\tau)}{4\pi r^2}$$

(4)

The solution of this problem can be determined by the method of separation of variables and applying the inverse Fourier transform (the detail about solving steps can be found in Feng and Fuentes, 2011). Once the equation is solved, the principal solution part is:

$$V_{(r,t|r',\tau)} = \frac{2H_{(t-\tau)}e^{-\alpha^2(t-\tau)}}{\pi r r'} \int_{\lambda=0}^{\infty} e^{-\lambda\alpha^2(t-\tau)} \sin(\lambda r) \sin(\lambda r') d\lambda ,$$ (5)

$H_{(t-\tau)}$ is the Heaviside step function. Equation (5) is a closed form of the analytical solution. Now, the final radial solution can be expressed as the following:

$$V_{(r,t|r',\tau)} = \frac{H_{(t-\tau)}e^{-\alpha^2(t-\tau)}}{2\pi r'\sqrt{\beta\pi(t-\tau)}} [e^{\frac{-(r-r')^2}{4\beta(t-\tau)}} - e^{\frac{-(r+r')^2}{4\beta(t-\tau)}}]$$

(6)

Now, the Green's function solution equation (GFSE) should be introduced, which is a general expression that provides a non-linear temperature distribution in a medium based on suitable boundary conditions. Here, for the particular case of Equation 4, the GFSE for spherical coordinate is given by Hahn and Özisik, 2012:

$$\theta(r,t) = \int r'^p V_{(r,t|r',\tau)\tau=0} F(r') dr' + \frac{\beta}{k}$$

$$\int_{\tau=0}^{t} \int r'^p V_{(r,t|r',\tau)} Q(r,\tau) dr \, d\tau$$

(7)

Where, $V_{(r,t|r',\tau)}$ is closed form solution at $\tau = 0$, $F(r')$ represents initial conditions. r'^p is the Sturm–Liouville weighting function.

A. Boundary conditions

For the boundary conditions, it is supposed that the core and the surrounding temperature of spherical tissue are constant before applying external source. By the same fact, the boundary temperature at any region far from the core of the sphere is also constant.

$$T(r,0) = T_c, \quad (\partial T(r,0))/\partial t = 0,$$ (8)

$$[T(r,t)]_{r\circledR\yen} = T_s, \quad ([\partial T(r,0)] \_(r \circledR \yen))/\partial t = 0$$ (9)

B. Laser heat source

As infrared laser beam (Nd: YaG) interacts with the region of interest, the diffusion of heat around the area of interaction mainly depends on the absorption coefficient μ_a (Milanic, et al., 2019, Ash, et al., 2017):

$$Q_{ext}(r) = \frac{\mu_a(1-R_l)Pe^{-(\mu_a+(1-g)\mu_s)r}}{4\pi r^2},$$

(10)

Where, R_l is the light reflectivity of tissue depends on laser wavelength, P is a laser power releasing, it is energy at close g point to the center $r = r_0$ and spontaneously at t=0. μ_s is scattering coefficient, g is anisotropy factor its range between −1 and 1, in case of forward scattering g = 1, isotropic scattering g = 0, and g = −1 for backward scattering.

C. Mathematical formulation

The temperature distribution in one-dimensional spherical coordinate can be represented as the following:

$$\frac{\partial\theta_{(r,t)}}{\partial t} = \frac{\beta}{r^2} \frac{\partial}{\partial r} (r^2 \frac{\partial\theta_{(r,t)}}{\partial r}) - \alpha^2\theta_{(r,t)} + \frac{Q}{\rho c},$$

$$(t > 0 ; 0 < r < \infty)$$

(11)

then, boundary conditions have been given in Equations (8) and (9)

$$\theta(r,0) = T_c, \quad (\partial\theta(r,0))/\partial t = 0.$$ (12)

$$[\theta(r,t)]_{r\to\infty} = T_c, \quad ([\partial\theta(r,0)] \_(r \to \infty))/\partial t = 0.$$ (13)

the homogenous solution in terms of Green's function is expressed as:

$$\theta1(r,t) = \int r'^p V_{(r,t|r',\tau)\tau=0} F(r') dr'.$$ (14)

and the inhomogeneous solution of Equation (7) in terms of Green's function is found to be:

$$\theta2(r,t) = \frac{\beta}{k} \int_{\tau=0}^{t} \int r'^p V_{(r,t|r',\tau)} Q(r) dr \, d\tau.$$ (15)

the overall temperature distribution solution is obtained by adding homogenous and inhomogeneous solutions:

$$\theta(r,t) = \int r'^p V_{(r,t|r',\tau)\tau=0} F(r') dr'$$

$$+\frac{\beta}{k} \int_{\tau=0}^{t} \int r'^p V_{(r,t|r',\tau)} Q(r) dr \, d\tau.$$

(16)

by putting the boundary conditions, putting P = 2 for the spherical coordinate system (Beck, et al, 1992), $F(r') = 0$, and laser heat source equation. The Equation (16) reduced to:

$$\theta(r,t) = \frac{\beta}{k} \int_{\tau=0}^{t} \int_{r'=0}^{\infty} r'^2 V_{(r,t|r',\tau)}$$

$$(Qm + \frac{\mu_a(1-R_l)Pe^{-\mu_a r}}{4\pi r^2}) dr' d\tau,$$

(17)

The integral $\int_{r'=0}^{\infty} r'^2 V_{(r,t|r',\tau)} (\frac{\mu_a(1-R_l)Pe^{-\mu_a r}}{4\pi r^2}) dr'$

can be solved numerically according to the given laser source equation along with $V_{(r,t|r',\tau)}$ into the equation, the expression becomes:

$$\frac{\beta}{k}\int_{\tau=0}^{t}\int_{r'=0}^{\infty} r'^2 V_{(r,t|r',\tau)}(\frac{\mu_a(1-R_l)Pe^{-\mu_a r}}{4\pi r^2})dr'd\tau$$

$$=\frac{\mu_a(1-R_l)P}{8\pi krr_0}\int_{\tau=0}^{t}\frac{e^{-(a^2(t-\tau)+\mu_a r)}}{\sqrt{\beta\pi(t-\tau)}}[e^{\frac{-(r-r')^2}{4\beta(t-\tau)}}-e^{\frac{-(r+r')^2}{4\beta(t-\tau)}}]\ d\tau$$

$$,$$
$$(18)$$

On the other hand, the radial integral of $\int_{\tau=0}^{t}\int_{r'=0}^{\infty} r'^2 V_{(r,t|r',\tau)}Qm\,dr'd\tau$ represents Dirac delta function integral over the entire line which is equal to one (Modest, and Mazumder, 2021) and the time integral can be determined analytically. Finally, the solution reduces to:

$$\frac{\beta}{k}\int_{\tau=0}^{t}\int_{r'=0}^{\infty} r'^2 V_{(r,t|r',\tau)}Qm\,dr'd\tau = \frac{\beta Qm}{k\alpha^2}(1-e^{-(\alpha^2 t)}). \quad (19)$$

by putting the results of last two equations into Equation (17), the final solution of temperature distribution is $T(r,t)=Tc+\theta(r,t)$ which can be expressed as:

$$T(r,t)=Tc+\frac{\beta Qm}{k\alpha^2}\left(1-e^{-(\alpha^2 t)}\right)+\frac{\mu_a(1-R_l)P}{8\pi krr_0}$$

$$\int_{\tau=0}^{t}\frac{e^{-(a^2(t-\tau)+\mu_a r)}}{\sqrt{\beta\pi(t-\tau)}}\left[e^{\frac{-(r-r')^2}{4\beta(t-\tau)}}-e^{\frac{-(r+r')^2}{4\beta(t-\tau)}}\right]d\tau \quad (20)$$

$$.$$

III. Results and Discussion

In this section, the variation of temperature in liver tissue is estimated according to different LITT laser powers, duration times, and blood perfusion rates. To obtain the results, a computer program by MATLAB R2016a language is written. All the figures are drawn which are obtained from final solution of Equation 20, then integration over time is applied numerically by Simpson approximation rule after sorting temperature for each r at a fixed t. In addition, we considered $r_0=0.00075\,m$ to get the solution from the final equation. Regarding to a Nd: YaG laser properties, the value of R_l is 5%. The values of blood perfusion rate have been repossessed from Niemz, 2019, Adeleh, et al., 2021. It usually changes between 0.001 s^{-1} and 0.0005 s^{-1}. Furthermore, the diffusion domain is considered a sphere of radius (R) of 2 cm. Exemplary values of thermodynamic and optical properties have been retrieved from reference (Blauth, et al., 2020) and supposed to be constant throughout the domain, Table I.

A. The effect of irradiation time

Fig. 2 shows the effect of the different laser irradiation time (τ =5.10 and 15 s) on the temperature profiles along the radial distance (r) when the laser power and blood diffusion rate are constants. Whenever the laser source is switched

TABLE I
PHYSICAL AND OPTICAL PARAMETERS FOR LASER INTERSTITIAL THERMAL THERAPY IN LIVER TISSUE

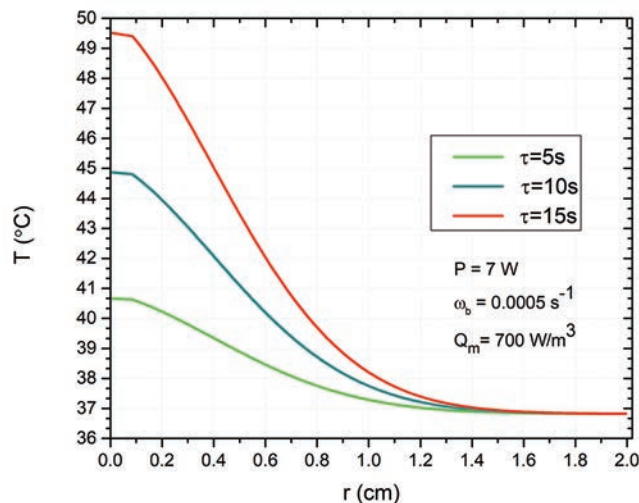| Parameter	Value
μ_a	50 m^{-1}
μ_s	8000 m^{-1}
G	0.97
ρ_b	1137 kg/m^3
C_b	3640 (J.kg)/k
K	0.518 W/(m.k)
P	1000 kg/m^3
C	4187 (J.kg)/k

Fig. 2. Transient temperature distribution as a function of radius in liver tissue for different values of laser irradiation time (laser on).

off, the maximum temperature is recorded at the center of the desired zone and then it decreases exponentially to the normal temperature T=36.8°C. Due to the long wavelength of used laser, which is located near-infrared region in the electromagnetic spectrum, an increase in time causes rise in temperature. It is clear that the longer exposure time produced a peak temperature just below 50°C which is the maximum limit of temperature to make hyperthermia whereas it is beyond that this value might cause reduction in enzyme activity resulting decrease to energy transfer between the cell and its surroundings, cell membrane disorder, and coagulation phenomena. As we can see even at maximum exposure time, the surrounding healthy cells are preserved.

The resulting of temperature change at 10 s laser intensity 120 Kw/m^2 is found to be very close to that achieved by Abbas, Hobiny, and Alzahrani, 2020. As can be noticed from Fig. 3, there is no much difference between temperatures in smaller radius whereas at 0.8–2 cm, the average change in temperature between this work and reference (Abbas, Hobiny, and Alzahrani, 2020) is only 1°C.

Alternately, at a given radius, the relation between the laser exposure time and temperature distribution is linear (Fig. 4). On the other side, when exposure time remains constant at 25 s, the variation of temperature between r = 0.3 cm and r
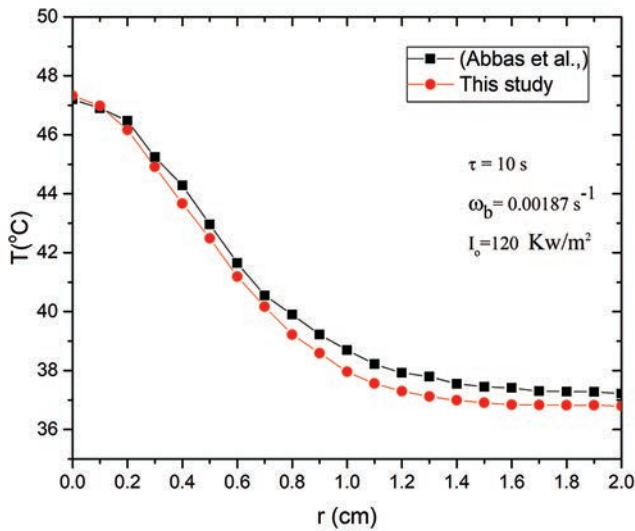
Fig. 3. Comparison between the centerline temperature obtained by this model and reference (Abbas, Hobiny, and Alzahrani, 2020) at constant laser intensity and exposure time.
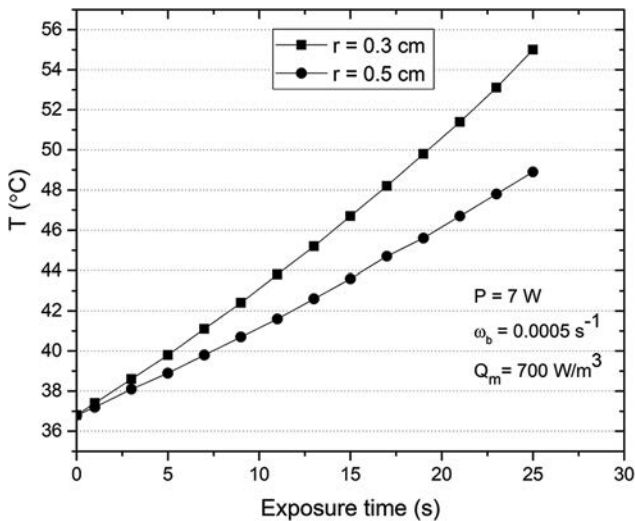

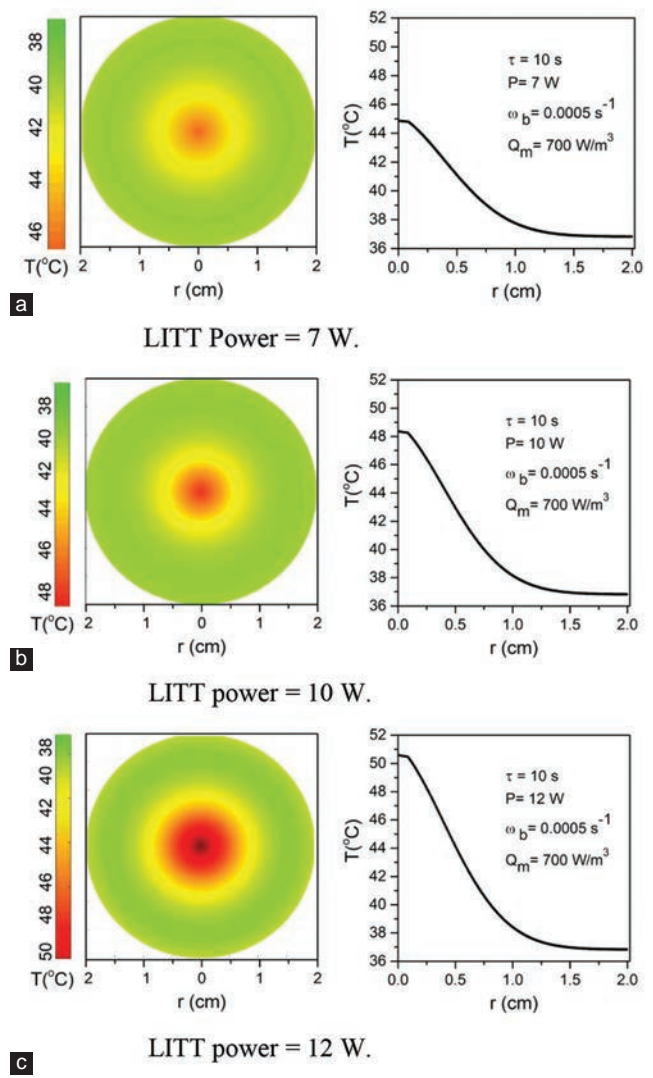Fig. 4. Relation between temperature achieved and exposure time for two different distances.

Fig. 5. Two-dimensional temperature contours and temperature profile as a function of distance for different laser powers. (a) and (b) achieved temperature change by laser interstitial thermal therapy with P = 7 and 10 W do not exceed hyperthermia conditions and (c) P=12 W more heat is generated and absorbed in the tissue which beyond hyperthermia condition.

= 0.5 cm is about 6°C as well as temperature still remains in hyperthermia range for smaller radius; meanwhile, it reaches the coagulation disorders in greater radius that necrotizing the healthy cells.

B. The effect of laser power

Fig. 5 illustrates the graphical representation of the temperature variation from the center of the desired target to sphere boarder. The tissue is irradiated by three different laser powers with beam radius 0.5 mm in constant exposure time and blood perfusion rate during the cooling process (laser off). It is to be noted as a power of laser irradiation increased, more energy is absorbed by the tissue which is subsequently transform into thermal energy, especially at the nearest surroundings. In contrast, low concentration laser power causes a weak thermal response. In Fig. 5a and b

as can be observed, under the given conditions, almost all healthy cells are not damaged and could be recovered after treatment whereas in Fig. 5c, the central part temperature profile higher than hyperthermia limit resulting repair mechanism of the cell is disabled. At a radial distance 2 cm from the irradiated point, the temperature is close to normal liver temperature for the suggested laser powers.

C. The effect of blood perfusion rate

Fig. 6 represents the effect of different blood perfusion rate on temperature distribution in a given laser power exposure time during the cooling process to investigate the hyperthermia phenomena.

It is shown that reducing blood perfusion rate from 0.0005 s^{-1} to 0 s^{-1} that causes temperature increased by approximately 5°C because decreasing blood perfusion rate means less energy
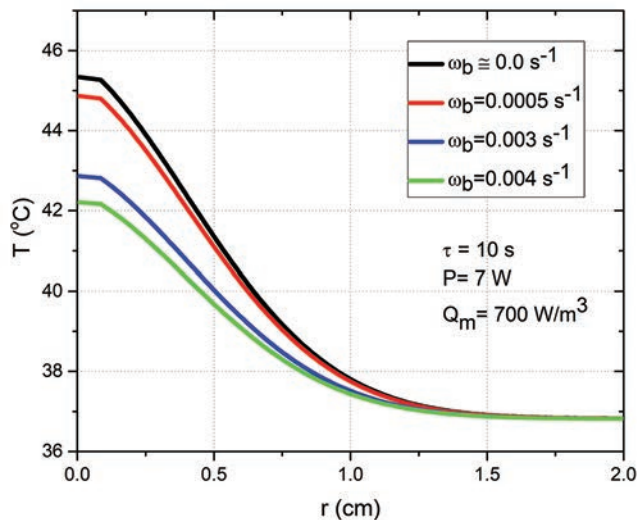
Fig. 6. Impact of blood perfusion rate on heat diffusion in a given time and power.

exchange between the blood capillaries and tissue surroundings due to implement less heat with low blood flow. Overall, the figure shows that under the given constant parameters change blood perfusion rate do not affect the healthy cells because the temperatures located in the range needed for hyperthermia.

IV. Conclusion

Bioheat transfer simulation is extensively useful in thermotherapy applications such as hyperthermia for accurately evaluating heat diffusion in tissue during the treatment. LITT is a minimally invasive therapeutic technique approved for heating a liver tissue and other anatomical regions. Still there does not exist a unique model to describe heat transfer in biological tissue by LITT because the process is involved with many complex phenomena. In this study, a Green function method is used to solve bioheat equation according to thermal and optical properties of LITT to predict amount of heat transfer to liver tissue during hyperthermia treatment. The model shows how the laser power and exposure time are controlled to protect the healthy cells around the damaged zone. The results from this study are compared with the results of other published data (Abbas, Hobiny, and Alzahrani, 2020). It can be observed increase LITT power and exposure time causes increase thermal damage volume. Furthermore, heat diffusion and rate of damage are inversely proportional with blood perfusion rate. The suggested Nd-YaG laser power for liver hyperthermia treatment is 7–10 W with exposure time of 10 s. This approach can be useful before beginning the treatment to correctly predict the diffused temperature and monitor it during the treatment. However, this model is simulated for hyperthermia treatment only, under semi-homogenous system and symmetrical geometry. In further study, more accurate conditions could be added to develop the model to shrink tumors by damaging tissue through thermal coagulation and get better estimation about heat diffusion.

References

Abbas, I., Hobiny, A. and Alzahrani, F. 2020. An analytical solution of the bioheat model in a spherical tissue due to laser irradiation. *Indian Journal of Physics*, 94(9), pp.1329-1334.

Adeleh, K., Reza, H., Mohammed, M. and Hossein, A. 2021. Numerical study on the effect of blood perfusion and tumor metabolism on tumor temperature for targeted hyperthermia considereing a realistic geometrical model of head layers using the finite element method. *SN Applied Sciences*, 3(4), pp.1-17.

Andres, M., Blauth, S., Leithäuser, C. and Siedow, N. 2020. Identification of the blood perfusion rate for laser-induced thermotherapy in the liver. *Journal of Mathematics in Industry*, 10(1), p.17.

Ash, C., Dubec, M., Donne, K. and Bashford, T. 2017. Effect of wavelength and beam width on penetration in light-tissue interaction using computational methods. *Lasers in Medical Science*, 32(8), pp.1909-1918.

Beck, J.V., Cole, K.D., Haji-Sheikh, A. and Litkouhl, B. 1992. *Heat Conduction Using Green Function*. Taylor and Francis., New York. p.552.

Bhowmik, A., Singh, R. and Repaka, R. 2013. Conventional and newly developed bioheat transport models in vascularized tissues: A review. *Journal of Thermal Biology*, 38(3), pp.107-125.

Blauth, S., Hübner, F., Leithäuser, C., Siedow, N. and Vogl, T.J. 2020. Mathematical modeling of vaporization during laser-induced thermotherapy in liver tissue. *Journal of Mathematics in Industry*, 10(1), pp. 1-16.

Chen, C., Lee, I., Tatsui, C., Elder, T. and Sloan, A.E. 2021. Laser interstitial thermotherapy (Litt) for the treatment of tumors of the brain and spine: A brief review. *Journal of Neuro Oncology*, 151(3), pp.429-442.

Chu, K. and Dupuy, D. 2014. Thermal ablation of tumours: Biological mechanisms and advances in therapy. *Nature Reviews Cancer*, 14(3), pp.199-208.

Dutta, J. and Kundu, B. 2018. Thermal wave propagation in blood perfused tissues under hyperthermia treatment for unique oscillatory heat flux at skin surface and appropriate initial condition. *Heat and Mass Transfer*, 54(11), pp.3199-3217.

Faryad, M. and Lakhtakia, A. 2018. *Infinite-Space Dyadic Green Functions in Electromagnetism*. Morgan and Claypool Publishers, San Rafael, California.

Feng, Y. and Fuentes, D. 2011. Model-based planning and real-time predictive control for laser-induced thermal therapy. *International Journal of Hyperthermia*, 27(8), pp.751-761.

Giordano, M., Gutierrez, G. and Rinaldi, C. 2010. Fundamental solutions to the bioheat equation and their application to magnetic fluid hyperthermia. *International Journal of Hyperthermia*, 26(5), pp.475-484.

Hahn, D. and Özisik, M. 2012. *Heat Conduction*, John Wiley and Sons, United States.

Soares, P.I.P., Ferreira, I.M.M., Igreja, R.A.G., Novo, C.M.M. and Borges, J.P.M. 2012. Application of hyperthermia for cancer treatment: Recent patents review. *Recent patents on anti-cancer drug discovery*, 7(1), pp.64-73.

Khaleel, Y., Yahya, S. and Ibrahim, R. 2019. Skin temperature distribution over human head due to handheld mobile phone call using thermal imaging camera. *Aro-The Scientific Journal of Koya University*, 7(2), pp.63-68.

Li, X., Qin, Q.H. and Tian, X. 2019. Thermomechanical response of porous biological tissue based on local thermal non-equilibrium. *Journal of Thermal Stresses*, 42(12), pp.1481-1498.

Milanic, M., Muc, B.T., Lukac, N. and Lukac, M. 2019. Numerical study of hyper-thermic laser lipolysis with 1,064 nm Nd: Yag laser in human subjects. *Lasers in Surgery and Medicine*, 51(10), pp.897-909.

Modest, F. and Mazumder, S. 2021. *Radiative Heat Transfer*, Academic Press, United States.

Niemz, M., 2019. *Laser-Tissue Interactions*, Springer, Germany.

O'neal, D., Hirsch, L. and Halas, N. 2004. Photo-thermal tumor ablation in mice

using near infrared-absorbing nanoparticles. *Cancer Letters*, 209(2), pp.171-176.

Özen, S., Helhel, S. and Cerezci, O. 2008. Heat analysis of biological tissue exposed to microwave by using thermal wave model of bio-heat transfer (TWMBT). *Burns*, 34(1), pp.45-49.

Pennes, H. 1948. Analysis of tissue and arterial blood temperatures in the resting human forearm. *Journal of Applied Physiology*, 1(2), pp.93-122.

Shibib, K., Munshid, A. and Lateef, H. 2017. The effect of laser power, blood perfusion, thermal and optical properties of human liver tissue on thermal damage in LITT. *Lasers in Medical Science*, 32(9), pp.2039-2046.

Skandalakis, G., Rivera, D., Rizea, C., Bouras, A., JesuRaj, J., Bozec, D. and Hadjipanayis, C. 2020. Hyperthermia treatment advances for brain tumors. *International Journal of Hyperthermia*, 37(2), pp.3-19.

Vogl, T., Straub, R., Zangos, S., Mack, M. and Eichler, K. 2004. Mr-guided laser-induced thermotherapy (LITT) of liver tumours: Experimental and clinical data. *International Journal of Hyperthermia*, 20(7), pp.713-724.

Wang, K., Tavakkoli, F., Wang, S. and Vafai, K. 2015. Analysis and analytical characterization of bioheat transfer during radiofrequency ablation. *Journal of Biomechanics*, 48(6), pp.930-940.

Mohammed, Y. and Verhey, J.F. 2005. A finite element method to simulate laser interstitial thermotherapy in anatomical inhomogenous regions. *Biomedical Engineering Online*, 4(1), p.2.

Zhou, J., Chen, J. and Zhang, Y. 2007. Theoretical analysis of thermal damage in biological tissues caused by laser irradiation. *Molecular and Cellular Biomechanics*, 4(1), pp.27.

Zhu, D., Luo, Q., Zhu, G. and Liu, W. 2002. Kinetic thermal response and damage in laser coagulation of tissue. *Lasers In Surgery And Medicine: The Official Journal of The American Society For Laser Medicine and Surgery*, 31(5), pp.313-321.

Data Analytics and Techniques: A Review

Safa S. Abdul-Jabbar[1] and Alaa K. Farhan[2]

[1]Department of Computer Science, College of Science for Women, University of Baghdad, Baghdad, Iraq
[2]Department of Computer Science, University of Technology, Baghdad, Iraq

Abstract—**Big data of different types, such as texts and images, are rapidly generated from the internet and other applications. Dealing with this data using traditional methods is not practical since it is available in various sizes, types, and processing speed requirements. Therefore, data analytics has become an important tool because only meaningful information is analyzed and extracted, which makes it essential for big data applications to analyze and extract useful information. This paper presents several innovative methods that use data analytics techniques to improve the analysis process and data management. Furthermore, this paper discusses how the revolution of data analytics based on artificial intelligence algorithms might provide improvements for many applications. In addition, critical challenges and research issues were provided based on published paper limitations to help researchers distinguish between various analytics techniques to develop highly consistent, logical, and information-rich analyses based on valuable features. Furthermore, the findings of this paper may be used to identify the best methods in each sector used in these publications, assist future researchers in their studies for more systematic and comprehensive analysis and identify areas for developing a unique or hybrid technique for data analysis.**

Index Terms—**Big data analysis, Data analytics, Data analysis, Data management, Machine learning**

I. Introduction

Every company collects a considerable amount of data from various sources. So two prerequisites are needed to secure this data and use techniques to extract useful information from this data (Khoshbakht, Shiranzaei and Quadri, 2021; Farhan and Ali, 2017). The use of big data has rapidly progressed from a theory to a reality with the rapid progression of data resources and the creation of companies specializing in big data (Zheng and Guo, 2020; Do Nascimento, et al., 2021; Mariani and Baggio, 2022). For example, clients struggle to find relevant and acceptable material that satisfies their needs because the amount of data on the internet is constantly rising. When a customer submits a query for information or data to an Internet search engine, the result is typically

many pages. Hence, he faces the repetitious task of locating the appropriate data from this flood of results. The term describing this problem is called "Data Overloading" (Kan and Klavans, 2002). Hence, the primary objective of this decade of electronic revolution is to construct and ensure a better manner of managing, collaborating, and developing via the use of computer and information technology-based knowledge and information-oriented services (Rajon, Shamim and Arif, 2011; Russell and Norvig, 2020). The process of analyzing and discovering hidden patterns, undiscovered correlations, and other valuable business information from a vast volume of data is known as big data analytics (Patel, Singh and Kazi, 2017; Faizan, et al., 2020). Therefore, data analytics is a crucial subject for many systems, such as those that work with strings or information retrieval operations (Abdul-jabbar and George, 2017). Furthermore, data analytics can be used to check the privacy issues in social media, such as tags and image uploading, as we can see on Flickr and Facebook (Smith, et al., 2013; Abkenar, et al., 2020). Besides social media applications, data analytics can provide many services for applications in different fields, such as audio and video (Verma and Agrawal, 2016).

This paper has three overarching goals:
1. It will provide a brief history of data analytics techniques and methods for documents and describe how data analytics tools utilize the knowledge from all input documents.
2. Presents how the previous studies are based on multi-algorithms and multi objective to optimize the traditional methods and explain the researcher with a comprehensive overview that helps him choose the suitable algorithms and integrate them into a model according to the task at hand.
3. Finally, this paper also illustrates the limitations of each proposed method to present new directions in future works.

The paper is structured as follows, Section II introduces the proposed data analytics techniques and methods, and Section III interprets and describes the significance of our findings in the published paper. Finally, Sections IV and V present a compelling discussion and conclusions that inform researchers on what they can learn from published research papers mentioned in this research.

II. Data Analytics and its Methods

Data analysis primarily entails big data analytical methodologies, systematic architecture, data mining, and analysis tools. The most crucial phase in big data is data

research, which involves examining significant values, making recommendations, and making judgments and decision support tools that have gained popularity, such as executive information systems and online analytical processing. Therefore, data analysis and interpretation complexity encourage researchers and companies to use algorithms that process real-time data, analyze it, and produce highly accurate analytics results. In addition, data analysis can be used to investigate potential values where this information can be used for business development and performance enhancement, such as predictive analytics that can make future predictions. Data analytics is a wide, dynamic and complex field because data comes in different types and grows significantly. Furthermore, the purpose of the analysis varies depending on the type of application required (Schwarz, Schwarz and Black, 2014; Harfouchi, et al., 2017; Rajaraman, 2016). Hence, data analytics aims to answer three categories of questions in general. As shown in Fig. 1, these elucidate what happened in the past, what is happening now, and what is anticipated (Ghavami, 2020).

As a result, processing and obtaining the necessary information from an extensive database cost a lot of time and processing power (Abdul Majeed, Kadhim and Subhi Ali, 2017). Moreover, interdisciplinary investigation makes it difficult for businesses to identify the specialist skills needed to conduct a large-scale reality check. Therefore, viable research provides critical features for completing this activity and overcoming the inaccessibility of analytical capabilities (Kashyap, 2019).

In other words, data analytics can be defined as a data science used to break data into individual components for personal inspection and integrate these components to create knowledge. Informally, Oracle and Cloudera have proposed a seven-step "value-chain" approach for extracting value using data analytics; these steps are as follows (Ghavami, 2020):

1. Objectives identification.
2. Business levers identification.
3. Data collecting.
4. Data cleaning.
5. Data modeling.
6. Data science team creating (i.e., building solid teams).
7. Optimize and repeat.

On the other hand, Dr. Carol Anne Hargreaves proposed another seven steps for the business analytics process in her data science process model, which also can be listed as follows (Ghavami, 2020):

The Past	**The Present**	**The Future**
Retrospective View	**Real-time View**	**Prospective View**
– What happened?	– What is happening now?	– What will happen next?
– Why it happened?	– Uses real-time data	– How can I intervene?
– Uses historical data	– Actionable dashboards	– Uses historical and real-time data
– Delivers static dashboards	– Alerts	– Predictive dashboards
	– Reminders	– Knowledge-based dashboards

Fig. 1. Big data analytics' temporal questions (Peter Ghavami, 2020).

1. Business needs identification.
2. Explore the data.
3. Analyze the data.
4. Predict what is likely to happen.
5. Optimize (find the best solution).
6. Make a decision and measure the outcome.
7. Update the system with the results of the decision.

All kinds of data analytics processes, including the traditional Knowledge Discovery in Databases (KDD) process, and others such as (Mishra and Sharma, 2014), who proposed six steps for data analytics and (Chen, Mao and Liu, 2014) suggested three primary steps only and many others. These proposed systems depend on big data analytics tools that provide valuable knowledge for enhancing business. Typically, these methods can be used in different analytics models that can be divided into the following types:

A. Advanced analytics and predictive modeling

Machine learning, data science, and predictive modeling have grown widespread in every area where data analysis plays a key role (Butcher and Smith, 2020). Data mining is a sophisticated technique for evaluating large amounts of data. There are two forms of data analytics: Supervised/ unsupervised data analytics. Based on the findings of previous research studies, predictive modeling works in different scopes with solid chances of achieving efficient results when used with unsupervised analytics than with supervised analytics (Fan, et al., 2018). A prediction model is built by learning a dataset with a known outcome (classified results) and then determining the effects of unclassified cases (Shouval, et al., 2014). De Fortuny, Martens and Provost, showed in 2013 that when predictive models are created depending on varied and accurate data, they can provide a performance improvement even on a large amount of data. In this study, the researchers trained models and made predictions on sparse datasets using the Naive Bayes classifier. Data from several different predictive modeling applications are used to test the proposed method. The proposed method can conclude that the system with big data might be more efficient for predictive analytics operations. Consequently, organizations with more data and better understanding may gain significant competitiveness (De Fortuny, Martens and Provost, 2013). The data analytics technologies can be used in health-care systems as in 2014 when Pourhomayoun, et al., 2014 proposed a new system for remote health monitoring (Pourhomayoun, et al., 2014).

On the other hand, machine learning is one of the most critical data analytics approaches with significant facilitators of knowledge-intensive automation that can be used in many applications (Mishra and Sharma, 2014; Cearley, et al., 2018). Therefore, ML Algorithms are used in different applications such as medical, roads and many other applications with the risk of facing many problems in robustness, monitoring, alignment and systemic safety that should be handled (Rajpurkar, et al., 2017; Hendrycks, et al., 2021).

As an example of using neural networks for advanced data analytics systems, in 2017, Jain presented the implementation details for detecting telecommunication fraud using Data Stream Analytics and Neural Network classification-based Data Mining. The proposed method depends on Microsoft Azure's Event Hub and Stream Analytics components for fraud detection using a self-coded algorithm and a Data Mining Neural Network Pattern Recognition tool. The findings indicate that the proposed methodologies are accurate and efficient and may be extended to various cloud analytics systems and provide a foundation for big data analytics and mining (Jain, 2017). Furthermore, Talasila, et al. (2020) presented a novel neural network-based method for medical data analytics and disease prediction in 2020. They employed rough set theory to choose the most significant characteristics and then fed them into a Recurrent Neural Network for disease forecasting. As a result, the new technique had a 98.57% accuracy, more than the current accuracy presented by the existing methods for the heart disease dataset (Talasila, et al., 2020). Furthermore, dealing with big data can be aided by deep learning, which has the potential to extract complicated abstractions (Vu, et al., 2021). A new analytics model for distant physiological data was proposed based on powerful clustering techniques and multi-model classification. The proposed model is decomposed into several steps. The first is remote health monitoring and body sensor networks, which collect the data and send it to the analytical system. Then the data preprocessing and feature extraction step should be done to the received data. Followed by data sample clustering and group-specific feature selection, the multiple model classification must be done as a final step in this model. The proposed model was evaluated using a subset of data acquired from 600 heart failure patients through a remote health monitoring system. The proposed model dramatically improved prediction accuracy and performance (Pourhomayoun, et al., 2014). Whereas in 2019, Corizzo, Ceci and Malerba, 2019 were inspired by the goals of scientific studies sponsored by the European Commission and several national governments (Corizzo, Ceci and Malerba, 2019). They employed recommended methodologies based on distributed architectures, big data analytics, and predictive modeling research domains. The results of the proposed system give accurate predictions (temporal and geographical) that are scalable in big data. While in 2021(Hamarashid, Saeed and Rashid, 2021), a new paper was published to present a novel model for predicting the next word depending on the N-gram method with a sufficient increase in the number of N-grams used to reduce the time for predicting the next word in Kurdish dataset. The proposed model achieved results with accuracy up to 96.3%. Also, in 2021 another research was presented to produce a prediction model for healthcare centers based on machine learning algorithms and analysis methods (Moharram, Altamimi and Alshammari, 2021). In this paper, they analyze the input data to reduce the number of training data. Then, three machine learning algorithms were applied (Logistic Regression, JRip, and Hoeffding tree) to compare the results and select the best one for the proposed system. The proposed model produces similar effects in predicting

appointment no-shows in pediatric outpatient clinics with roughly 90% classification accuracy. Furthermore, in the same year, Rocha, et al. (2021) used Principal Component Analysis techniques and unsupervised algorithms to perform better clustering. As a result, K-mean clustering algorithm shows the best results for clustering operation (Rocha, et al., 2021).

B. Model accuracy and optimization

There are several optimization strategies available by multiobjective optimization approaches (Zarchi and Attaran, 2019), (Wang, et al., 2011), (Jaouadi, et al., 2020). In 2020, Castellanos, et al. showed how to specify, deploy and track performance metrics in big data analytics applications based on domain-specific modelling and DevOps using a design process methodology based on the Attribute-Driven and Architecture analysis method technique (Castellanos, et al., 2020). Furthermore, many researchers employ the approximation model instead of the accurate numerical simulation model to improve the effectiveness of the current multiobjective optimization approaches in dealing with complicated engineering issues (Choi, Cho and Kim, 2018). Therefore, employing optimization techniques is the best method for identifying suitable model parameters (Kumar, et al., 2018). A data-driven predictive modeling strategy for forecasting surface roughness in additive manufacturing is developed to optimize the integrity of fabricated components. Various sensors of various sorts are used to collect data on temperature and vibration. An ensemble learning approach is used to train the surface roughness prediction model. A subset of these characteristics is chosen to enhance computational complexity and accuracy rate. As a result, the proposed model can provide accurate predicting results. At the same time, the frequency amplitude of the build plate vibrations, the extruder vibrations, and the temperatures influence the outcome (Li, et al., 2019). Whereas in the education sector, Tran, et al. (2019) have published a paper that described the benefits of Federated Learning and suggested a new system by establishing Federated Learning over a wireless network. This paper fills the trade-offs between computation and communication latencies caused by learning accuracy level, Federated Learning time, and energy consumption of mobile user equipment. They found the globally optimal solution by finding the confined methods to all sub-problems. This solution provides exciting insights into design issues through the ideal Federated Learning over wireless network learning (time, accuracy, and user equipment's energy cost) obtained through numerical and theoretical analysis (Tran, et al., 2019). In 2019 Zou, et al., proposed a new vehicle evaluation prediction model (Zou, et al., 2019). This model is used to optimize the traditional logistic regression algorithm by studying the logistic mathematical model, designing the error function, using the gradient descent method to discover the regression coefficient, and optimizing the sigmoid function. Consequently, the training time and classification effect were enhanced, and the accuracy is maintained.

In 2020, Liu, et al. developed a new adaptive model for efficient multiobjective optimization. This model depends

on micro multi objective genetics to improve performance. The optimization results further demonstrate the proposed model's usefulness in real-life applications. However, this model needs more samples and local-densifying iterations to provide reliable optimization results (Liu, et al., 2020). Also, in 2020 Ben Seghier, et al. (2020) proposed a hybrid Artificial Intelligence model that aims to create a hybrid framework for predicting and analyzing stress intensity factors. This framework was built by building an adaptive neuro-fuzzy inference system, tuned using two meta-heuristic algorithms: genetic algorithm and particle swarm optimization. The proposed model outperformed the other AI models for accurate prediction, with R^2 = 0.9913, RMSE = 23.6, and MAE = 18.07. However, increasing the datasets generated based on actual test results or bigger finite element method computations that include a variety of ranges and materials might enhance prediction performance (Seghier, et al., 2020).

C. Natural language processing

Researchers in the discipline can leverage techniques developed to appropriately and accurately analyze language. For example, natural language techniques have computational assessments of various language features about specific goals, and deep learning techniques such as CNN are widely used in this area (McNamara, et al., 2017; Shamsaldin, et al., 2019). Hence, the Natural Processing techniques allow researchers to collect and analyze data to extract the information (Rajput, 2019). However, one of the significant obstacles in text categorization is the optimization problem. This problem can consider an analytical issue for document summarization, prompting a group of academics to create a nature-inspired optimization technique based on a multi-criteria optimization model linked to Artificial Bee Colonies (ABC). The suggested technique in 2018 yielded significant gains, with average increases of 31.09% (8.43%) and 18.63% (6.09%) in ROUGE-2 (ROUGE-L) compared to the best single-objective and multiobjective findings in the published studies (Sanchez-Gomez, Vega-Rodríguez and Pérez, 2018). In the same year (Rashid, Mustafa and Saeed, 2018), Rashid, Mustafa and Saeed (2018) applied a stemmer to Kurdish text documents (KDC-4007 dataset). They used three algorithms: Support Vector Machine, Naïve Bays, and Decision Tree, to classify Kurdish text. After the preprocessing phase, they found that the support vector machine achieved the best accuracy among all the applied algorithms. In 2019, researcher Sanchez-Gomez, Vega-Rodríguez and Pérez, 2018 continued developing the research proposed in the previous year by creating an indicator based on a multi objective Artificial Bee Colony. The developed system was tested on several datasets (the same datasets used in their previous research) and evaluated the results using a variety of measures. Consequently, the results for ROUGE-2 and ROUGE-L have improved to between 7.37% and 40.76% and 2.59% and 11.24%, respectively (Sanchez-Gomez, et al., 2019).

On the other hand, Yadav and Chatterjee (2016) describe an efficient and robust summarizing approach based on the meaning of essential words in the content for text summarization. Sentiment analysis is constantly utilized for large-scale text data analysis and subjectivity analysis. This study demonstrates that sentiment analysis may be used well for text summarization and provides an efficient way to summarize the content, particularly for 50% (Yadav and Chatterjee, 2016). Furthermore, researchers can employ a lexicon-based technique to examine students' responses. A new algorithm has been suggested to establish teachers' opinion results by extracting semantic meaning from students' comments, including intensifier words, and determining the amount of positive or negative thoughts. This method displays the instructors' opinion results, categorized according to the strength of the positive or negative sentences. However, utilizing a lexicon approach to sentiment analysis is not optimal because some crucial details might be lost (Aung and Myo, 2017).

A summarization system can be designed depending on the dataset's similarity or dissimilarity measures. The research performed by Saini, et al., 2019 presented effective feature summarization for text as a binary analysis issue. They use a multi objective binary differential evolution-based optimization technique. Differential evolution's solutions encode a potential subcategory of sentences to be included in the summary, and then assessed using objective functions such as the sentence's location in the document. The results show that good improvements were obtained depending on the dataset used and the objective function (Saini, et al., 2019). In 2021, another paper was proposed to perform data analysis using state-of-the-art techniques. Using syntax analysis, they developed a method capable of extracting the recent Toolkit for ATM Incidence investigative process taxonomy factors from free-text safety reports. Finally, they modify a Data-Driven Method capable of automatically determining the cause of the aircraft accident. The results demonstrate that when merely elevated predictions are considered, the model provided pilots' contribution is around 97% accurate and 94% for ATCo (Buselli, et al., 2021). In the same year, Vargas-Calderón, et al. (2021) presented a model used in healthcare applications to evaluate the quality of service in hospitals depending on client reviews. After the text extraction and cleaning step, the model was designed depending on multi-ML algorithms (Vargas-Calderón, et al., 2021). Furthermore, in 2021 Hryshchenko and Yaremenko implemented the bloom filter, naïve Bayesian classifier, and neural networks to categorize a batch of text data and determine both disadvantages and advantages of each method (Hryshchenko and Yaremenko, 2021). At the same time, Yaremenko, Rogoza and Spitkovskyi (2021) developed a neural network architecture that can process a large amount of data in real-time systems and handle the determined limitation of the applied mathematical models of the standard Neural Networks and Naive Bays (Yaremenko, Rogoza and Spitkovskyi, 2021).

D. Quantitative analysis (prediction and prognostics)

Quantitative analysis is concerned with quantifying and analyzing variables to arrive at conclusions. It entails using

statistical tools to analyze numerical data for answering questions such as who and when. Apuke published his work on predictor measurement heterogeneity by altering the degree of measurement error across derivation and validation scenarios. Hence, he generated hybrid predictor measurements using measurement error models (Apuke, 2017; Pajouheshnia, et al., 2019; Luijken, et al., 2019; and Luijken, et al., 2020). In 2021, Admiraal, et al. used 12 quantitative features gathered from various patient situations to train several types of machine learning algorithms. The research results show that machine learning employing quantitative features derived from collected data has a better precision than visual data analysis in predicting poor prognosis following cardiac arrest, making it a potential alternative to visual analysis (Admiraal, et al., 2021). In 2022, Luijken, Song and Groenwold proposed a paper to analyze the expected predictor measurement diversity impact. In period outcome data, simulation research was conducted to examine the influence of predictor measurement variation across validation and implementation settings. The application of quantitative prediction error analysis was demonstrated with an illustration of forecasting the 6-year probability of acquiring second type diabetes with variability in the predictor body mass-index measurement. As a result of this paper, all situations of predictor measurement variability resulted in the poor measurement of prediction models, and overall accuracy was lowered. Furthermore, it increased random predictor measurement variability (Luijken, Song and Groenwold, 2022).

Moreover, artificial systems generate high-level data representations from large-scale data, particularly unlabeled data, which is plentiful in Big Data (Chen and Lin, 2014), (Najafabadi, et al., 2015). In 2020, Zhong, Yu and Ai proposed the big data-based hierarchical deep learning system in the context of employing deep learning for data analytics. This system uses behavioral and content features to interpret network traffic patterns and information encoded in the payload. When several machines are deployed, the findings of this suggested system demonstrate that it may boost the detection rate of intrusive attacks and reduce the time spent significantly (Zhong, Yu and Ai, 2020).

E. Ensemble of models (data analytics prediction framework)

In many real-world applications, the availability of classified data is restricted, making it challenging to detect and eliminate duplicate and unnecessary variables from the feature-set, particularly in high-dimensional applications. This circumstance naturally happens in many real-life situations when a large amount of data can be acquired inexpensively and quickly. Yet, the manual classification of samples is time-intensive and cannot be assumed. Many approaches were suggested to improve accuracy in machine learning; one of these approaches is to aggregate the output of several learners. Ensemble Learning is a term used to describe this approach. Bagging, boosting, stacking, and error-correcting output are the four methods for merging

several models (Wang, et al., 2014). The learning under supervision finds clusters with high probability densities in individual classes. It is employed when there is a reference value and training set with the variables to cluster (Dean, 2014). Whereas in unsupervised learning, feature selection seeks to locate meaningful subsets of features that yield best groupings clustering by clustering "similar" items together using any similarity metric (Nag and Mitra, 2002), (Dy and Brodley, 2004), (Hong, et al., 2008), (Elghazel and Aussem, 2010). In 2011, Rajon, Shamim and Arif proposed a complete framework that designed and implemented a generic product-independent e-market model for emerging economies. This paper's fundamental contribution is creating and executing a generic e-marketplace model for emerging economies where agriculture is widely practiced and a thriving manufacturing sector. A comprehensive examination of the utility and efficacy of establishing e-Commerce and e-Commerce services has also been presented (Rajon, Shamim and Arif, 2011). Rajon, Shamim and Arif proposed a method based on the random sample partition model that retains the statistical features of the data set in each data block in 2018. They presented the Alpha framework, which consists of three primary layers for data administration, batch management, and data analysis, to enhance the efficiency of big data analysis with Random Sample Partition blocks. The results show that the proposed method can provide approximate results for data analysis tasks such as data summarization and the Alpha framework for Big Data Analysis tasks (Salloum, et al., 2018). In the same year, Yu, et al. (2018) design a model to demonstrate how boosting and bagging approaches can be compared to produce better explanatory models to prove that the ensemble approaches are more suitable for some problems than other approaches (Yu, et al., 2018).

On the other hand, Kumar, Singh and Buyya proposed a new ensemble learning-based workload prediction model in 2020, which makes use of excessive learning machines and weights their estimates with a voting engine. The optimized weights are chosen using a metaheuristic algorithm motivated by the black hole theory. The results demonstrate the approach's superiority over conventional methods, with a reduction in RMSE of up to 99.20% (Kumar, Singh and Buyya, 2020). Whereas in 2021 a new framework based on features modeling and ensemble learning to predict query performance was proposed by Zaghloul, Salem and Ali-Eldin (2021) using Machine learning algorithm attempting to predict a performance metric based on the amount of time elapsed and ensemble learning (Zaghloul, Salem and Ali-Eldin, 2021).

III. Summary and Comparisons

Data analytics methods and techniques have more and more applications in life, and performance enhancement solutions are widely applied. It is essential to improve the efficiency of any application when dealing with big data by enhancing the result accuracy and processing time; this will be done by analyzing the input data and extracting

only the relevant information that the application needs. Depending on this principle, many types of research papers in different scopes were published to propose new techniques that can be used to enhance the analysis results. This paper provides a survey of these research papers as summarized in Table I. The research papers that proposed new methods in the literature review section were presented in this table; to illustrate the research scope: The field in which this research was developed, the main research issues: Used to show the problems that the research tries to solve, the research

techniques used to describe the method and tools that used to implement this research, and the main research findings used to describe the most important results obtained or concluded from the published research.

It can be seen from the summary of the research papers presented in Table I that most researchers suggested methods that used machine learning algorithms. Accordingly, this paper discusses the proposed methods' details and their impact on the results, as shown in Table II. This table focuses on the strengths or offers the capabilities of ML algorithms in

TABLE I
The Categorization of the Discussed Previous Works in Terms of Researches Scope, Main Research Issues, Main Research Techniques, and Main Research Findings

References	Research scope	Main research issues	Main research techniques	Main research findings
De Fortuny, Martens, and Provost, 2013	Big data	Critical prediction jobs	Multivariate Bernoulli Naive Bayes	Organizations with more data selection provide a better understanding of improving the system performance
Pourhomayoun, et al., 2014	Healthcare systems	Remote health monitoring	Multi-model approach	Improved the prediction accuracy and performance
Corizzo, Ceci and Malerba, 2019	Energy sector	Accurate big data analytics and predictive modeling	Multi-model approach	Accurate results of predictions with big data
Moharram, Altamimi, and Alshammari, 2021	Healthcare systems	Reduce the number of training data	ML algorithms	The best algorithm used was the Hoeffding tree, with a classification accuracy of 90%
Rocha, et al., 2021	Human development indicators	Classify the departments of peru according to their human development index using clustering techniques	Multi-model approach	
Li, et al., 2019	Extrusion-based additive manufacturing	Optimize the integrity of fabricated components	ML algorithms	Providing accurately predicted results
Tran, et al., 2019	Education sector	Analyze the trade-offs between (computation *vs*. communication latencies) (Learning time *vs*. energy consumption)	Confined methods to all sub-problems	Provide effective cost, time, and accuracy
Zou, et al., 2019	Vehicle evaluation	Many iterations and training vast amounts of data take a long time	ML algorithms	The training time is reduced, the classification effect is enhanced, and the accuracy is maintained
Liu, et al., 2020	Multiobjective optimization	High computational cost	Multi-model approach	The proposed model's useful in real-life applications
El and Ben, 2020	Stress intensity factor	Prediction of stress intensity factor	ML algorithms	Provide accurate predictions
Sanchez-Gomez, Vega-Rodríguez and Pérez, 2018	Text summarization	Find the essential information from a document collection	Optimization algorithms	Provide essential improvements compared to the traditional approaches
Sanchez-Gomez, et al., 2019	Text summarization	Find the essential information from a document collection	Optimization algorithms	Provide improved performance and accurate results compared to the previous version of this proposed model
Yadav and Chatterjee, 2016	Text summarization	Find an efficient way to summarize texts	Sentiment analysis	Provide an efficient and robust summarizing approach based on feelings of significant words
Aung and Myo, 2017	Education system	Analysis of students' comment	Sentiment analysis	Provide good results but not optimal because some crucial details might be lost
Saini, et al., 2019	Text summarization	Design an automatic text-summarization	ML algorithms	Provide good improvements to the traditional approaches
Buselli, et al., 2021	Air traffic management	Automation and digitization to maintain safety aviation	Multi-model approach	The model provides pilots' contribution is around 97% accurate and 94% for ATCo
Admiraal, et al., 2021	Healthcare systems	electroencephalography reactivity quantitative analysis of neurological prognostication following cardiac arrest	ML algorithms	Provide better precision than visual data analysis in predicting poor prognosis following cardiac arrest

(Contd...)

TABLE I
(Continued)

References	Research scope	Main research issues	Main research techniques	Main research findings
Luijken, Song and Groenwold, 2022	Healthcare systems	Analyze the effect predictor measurement heterogeneity	Quantitative prediction error analysis	Proves that the increasing random predictor measurement heterogeneity will decrease the model discrimination
Rajon, Shamim and Arif, 2011	Electronic commerce	Design a general prototype for the e-marketplace	General framework designing tools	They are creating and developing an e-market prototype that is product-agnostic
Salloum, et al., 2018	Big data	Making the extensive data analysis is feasible when the volume of data exceeds the available computation power	ML algorithms	Enhance the efficiency of extensive data analysis
Kumar, Singh, and Buyya, 2020	Cloud systems	These systems must allocate and deallocate resources with low operational cost and maintain the quality of services	ML algorithms	Provide accurate predictions by reducing the error prediction
Zaghloul, Salem, and Ali-Eldin, 2021	Query optimizer	Attempt to predict a performance metric	ML algorithms	Provide effective query performance metric
Jain, 2017	Fraud detection	Provide an efficient method for fraud detection using a self-coded	ML algorithms	Provide accurate and efficient cloud analytics
Talasila, et al., 2020	Healthcare systems	Provide an accurate method for diseases prediction	ML algorithms	Provide prediction accuracy up to 98.57%
Zhong, Yu, and Ai, 2020	Intrusion detection	Intrusion detection depends on data analytics	ML algorithms	Boost the detection rate of intrusive attacks and reduce the time spent significantly

ML: Machine learning

the proposed methods. It has been formulated to include: the ML algorithms used to design the proposed method and the techniques used to describe how these algorithms are used to build the proposed method. Furthermore, this table illustrates the effect of the proposed method on the system performance and the efficiency of the results. The most vital limitations are also presented in this table to describe the challenges and problems in the proposed methods. As we have seen in (Admiraal, et al., 2021) and (Zhong, Yu, and Ai, 2020), there are limitations in the number of models used to test the proposed method, so it may have some implementation issues when used with a large amount of data. Furthermore, (Kumar, Singh, and Buyya, 2020), there is another issue: people must determine the number of networks and nodes in hidden networks. This causes the system to need human intervention at the data entry stage and is not entirely automated. While in (Jain, 2017), the proposed model was designed without any dynamic implementation, which causes the possibility of facing several problems when applied to real-time data.

IV. Discussion

Data analytics aims to provide meaningful and relevant information. However, many users are uncertain about the sort of analysis to perform on data collection and which kinds of visual data presentation are appropriate. This paper presented a comprehensive review of data analytics techniques to help researchers construct an accurate and effective analytics tool to be used efficiently in the intentional system. This leads to utilizing these
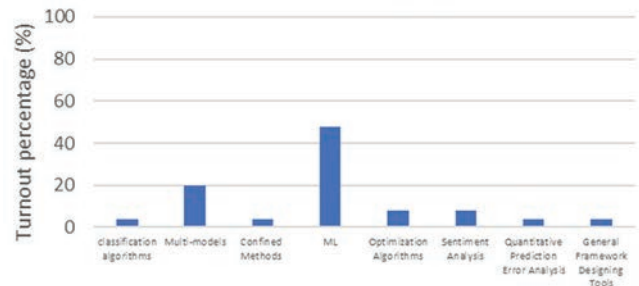
Fig. 2. The percentage of turnout for each method.

analytics tools in the best way to provide better privacy, less power, more efficient results, and economic services rather than relying only on standard analytics tools to solve the research problem in different scopes. According to Table I, ML algorithms have shown the highest usability in data analysis systems than other algorithms because they provide good accuracy with higher performance capacity in multiple areas of industrial, commercial, agricultural, health, education, and mining activity. As a result, these algorithms contribute to the development, increased employment, the contribution of mining techniques and increased business investment. Furthermore, Fig. 2 illustrates the turnout percentage, which shows that ML algorithms are superior to other methods. Therefore, the ML improvements were presented in Table II to show the effectiveness of each proposed method.

Finally, it is worth mentioning that many challenges can be faced when adopting ML algorithms in data analytics systems. For example, processing time, computation power… etc. However, these challenges can be addressed using

TABLE II
MACHINE LEARNING ALGORITHMS IN EACH REFERENCE AND ITS CONTRIBUTIONS

References	The used ML algorithms	The design techniques	Improvement effects			Research limitation
			Time	Accuracy	Processing	
Moharram, Altamimi and Alshammari, 2021	logistic regression, JRip, and Hoeffding tree	The optimum machine-learning method was discovered by comparing three ML algorithms by computing recall and prescience for each	*	*	*	-
Li, et al., 2019	RF, AdaBoost, CART, SVR, RR, RVFL	Data was collected depending on multiple sensors, extracted and selected features from these data, then applied selective algorithms to these features		*		-
Zou, et al., 2019	Logistic regression	Using the logistic regression algorithm based on the gradient descent method and optimizing the sigmoid function	*	*	*	-
Seghier, et al., 2020	Genetic and particle swarm algorithms	Hybrid AI model (GA and PSO) for accurate prediction trained on data values obtained using the finite element method calculations	*	*	*	-
Saini, et al., 2019	Genetic operation inspired form genetic algorithm	Using the binary differential evolution and self-organizing map (based on genetic operations) to select a subset of sentences, then evaluate these sentences using the objective functions		*	*	-
Admiraal, et al., 2021	LR, SVM, RF, NN, and GTB algorithms	The RF classifier was the best machine-learning classifier after ML models were trained with twelve quantitative variables derived from data of 134 adult cardiac arrest patients		*		A limited number of samples for training operation
Salloum, et al., 2018	Hadoop MapReduce and data mining and machine learning algorithms	The data is assigned to the blocks sampling algorithm, followed by blocks analysis and ensemble estimates techniques. Finally, an ensemble model evaluation was applied to check whether it fits the criteria	*	*	*	-
Kumar, Singh, and Buyya, 2020	Extreme learning machines	They were using extreme learning machines and a voting engine to optimize the weights of the prediction outcomes. These weights are defined using a black hole-inspired metaheuristic algorithm		*	*	Heuristics are used to manually define the number of networks and the number of hidden nodes in each network
Zaghloul, Salem, and Ali-Eldin, 2021	XGBoost algorithm	They used real-world datasets to predict query performance using feature modelling and ensemble learning based on the duration of time elapsed	*	*	*	-
Jain, 2017	Neural network	The event hub and Stream analytics of microsoft azure with the neural network and self-coded algorithms were employed to detect fraud telecommunication		*	*	It cannot be applied to real-time dynamic data
Talasila, et al., 2020	Recurrent Neural Network	The most significant characteristics were selected using rough set and then fed into a recurrent neural network to predicate the system results	*	*		-
Zhong, Yu, and Ai, 2020	Deep learning	The hierarchical deep learning system used the behavioural and content features to interpret network traffic patterns and information encoded in the payload	*	*	*	Effective when applied to a small number of samples

ML: Machine learning

parallel and distributed frameworks and choosing appropriate algorithms to implement for each system.

V. CONCLUSION

In this paper, several articles are reviewed in different analytics models. In addition, this study addressed the specified standards algorithms used for designing each system. Furthermore, it highlighted the advantages and disadvantages of the specified big data analytics methodologies, evaluating them in terms of scalability, efficiency, precision, and privacy. Furthermore, the suitable employment of data analytics in different scopes and applications has been adjusted to construct how it can be used to provide a high-quality performance. It should be noted that this paper recognizes that the core function of machine learning is to offer analytical answers that can be developed based on

the behavior of previous data models. As a result, this paper intends to provide simple research examining several proposed analytics technologies from various perspectives and fills in the gaps clearly in unknown information. From this examination, we can conclude that when the ML algorithms and data analytics self-tuning system feature selections have been used, they will improve performance compared to other approaches and techniques. Several paper investigations in many sectors have demonstrated the possibility of using machine learning algorithms in data analytics systems to improve performance speed and accuracy.

REFERENCES

Abdul Majeed, G., Kadhim, A. and Subhi Ali, R. (2017). Retrieving encrypted query from encrypted database depending on symmetric encrypted cipher system method. *Diyala Journal For Pure Science*, 13(1), pp.183-207.

Abdul-jabbar, S.S. and George, L.E. (2017). Fast text analysis using symbol enumeration and Hashing methodology. *Fast Strings Search Process*, 58(1), pp.345-354.

Abkenar, S.B. Kashani, M.H., Mahdipour, E. and Jameii, S.M. (2020). Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. *Telematics and Informatics*, 57, 101517.

Admiraal, M.M., Ramos, L.A., Delgado Olabarriaga, S., Marquering, H.A., Horn, J. and van Rootselaar, A.F. (2021). Quantitative analysis of EEG reactivity for neurological prognostication after cardiac arrest. *Clinical Neurophysiology*, 132(9), pp.2240-2247.

Apuke, O.D. (2017). Quantitative research methods : A synopsis approach. *Kuwait Chapter of Arabian Journal of Business and Management Review*, 6(11), pp.40-47.

Aung, K.Z. Myo, N.N. (2017). Sentiment Analysis of Students'Comment Using Lexicon Based Approach. *IEEE/ACIS 16th International Conference on Computer and Information Science*, pp.149-154.

Ben Seghier, M., Carvalho, H., Keshtegar, B. and Correia, J.A.F. (2020). Novel hybridized adaptive neuro-fuzzy inference system models based particle swarm optimization and genetic algorithms for accurate prediction of stress intensity factor. *FFEMS*, 43(11), pp.2653-2667.

Buselli, I., Oneto, L., Dambra, C., Gallego, C.V., Martínez, M.G., Smoker, A. and Martino, P.R. (2021). Natural Language Processing and Data-Driven Methods for Aviation Safety and Resilience : From Extant Knowledge to Potential Precursors. Open Research Europe.

Butcher, B. and Smith, B.J. (2020). Feature engineering and selection: A practical approach for predictive models. *The American Statistician*, 74(3), pp.308-309.

Castellanos, C., Pérez, B., Varela, C.A. and Correal, D. (2020). A Model-Driven Architectural Design Method for Big Data Analytics Applications. *Proceedings 2020 IEEE International Conference on Software Architecture Companion*, *ICSA-C 2020*, pp.89-94.

Cearley, D.W., Natis, Y., Walker, M. and Burke, B. (2018). Top 10 Strategic Technology Trends for 2018. Gartner, Stamford. Available fromt: https://www.gartner.com/ngw/globalassets/en/information-technology/documents/top-10-strategic-technology-trends-for-2018.pdf [Last accessed on 2017 Oct 3].

Chen, M., Mao, S. and Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2), pp.171-209.

Chen, X.W. and Lin, X. (2014). Big data deep learning: Challenges and perspectives. *IEEE Access*, 2, pp.514-525.

Choi, B.C., Cho, S. and Kim, C.W. (2018). Kriging Model Based Optimization of MacPherson Strut Suspension for Minimizing Side Load using Flexible Multi-Body Dynamics. *International Journal of Precision Engineering and Manufacturing*, 19(6), pp. 873-879.

Corizzo, R., Ceci, M. and Malerba, D. (2019). Big Data Analytics and Predictive Modeling Approaches for the Energy Sector. *2019 IEEE International Congress on Big Data* (*BigDataCongress*), pp.55-63.

De Fortuny, E.J., Martens, D. and Provost, F. (2013). Predictive modeling with big data: Is bigger really better? *Big Data*, 1(4), pp.215-226.

Dean, J. (2014). Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. John Wiley and Sons, Hoboken. Available from: https://doc.lagout.org/Others/Data Mining/Big Data, Data Mining, and Machine Learning_Value Creation for Business Leaders and Practitioners %5BDean 2014-05-27%5D.pdf. [Last accessed on 2022 Apr 01].

Do Nascimento, I.J.B., Marcolino, MS., Abdulazeem, H.M, Weerasekara, I., Azzopardi-Muscat, N., Gonçalves, M.A. and Novillo-Ortiz D. (2021). Impact of big data analytics on people's health: Overview of systematic reviews and recommendations for future studies. *Journal of Medical Internet Research*, 23(4), p.e27275.

Dy, J.G. and Brodley, C.E. (2004). Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5, pp. 848-889.

Elghazel, H. and Aussem, A. (2010). Feature selection for unsupervised learning using random cluster ensembles. *Proceedings IEEE International Conference on Data Mining ICDM*, pp.168-175. Doi: 10.1109/ICDM.2010.137.

Faizan, M. Zuhairi, M.F., Ismail, S. and Sultan, S. (2020). Applications of Clustering Techniques in Data Mining: A Comparative Study. *International Journal of Advanced Computer Science and Applications*, 11(12), pp.146-153.

Fan, C., Xiao, F., Li, Z. and Wang, J. (2018). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. *Energy and Buildings*, 159, pp.296-308.

Farhan, K.A. and Ali, M.A. (2017). Database Protection System Depend on Modified Hash Function. *2nd International Conference of Cihan University-Erbil on Communication Engineering and Computer Science*, p.2520-4777.

Ghavami P. (2020). *Big Data Analytics Methods*. 2nd ed. De Gruyter, Berlin.

Hamarashid, H.K., Saeed, S.A. and Rashid, T.A. (2021). Next word prediction based on the N-gram model for Kurdish Sorani and Kurmanji. *Neural Computing and Applications*, 33(9), pp. 4547-4566.

Harfouchi, F., Habbi, H., Ozturk, C. and Karaboga, D. (2017). Modified multiple search cooperative foraging strategy for improved artificial bee colony optimization with robustness analysis. *Soft Computing A Fusion of Foundations Methodologies and Applications*, 22(19), pp.6371-6394.

Hendrycks, D. Carlini, N., Schulman, J., Steinhardt, J. (2021) Unsolved Problems in ML Safety. ArXiv, Cornell Tech, pp.1-28. Available from: https://arxiv.org/abs/2109.13916

Hong, Z., Smart, G., Dawood, M., Kaita, K., Wen, S.W., Gomes, J. and Wu, J. (2008). Hepatitis C Infection and Survivals of Liver Transplant Patients in Canada, 1997-2003. *Transplantation Proceedings*, 40(5), pp.1466-1470.

Hryshchenko, O. and Yaremenko, V. (2021). A comparative analysis of text data classification accuracy and speed using neural networks, Bloom filter and naive Bayes. *Technology Audit and Production Reserves*, 5(2(61), pp.6-8.

Jain, V. (2017). Perspective analysis of telecommunication fraud detection using data stream analytics and neural network classification based data mining. *International Journal of Information Technology*, 9(3), p.1-8.

Jaouadi, Z., Abbas, T., Morgenthal, G. and Lahemer, T. (2020). Single and multi-objective shape optimization of streamlined bridge decks. *Structural and Multidisciplinary Optimization*, 61(4), pp.1495-1514.

Kan, M.Y. and Klavans, J.L., (2002). Using Librarian Techniques in Automatic Text Summarization for Information Retrieval. *Proceedings of the ACM International Conference on Digital Libraries*, Wuhan, pp.36-45.

Kashyap, R., (2019). Big data analytics challenges and solutions. In: *Big Data Analytics for Intelligent Healthcare Management*, Academic Press, Cambridge, pp.19-41.

Khoshbakht, F., Shiranzaei, A. and Quadri, S.M.K., (2021). Role of the big data analytic framework in business intelligence and its impact : Need and benefits. *Turkish Journal of Computer and Mathematics Education*, 12(10), pp.560-566.

Kumar, D.U., Soon, T.K., Saad, M., Idna, I.M.Y., Mehdi, S. and Bend, H. (2018). Forecasting of photovoltaic power generation and model optimization : A review. *Renewable and Sustainable Energy Reviews*, 81, pp.912-928.

Kumar, J., Singh, A.K. and Buyya, R. (2020). Ensemble learning based predictive framework for virtual machine resource request prediction. *Neurocomputing*, 397, p.20-30.

Li, Z., Zhang, Z., Shi, J. and Wu, D., (2019). Prediction of surface roughness in extrusion-based additive manufacturing with machine learning. *Robotics and Computer Integrated Manufacturing*, 57, pp.488-495.

Liu, X., Liu, X., Zhu, Z. and Hu, L., (2020). An efficient multi-objective optimization method based on the adaptive approximation model of the radial basis function. *Structural and Multidisciplinary Optimization*, 63(4), p.1-19.

Luijken, K., Groenwold, R.H.H., Van Calster, B.E.W., Steyerberg, E.W. and Van Smeden, M. (2019). Impact of predictor measurement heterogeneity

across settings on the performance of prediction models: A measurement error perspective. *Statistics in Medicine*, 38(18), pp.3444-3459.

Luijken, K., Song, J. and Groenwold, R.H.H., (2022). Quantitative prediction error analysis to investigate predictive performance under predictor measurement heterogeneity at model implementation. *Diagnostic and Prognostic Research*, 1, pp.1-11.

Luijken, K., Wynants, L., van Smeden, M., Van Calster, B., Steyerberg, E.W. and Groenwold, R.H.H., (2020). Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *Journal of Clinical Epidemiology*, 119, pp.7-18.

Mariani, M. and Baggio, R. (2022). Big data and analytics in hospitality and tourism: A systematic literature review. *International Journal of Contemporary Hospitality Management*, 34(1), pp.231-278.

McNamara, D.S., Allen, L.K., Crossley, S.A., Dascalu, M. and Perret, C.A., (2017). Natural language processing and learning analytics. In: *Handbook of Learning Analytics*. Ch. 8. Society for Learning Analytics Research, Alberta, pp.93-104.

Mishra, R. and Sharma, R. (2014). Big data opportunities and challenges: Discussions from data analytics perspectives. *International Journal of Computer Science and Mobile Computing*, 46(6), pp.27-35.

Moharram, A., Altamimi, S. and Alshammari, R., (2021). Data Analytics and Predictive Modeling for Appointments No-show at a Tertiary Care Hospital. *2021 1st International Conference on Artificial Intelligence and Data Analytics, CAIDA 2021*, pp.275-277.

Nag, A.K. and Mitra, A., (2002). Forecasting daily foreign exchange rates using genetically optimized neural networks. *Journal of Forecasting*, 21(7), pp.501-511.

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R. and Muharemagic, E., (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), pp.1-21.

Pajouheshnia, R., van Smeden, M., Peelen, L.M., Groenwold, R.H.H., (2019). How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *Journal of Clinical Epidemiology*, 105, pp.136-141.

Patel, A., Singh, N.M. and Kazi, F. (2017). *Internet of Things and Big Data Technologies for Next Generation Healthcare*. Springer Cham, Berlin.

Pourhomayoun, M., Alshurafa, N., Mortazavi, B., Ghasemzadeh, H., Sideris, K., Sadeghi, B., Ong, M., Evangelista, L., Romano, P., Auerbach, A., Kimchi, A. and Sarrafzadeh, M. (2014). Multiple model analytics for adverse event prediction in remote health monitoring systems. In: *2014 IEEE Healthcare Innovation Conference, HIC 2014*, pp.106-110.

Rajaraman, V., (2016). Big data analytics. *Resonance*, 21(8), pp.695-716.

Rajon, S.A.A., Shamim, A. and Arif, M., (2011). A Generic Framework for Implementing Electronic Commerce in Developing Countries. *International Journal of Computer and Information Technology*, 1(2).

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P. and Na, A.Y., (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. ArXiv.

Rajput, A., (2019). Natural language processing, sentiment analysis, and clinical analytics. In: *Innovation in Health Informatics: A Smart Healthcare Primer*, Academic Press, USA, pp.79-97.

Rashid, T.A., Mustafa, A.M. and Saeed, A.M. (2018). Automatic kurdish text classification using KDC 4007 dataset. In: Advances in Internetworking, Data and Web Technologies, Barolli, L., Zhang, M. and Wang, Z., editors. *Lecture Notes on Data Engineering and Communications Technologies*. Vol. 6, Springer, Cham, Berlin, pp.187-198.

Rocha, J.L.M., Zela, M.A.C., Torres, N.I.V. and Medina, G.S. (2021). Analogy of the application of clustering and K-means techniques for the approximation of values of human development indicators. *International Journal of Advanced Computer Science and Applications*, 12(9), pp.526-532.

Russell, S. and Norvig, P. (2020). *Artificial Intelligence a Modern Approach*. 4th ed. Prentice Hall, Hoboken.

Saini, N., Saha, S., Chakraborty, D. and Bhattacharyya, B., (2019). Extractive single document summarization using binary differential evolution optimization of different sentence quality measures. *PLoS One*, 14(11), p.e0223477. [Last accessed on 2022 Apr 01].

Salloum, S., Huang, J.Z., He, Y. and Chen, X. (2018). An asymptotic ensemble learning framework for big data analysis. *IEEE Access*, 7(c), pp.3675-3693.

Sanchez-Gomez, J.M., Vega-Rodriguez, M.A. and C., Perez, C.J. (2019). An Indicator-based Multi-objective optimizationapproach applied to extractive multi-documenttext summarization. *IEEE Latin America Transactions*, 17(8), pp.1291-1299.

Sanchez-Gomez, J.M., Vega-Rodríguez, M.A. and Pérez, C.J. (2018). Extractive multidocument text summarization using a multiobjective artificial bee colony optimization approach. *Knowledge-Based Systems*, 159, pp.1-8.

Schwarz, C., Schwarz, A. and Black, W.C., (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems*, 34(1), pp.1191-1208.

Shamsaldin, A., Rashid, T.A., Fattah, P. and Al-Salihi, N.K., (2019). A study of the convolutional neural networks applications. *UKH Journal of Science and Engineering*, 3(2), pp.31-40.

Shouval, R., Bondi, O., Mishan, H., Shimoni, A., Unger, R. and Nagler, A., (2014). Application of machine learning algorithms for clinical predictive modeling: A data-mining approach in SCT. *Bone Marrow Transplantation*, 49(3), pp.332-337.

Smith, M., Szongott, C., Henne, B. and von Voigt, G., (2013). Big Data Privacy Issues in Public Social Media. *IEEE International Conference on Digital Ecosystems and Technologies* [Preprint].

Talasila, V., Madhubabu, K., Mahadasyam, M.C., Atchala, N.J. and Kande, L.S., (2020). The prediction of diseases using rough set theory with recurrent neural network in big data analytics. *International Journal of Intelligent Engineering and Systems*, 13(5), pp.10-18.

Tran, N.H., Bao, W., Zomaya, A., Nguyen Minh, N.H. and Hong, C.S., (2019). Federated Learning over Wireless Networks: Optimization Model Design and Analysis. *Proceedings-IEEE INFOCOM, 2019-April(1)*, pp.1387-1395.

Vargas-Calderón, V., Ochoa, A.M., Nieto, G.Y.C. and Camargo, J.E., (2021). Machine learning for assessing quality of service in the hospitality sector based on customer reviews. *Information Technology and Tourism*, 23(3), pp.351-379.

Verma, J.P. and Agrawal, S., Patel, P. and Patel, A., (2016). Big data analytics: challenges and applications for text, audio, video, and social media data. *International Journal on Soft Computing Artificial Intelligence and Applications*, 5(1), pp.41-51.

Vu, T., Belussi, A., Migliorini, S. and Eldway, A., (2021). Using deep learning for big spatial data partitioning. *ACM Transactions on Spatial Algorithms and Systems* (*TSAS*), 7(1), pp.1-37.

Wang, J., Tang, Y., Nguyen, M. and Altintas, I., (2014). A Scalable Data Science Workflow Approach for Big Data Bayesian Network Learning. *Proceedings of the 2014 International Symposium on Big Data Computing, BDC 2014*, pp.16-25.

Wang, X.D., Hirsch, C., Kang, S. and Lacor, C., (2011). Multi-objective optimization of turbomachinery using improved NSGA-II and approximation model. *Computer Methods in Applied Mechanics and Engineering*, 200(9-12), pp.883-895.

Yadav, N. and Chatterjee, N., (2016). Text Summarization using Sentiment Analysis for DUC Data. *International Conference on Information Technology*, pp.5.

Yaremenko, V.S., Rogoza, W.S. and Spitkovskyi, V.I., (2021). Application of neural network algorithms and naive bayes for text classification. *Journal of*

Theoretical and Applied Information Technology, 99(1), pp.125-134.

Yu, C.H. Lee, H.S., Lara, E. and Gan, S., (2018). The ensemble and model comparison approaches for big data analytics in social sciences. *Practical Assessment Research and Evaluation*, 23(17).

Zaghloul, M., Salem, M. and Ali-Eldin, A., (2021). A new framework based on features modeling and ensemble learning to predict query performance. *PLoS One*, 16(10), pp.1-18.

Zarchi, M. and Attaran, B., (2019). Improved design of an active landing gear for a passenger aircraft using multi-objective optimization technique. *Structural and Multidisciplinary Optimization*, 59(5), pp.1813-1833.

Zheng, L. and Guo, L., (2020). Application of big data technology in insurance innovation. *Journal of Physics Conference Series*, 1682(1), pp.285-294.

Zhong, W., Yu, N. and Ai, C., (2020). Applying big data based deep learning system to intrusion detection. *Big Data Mining and Analytices*, 3(3), pp.181-195.

Zou, X., et al. (2019). Logistic Regression Model Optimization and Case Analysis, Proceedings of IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT 2019, pp.135-139.

Toxic Metals in Some Decorative Cosmetics and Nail Products: Analysis, Evaluation, and Mitigation

Bashdar I. Meena[1], F. Tahir Tara[2], Shalaw Z. Sedeeq[3] and Khalid N. Sediq[3]

[1]Department of Chemistry, Faculty of Science and Health, Koya University,
Koya KOY45, Kurdistan Region, F.R. Iraq

[2]Department of Medical Microbiology, Faculty of Science and Health, Koya University,
Koya KOY45, Kurdistan Region, F.R. Iraq

[3]Department of Physics, Faculty of Science and Health, Koya University,
Koya KOY45, Kurdistan Region, F.R. Iraq

Abstract—Cosmetic marketing is one of the most profitable and fast increasing markets in Kurdistan Region of Iraq. In recent years, the use of cosmetics has witnessed a rapid increase, especially with the emergence of social media and its impact on this trade. The market is full of different cosmetic brands and nail products. Moderate and low-quality brands of cosmetic samples that available in the local markets were selected to investigate their heavy metals and chemical composition. Samples from face foundation, eye shadow, and nail polish products were taken and examined to evaluate the concentration of metals, that is, Hg, Pb, Cd, As, Mn, Cr, Ni, Co, Fe, Zn, Cu, and Al ions, using X-ray diffraction and X-ray fluorescence techniques. The examination results show high concentrations of Fe and Al metals in the lipstick samples whereas the Hg, Cd, Cr, and Ni were out of detection limit. Moreover, the results show contamination of Hg heavy metal in one of the examined nail polishes brands, whereas the rest of foundation and eye shadow samples show a higher concentration of Al and Fe. Curcumin, as a natural bio-friendly chelate, has been used to deplete metal ions using ultraviolet-visible Spectrophotometer.

Index Terms—Cosmetics, Curcumin, Metal ions, Nail polish, X-ray diffraction, X-ray fluorescence

I. Introduction

Cosmetics are defined according to the U.S. Food and Drug Administration as "articles intended to be rubbed, poured, sprinkled, or sprayed on, introduced into, or otherwise applied to the human body or any part thereof for cleansing, beautifying, promoting attractiveness, or altering the appearance" (Claudia and Giovanni, 2017).

Essentially, cosmetics are classified mainly into leave-on and rinse-off products. Examples of leave-on cosmetics are perfumes, decorative cosmetics, body and face creams, and antiperspirants which are products that are designed for the purpose of spread out on the skin for a quit prolonged time. Meanwhile, the rinse-off cosmetics are products formed to be removed after a short time of stay on the skin or mucous membranes and examples are shampoos, soaps, shower gels, and toothpastes (Claudia and Giovanni, 2017).

Decorative cosmetics handling in addition to pharmaceuticals and many personal care products have been raised especially in the past few years with evolving of social media. These products do not undergo regulations of disinfectants, insect repellents, and dietary supplements; therefore, they need serious consideration due to the pollutants which are continuously discharged into the environment. Large amounts of the decorative cosmetics are used which are related to their environmental impact due to the bioaccumulation ability of these bioactive persistent (Brausch and Rand, 2011, Adepoju-Bello, et al., 2012). Heavy metals assessed in the decorative cosmetics are of substantial heed with different levels as ingredients or contaminants (Al-Dayel, Hefne, and Al-Ajyan, 2011, Ekere, et al., 2014, Golnaz and Parisa 2015; Chen, et al., 2018).

Hazards of the decorative cosmetics to humans are through the ability of the cosmetic molecules to enter the systemic circulation because they can cross the cutaneous barrier when it is applied to the skin or by contact with the mucous membranes or ingestion, for example, lipstick (Wormuth, et al., 2006, Hwang, et al., 2009, Chevillotte, et al., 2014, Ouremi and Ayodele, 2014). Impurities in the cosmetics could be exist as heavy metals (Morra, 2014, Riyadh, 2020). Daily apply of the decorative cosmetics on the skin and nail is enhanced absorption of the heavy metals that pass through hence accumulate in multiple organs and provoke damage of the target organ (Borowska and Brzoska, 2015). Toxic heavy metals are damaged the target organ even with lower level of disclosure and they are recognized as human carcinogens (arsenic, cadmium, chromium, lead, and mercury) (Chauhan, et al., 2010, Borowska and Brzoska, 2015, Muhammed and Hamsa, 2017). Essential heavy metals (manganese, cobalt, and

nickel, iron, zinc, and copper) are important for human organ function in a low limit, and they turn to toxic and harmful metals when they exceed the permissible limit; however, a light metal as aluminum can cause Alzheimer disease when there is a continuous and high exposure to this metal (Lakshmi, Sudhakar, and Prakash, 2015).

Regulations of the compositions and contents of the cosmetics are varied in different countries. Absorbing cosmetic products through skin and disposing the expired products exposes human life and the environment to the risk of pollution and the diseases resulting from it and this urges to have a data base for the toxicological sources in the cosmetic products (make up and nail polish) that are existed in the local market. As well as finding the eco and bio friendly solution to reduce the risk of the harmful metal and heavy ions if they are existed in high concentration.

II. Experimental

A. Apparatus

An energy dispersive X-ray fluorescence (EDXRF), Rigaku NEX CG with RX9, Mo, Cu, and Al targets was used for quantitative measurements of the metal Hg, Pb, Cd, As, Mn, Cr, Ni, Co, Fe, Zn, Cu, and Al ions. A 200 s and 100 s were the X-ray measuring time for the Al target and other targets, respectively.

The chemical compositions and intensity versus energy is measured from the signal outcome of XRF which is processed to a computer program. X-ray diffraction, Rigaku with a Cu target was applied to the samples to identify the type of material as well as its phase and crystalline properties.

Agilent Cary Eclipse ultraviolet (UV)-visible Spectrophotometer, USA, with a quartz cell of 1.0 cm optimal path length was operated for the spectra measurements of the curcumin-metal ions complexes.

B. Material and reagent

All nitrate salts of the metal ions were of analytical grades. Double distilled water was utilized for the preparation. Curcumin plant was dissolved in absolute ethanol after grinding it as a powder.

C. Preparation of the metal ions and samples

A 1×10^{-4} M concentration solution was prepared for the metal ions in a 100 mL volumetric flask using a doubled distilled water. Curcumin plant was grinded first as a yellow powder and 1×10^{-4} M concentration solution was prepared by dissolving it in absolute ethanol, then filtered using Whatman No. 41 membrane filter paper to remove undissolved fibers of the plant and completed the volume to 100 mL in a volumetric flask. Curcumin-metal ion complex was prepared by mixing 3 mL of curcumin ethanolic solution with 1 mL of each metal ion solution, individually and the volume is completed to 25 mL in a volumetric flask by ethanol.

Decorative cosmetic samples of foundations, eye shadows, lipsticks, and nail polishes were collected from the local

market of Kurdistan Region of Iraq. The liquid state sample such as nail polish and lipstick liquid was dried on a microscope slide sheet in an oven of 70°C. The rest of solid samples were applied directly for the analysis. Table I shows the type, brand, and origin country of the decorative cosmetic samples.

III. Results and Discussion

A. XRF analysis

XRF spectroscopy is based on simple relation of the fundamental physics comprising of atom radiation interaction. It is a highly sensitive analytical tool used for the analysis of the metals. The main advantage of XRF is the ability to perform accurate quantitative analysis of a wide range of elements. To determine the contents of metal ions in the cosmetic samples, the levels of heavy, essential, and light metals such as (Hg, Pb, Cd, As, Mn, Cr, Ni, Co, Fe, Zn, Cu, and Al) were evaluated using XRF.

TABLE I
INFORMATION OF THE COLLECTED DECORATIVE COSMETIC SAMPLES

| Type	Sample code	Brand name	Color	Country
Lipstick	LP1	3Q beauty	Pink	China
	LP2	Velvet matte	Pink	India
	LP3	Hudamoji	Red	UK
Nail polish	NP4	Mekyach	Black	PRC
	NP5	Flormar	Red	Turkey
	NP12	Asma beauty	Light brown	PRC
	NP13	Gel Polish	Green	UK
	NP17	Rival	Light red	Germany
Foundation	Fd6	NYX	Light gray	USA
	Fd7	Bonjour	Light pink	Canada
	Fd8	Trois coulerurs	Pink	PRC
	Fd9	MAC	Light pink	Canada
	Fd10	Essence	Light pink	Poland
	Fd11	ZARA	Pink	PRC
Eye shadow	ES14	Cien	Light brown	Germany
	ES15	Statement Eye	Purple	UK
	ES16	Malva	Light navy blue	Ukraine
	ES18	MSYAHO	White	China
	ES19	VIVO	Light brown	Taiwan
	ES20	EMERALD	Bronzy	Germany
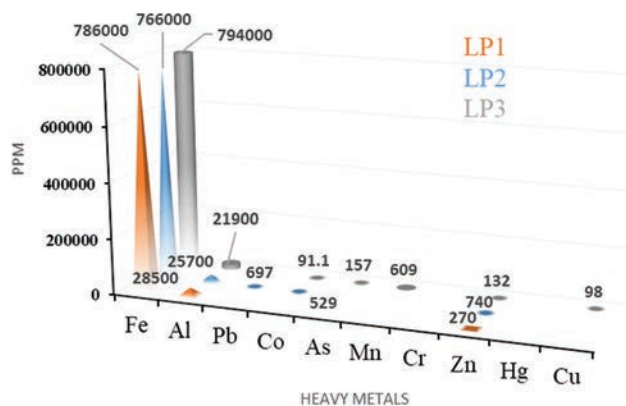
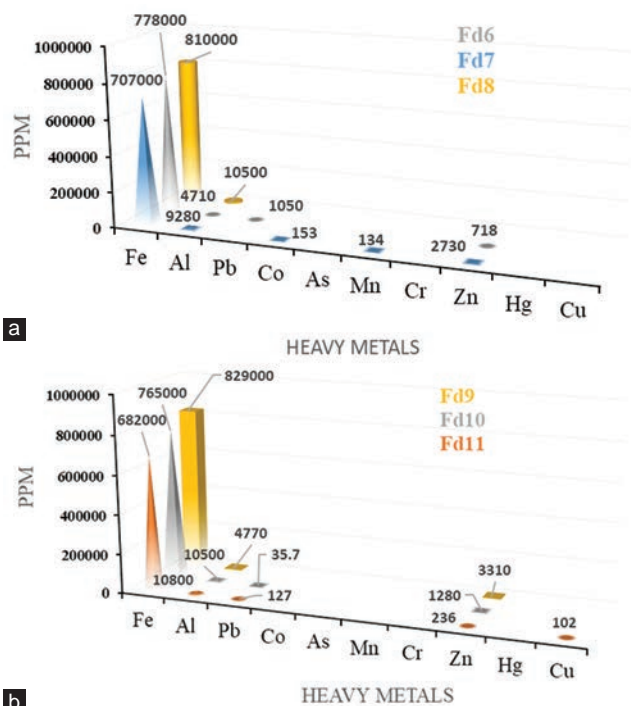Fig. 1. Concentrations of metals in ppm in the lipstick samples (LP1, LP2, and LP3).

Fig. 2. Concentrations of metals in ppm in the foundation samples (a) Fd6, Fd7, and Fd8, (b) Fd9, Fd10, and Fd11.
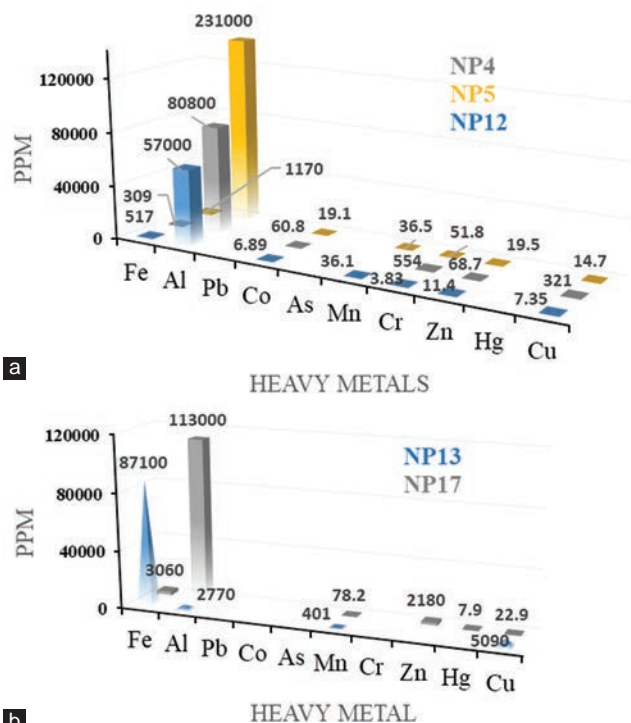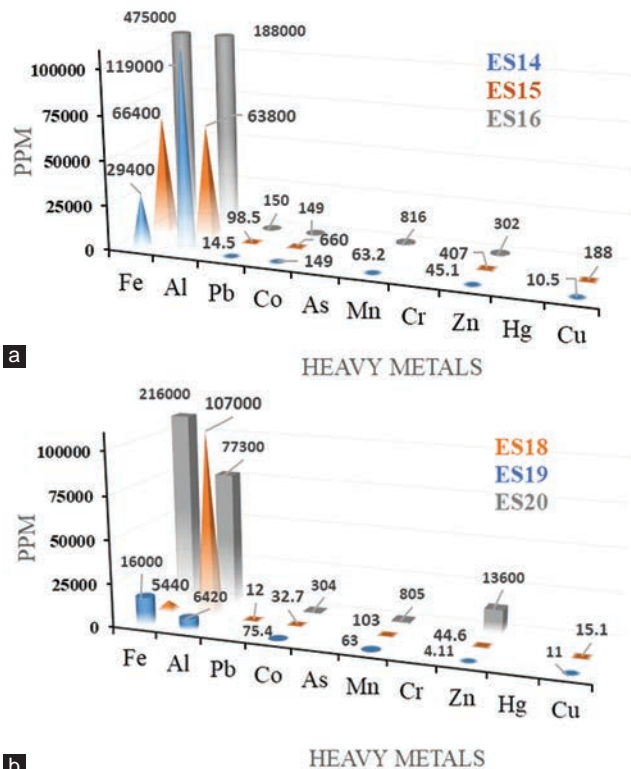

Fig. 3. Concentrations of metals in ppm in the eye shadow samples (a) ES14, ES15, and ES16, (b) ES18, ES19, and ES20.

Fig. 4. Concentrations of metals in ppm in the nail polish samples (a) NP4, NP5, and NP12, (b) NP13 and NP17.

TABLE II
HEAVY METALS ANALYSIS DATE OF THE DECORATIVE COSMETIC SAMPLES USING X-RAY FLORESCENCE (XRF) SPECTROMETER

S/M	Fe	Al	Pb	Co	As	Mn	Cr	Zn	Hg	Cu
LP1	786000	28500						270		
LP2	766000	25700	697	529				740		
LP3	794000	21900		91.1	157	609		132		98
NP4	309	80800		60.8			554	68.7		321
NP5	517	57000		6.89		36.1	3.83	11.4		7.35
NP12	1170	231000		19.1		36.5	51.8	19.5		14.7
NP13	87100	2770				401				5090
NP17	3060	113000				78.2		2180	7.9	22.9
Fd6	778000	4710	1050					718		
Fd7	707000	9280		153		134		2730		
Fd8	810000	10500								
Fd9	829000	4770						3310		
Fd10	765000	10500	35.7					1280		
Fd11	682000	10800	127					236		102
ES14	29400	119000	14.5	149		63.2		45.1		10.5
ES15	66400	63800	98.5	660				407		188
ES16	5440	188000	150	149		816		302		
ES18	5440	107000	12	32.7		103		44.6		15.1
ES19	16000	6420		75.4		63		4.11		11
ES20	475000	77300		304		805		13600		

*LP: Lipstick, NP: Nail polish, Fd: Foundation, ES: Eye shadow

In different brands of lipsticks (LP1, LP2, and LP3), heavy metals such as Hg, Cd, and Cr are not significant in the selected samples. Fig. 1 shows that LP3 depicted the highest concentration of Fe (794000 ppm), and LP2 displayed the lowest concentration (766000 ppm). Concentration of Fe in all lipstick samples is virtually high and contains more than 76% of iron, that is hazards to human health and exceeded the permissible limit according to the WHO. Aluminum is the second high metal ion concentration that reached 28500 ppm in LP1. Toxic heavy metals of Pb and As were assessed only in LP2 and LP3 samples in the concentration of 697 and 157 ppm, respectively.

Foundation samples of (Fd6, Fd7, Fd8, Fd9, Fd10, and Fd11) appraised different results of metal ions as shown in Fig. 2a and b. Among different brands of foundation samples Fe, was the highest concentration (829000 ppm), in Fd9 sample and the lowest concentration (682000 ppm) carried by Fd11 sample. High concentration of aluminum was observed in all samples that out passed the allowable limit. Al concentration range in the foundation samples was between (10800-4710) ppm (Table II). The lone toxic metal ion detected in some foundation samples was Pb as in Fd6 (1050 ppm), Fd11 (127 ppm), and Fd10 (35.7 ppm), however, their value is beyond the 10 ppm limit of Pb which set by the WHO. The rest of the noxious heavy metals Hg, As, and Cr vanished in all foundation samples.

Like foundation samples there was no trace of the poison heavy metals Hg, As, and Cr found in the eye shadow samples (Fig. 3a and b). The metal ions of Fe, Al, Zn, Mn, Co, and Cu were detected in all the investigated cosmetic samples (ES14, ES15, ES16, ES18, ES19, and ES20). Data presented in Table II reveal that the iron concentration in sample ES16 had the highest (475000 ppm) whereas in ES18 and ES14 had the lowest (5440 ppm). In all eye shadow samples the concentration of Al was very high in the range of (188000-6420) ppm in

the samples ES16 and ES19, respectively. The presence of the poison metal Pb was detected in the samples ES14, ES15, ES16, and ES18 that reached 150 ppm although Pb level was out of detection limit in the samples ES19 and ES20.

Unlike makeup samples highest concentration of the measurable metal ions was for Al in all nail polish samples as shown in Fig. 4a and b. The highest Al concentrations in the nail polish samples (NP4, NP5, NP12, NP13, and NP17) reached 231000 ppm in NP12 verses the lowest concentration of 2770 ppm in NP13 sample. The main health problem of aluminum toxicity is neurotoxicity effects such as neuronal atrophy in the locus ceruleus, striatum and substantia nigra (Lakshmi, Sudhakar and Prakash, 2015). Fe concentration level ranged from 309 to 87100 ppm. The highest concentration of Cr and Cu 554 ppm and 5090 ppm were detected in NP4 and NP13, respectively. Mercury metal ion was contaminated within one sample NP17 of 7.9 ppm concentration at the same time Pb and As abandoned in all nail polish samples (Table II).

B. X-ray diffraction (XRD) analysis

XRD is one of the powerful techniques applied to the materials analysis and interpretation of inorganic mineralized artifacts such as metal, ceramics, and pigments. It is a non-destructive and
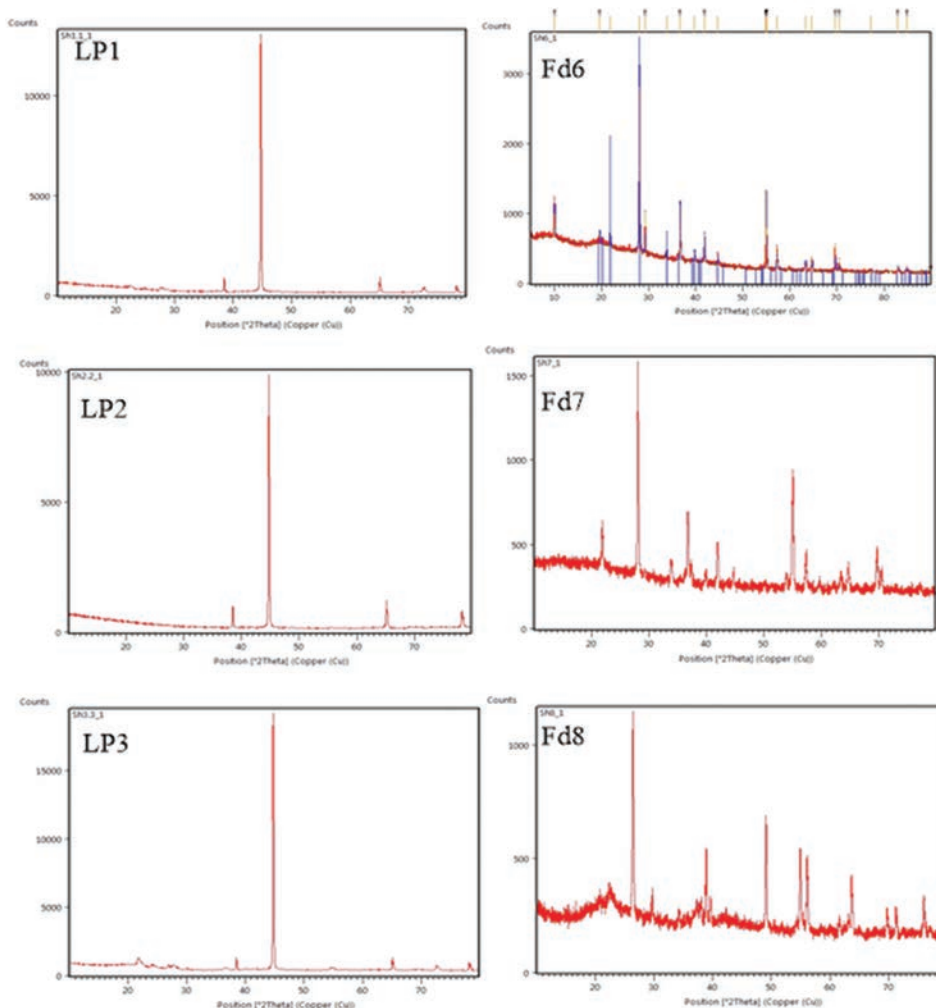
Fig. 5. X-ray diffraction patterns of some cosmetic samples.
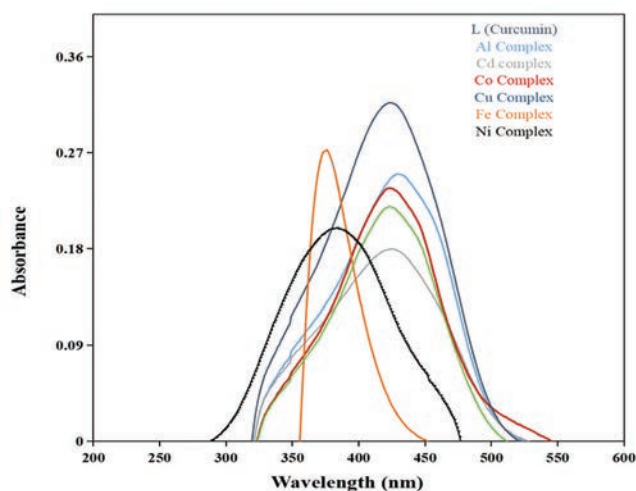
Fig. 6. Comparative ultraviolet-visible spectra of curcumin and its metal complexes.

versatile analytical technique that can rapidly get elaborate phase information and structure of materials. XRD patterns of materials in Fig. 5 shows the characteristic diffraction peaks in the samples LP1, LP2, and PL3 that indicated the major component of the sample were Fe and Al and Fe was almost 80% composed of the chemical composition. The samples Fd7, Fd8, and Fd9 show the same XRD patterns which means that they contain same chemical composition and the majority consist of (Al, 75%), which confirmed by the XRF technique.

C. UV-visible spectra

A vital step to reduce the high concentration of the metal ions during the analysis of cosmetic samples is to have a powerful natural chelating agent as curcumin with its safety assessment even distributed at the high doses in humans (Subhan, et al., 2014). The complexation of curcumin with the most metals has attracted interest over the past year; all the complexes are synthesized by mixing curcumin with metal ions at a molar ratio 3:1 in an ethanolic solution (Subhan, et al., 2014). UV-Visible spectrophotometric technique is the adequate and simplest method to prove the binding and complexation of curcumin with the metal ion.

The electronic spectra of the curcumin were recorded in ethanol and comparative electronic absorption spectra of curcumin with its metal complexes in ethanol are shown in Fig. 6. Curcumin ligand showed that the main absorption band at 420–434 nm, which is indicated to π-π* transition. The spectrum of the curcumin compared with its complex's spectra where most of the absorption spectra shifted by (2–35 nm) and indicated the participation of the carbonyl group of curcumin in metal complexation. The variation of the absorption peak of curcumin and all complexes depends on the nature of metal (M⁺) ion.

IV. Conclusion

High accuracy techniques such as XRF and XRD were used to measure the concentrations of Hg, Pb, Cd, As, Mn, Cr,

Ni, Co, Fe, Zn, Cu, and Al in cosmetic samples such as, lipstick, face foundation, eye shadow, and nail polish that were collected from the local market of the region. The essential findings of this work are that concentrations of Fe and Al were found in most of cosmetic samples in higher level than the safe allowable limits. Fortunately, heavy metals of Cd, Hg, and As were not existed in risky level in all samples. Curcumin was used as a natural bidentate ligand to coordinate with various heavy metals, which is a natural and bio-friendly solution to reduce the risk of the high concentrations of the metals.

Declaration

We declare that this study is not to defame any product or harm the reputation of the product.

References

Adepoju-Bello, A.A., Oguntibeju, O.O., Adebisi, R.A. Okpala, N. and Coker, H.A.B. 2012. Evaluation of the concentration of toxic metals in cosmetic products in Nigeria. *African Journal of Biotechnology*, 11(97), pp.16360-16364.

Al-Dayel, O., Hefne, J. and Al-Ajyan, T. 2011. Human exposure to heavy metals from cosmetics. *Oriental Journal of Chemistry*, 27, pp.1-11.

Borowska, S. and Brzoska, M.M. 2015. Metals in cosmetics: Implication for human health. *Journal of Applied Toxicology*, 35(6), pp.551-572.

Brausch, J.M. and Rand, G.M. 2011. A review of personal care products in the aquatic environment: Environmental concentrations and toxicity. *Chemosphere*, 82(11), pp.1518-1532.

Chauhan, A.S., Bhadauria, R., Singh, A.K., Lodhi, S.S., Chaturvedi, D.K. and Tomar, V.S. 2010. Determination of lead and cadmium in cosmetic products. *Journal of Chemical and Pharmaceutical Research*, 2(6), pp.92-97.

Chen, X., Sullivan, D.A., Sullivan, A.G., Kam, W.R. and Liu, Y. 2018. Toxicity of cosmetic preservatives on human ocular surface and adnexal cells. *Experimental Eye Research*, 170, pp. 188-197.

Chevillotte, G.C., Ficheux, A.S., Morisset, T. and Roudot, A.C. 2014. Exposure method development for risk assessment to cosmetic products using a standard composition. *Food and Chemical Toxicology*, 68, pp.108-116.

Claudia, J. and Giovanni, A.M. 2017. Review: Cosmetic ingredients as emerging pollutants of environmental and health concern. *Cosmetics*, 4(11), pp. 2-18.

Ekere, N.R., Ihedioha, J.N., Oparanozie, T.I., Ogbuefi-Chima, F.I. and Ayogu, J. 2014. Assessment of some heavy metals in facial cosmetic products. *Journal of Chemical and Pharmaceutical Research*, 6(8), pp.561-564.

Golnaz, K. and Parisa, Z. 2015. Heavy metal contamination of popular nail polishes in Iran. *Iranian Journal of Toxicology*, 9(29), pp.1290-1295.

Hwang, M., Yoon, W.K., Kim, J.Y., Son, B.K., Yang, S.J., Yun, M.O., Choi, S.S., Jang, D.D. and Yoo, T.M. 2009. Safety assessment of chromium by exposure from cosmetic products. *Archives of Pharmacal Research*, 32, pp.235-241.

Lakshmi, B.V.S., Sudhakar, M. and Prakash, K.S. 2015. Protective effect of selenium against aluminum chloride-induced Alzheimer's disease: Behavioral and biochemical alterations in rats. *Biological Trace Element Research*, 165(1), pp.67-74.

Morra, O.Z. 2014. *Levels of selected heavy metals in Aloe vera skin cosmetics*. A Thesis School of Pure and Applied Sciences. Kenya: Kenyatta University. pp.69.

Muhammed, N.A.A. and Hamsa, T.A.A. 2017. Detection of cadmium and chromium in some facial cosmetics. *Iraqi Journal of Science*, 58(1), pp.408-416.

Ouremi, O.I. and Ayodele, O.E. 2014. Lipsticks and nail polishes: Potential sources of heavy metal in human body. *International Journal of Pharmaceutical Research and Allied Sciences*, 3(4), pp.45-51.

Riyadh, M.J. 2020. Determination of some heavy metals in selected cosmetic products sold out at Iraqi markets. *Systematic Reviews in Pharmacy*, 11(12), pp.1632-1635.

Subhan, M.A., Alam, K., Rahman, M.S., Rahman, M.A. and Awal, M.R. 2014. Synthesis and characterization of metal complexes containing curcumin $C_{21}H_{20}O_6$ and study of their anti-microbial activities and DNA bonding properties. *Journal of Scientific Research*, 6(1), pp.97-109.

Wormuth, M., Scheringer, M., Vollnweider, M. and HurgerbÜhler, K. 2006. What are the sources of exposure to eight frequently used phthalic acid esters in European? *Risk Analysis*, 26(3), pp.803-824.

Determination of the Potassium Content in Fruit Samples by Gamma Spectrometry to Emphasize its Health Implications

Dedawan S. Saleh[1]*, Saddon T. Ahmad[2] and Sarmad R. Kareem[2]

[1]Department of Physics, Faculty of Science and Health, Koya University,
Koya KOY45, Kurdistan Region - F.R., Iraq

[2]Faculty of Medicine, Koya University,
Koya KOY45, Kurdistan Region - F.R., Iraq

Abstract—In this study, the activity concentration of ^{40}K and its' concentrations in 24 different types of fruits were determined using high purity germanium and sodium iodide scintillation (NaI) detectors. The results of the two measurements are consistent. The maximum and minimum activities of ^{40}K in dry samples were 750.61 ± 11.88 and 15.64 ± 0.86 Bq/kg in apricot and olive, respectively, while in fresh samples they were 152.27 ± 2.12 and 1.99 ± 0.11 Bq/kg in dates and olive, respectively. The highest and lowest potassium contents were 489.81 and 6.42 mg/100 g in fresh dates and olives, respectively. Drupe and tropical fruits, as a fruit family, typically had the highest level of ^{40}K activity and potassium concentration, whereas pome fruits showed the lowest levels. Many of these commonly consumed fresh fruits with rich potassium and water contents are lowering hypertension and improving the hydration status in people's nutrition. The rate of potassium-40 and total potassium concentration intake for a single unit or portion of the fruits was calculated.

Index Terms—^{40}K, Health, HPGe, NaI, Fruits, Potassium

I. Introduction

Potassium is a delicate, silver-white metal that is abundant in nature and found in every tissue of plants and animals. Three isotopes of the potassium element make up the element: ^{39}K (93.3%), ^{40}K (0.012%, 120 ppm), and ^{41}K (6.7%). While potassium-40 is radioactive with a half-life of 1.28×10^9 years (one of the few naturally occurring radioisotopes with an atomic number below 82), potassium-39 and potassium-41 are stable isotopes to radioactive decay (Peterson, 1996). One of the most prominent naturally occurring radionuclides is potassium-40 (Tolstykh, et al., 2016). Naturally occurring radioactive nuclides can be found in human habitats such as soil, water, food, and air, as well as in our bodies (Aswood, Jaafar, and Bauk, 2013). These radioactive nuclides can enter plants through their roots or leaves, where they are subsequently directly absorbed by humans as food (Alharbi and El-Taher, 2013).

While 89.3% of the ^{40}K decay through beta-emission to the ground state of calcium without gamma emission, the remaining 10.67% of the ^{40}K will decay by electron capture to an excited state of argon, resulting in a 1460.8 keV gamma decay to the ground state. The measurement of 1460.8 keV serves as a signature and verification for potassium-40 (Arena, 1969; Leutz, Schulz and Wenninger, 1965). Both internal and external health risks are present with ^{40}K. External exposure to this radioisotope raises concerns due to its intense gamma radiation (E = 1460.8 keV); While inside the body, ^{40}K poses a health risk due to the emission of beta particles (E_{max} = 1.35 MeV) and gamma rays, which are linked to cell damage and the potential to cause cancer in the future (Sarayegord, et al., 2009; Santos Jr., et al., 2005).

Numerous clinical and observational studies have demonstrated the risk of hypertension being reduced by consuming less sodium and more potassium. An increased potassium consumption lowers blood pressure in hypertensive populations (Perez and Chang, 2014). Hypertension (high blood pressure) increases a person's risk of cardiovascular disease (CVD) (Steddon, et al., 2014). Blood pressure ≥140 mm Hg systolic and 90 mm Hg diastolic (>140/90 mmHg) can be classified as hypertension. Aside from increasing the risk of CVD, hypertension has been linked to stroke and other chronic conditions such as cerebrovascular disease, ischemic heart disease, and renal failure (Pawloski, 2015). The World Health Organization (WHO) strongly recommends increasing potassium intake to at least 90 mmol/day (3.5 g/day) in adults to reduce blood pressure and the risk of cardiovascular disease (CVD), stroke, and coronary heart disease (CHD) (Burnier, 2019). Although most intervention studies have focused on high levels of potassium intake, observational studies show that increasing potassium intake by 750–1000 mg/day can reduce

blood pressure by 2–3 mm Hg. This reduction in blood pressure (BP) benefits the cardiovascular system by lowering the risk of cerebral vascular accidents (CVA) and other CVD events. Reduced sodium intake, increased calcium, potassium, and magnesium intake, a diet rich in fruits and vegetables, various antioxidants, and other "designer foods" and nutritional supplements are all known to the lower blood pressure (Houston, 2011).

In the literature, several techniques have been used by researchers to determine the potassium content in several types of food. One of these is the gamma measurements of ^{40}K which is a clean, accurate, and, low-cost method. Some used the high purity germanium (HPGe) detectors in food measurements (Abt, et al., 2016; Acquah and Pooko-Aikins, 2013), others used Sodium Iodide Based Detectors (Hoeling, Reed and Siegel, 1999; Escareño-Juárez and Vega-Carrillo, 2012). The Fluorimeters technique has been implemented to obtain potassium content in milk again, through the measurement of the ^{40}K (Emumejaye, 2012). Measurements of the potassium content of food samples also have been performed using atomic absorption spectrometers (Vaessen and van de Kamp, 1989).

The current study aims to estimate the ^{40}K activity concentration and total potassium contents in 24 sample fruits. These fruits are regularly consumed by residents of the Koya district. Knowing the potassium content of these fruits allows for making better recommendations for a better nutritional system that helps control blood pressure.

II. Materials and Methods

A. Samples collection and preparation

Twenty-four types of fruits were collected from local markets in Koya city. It is about 57.32 km far from Erbil. Erbil is the capital of Kurdistan Region-Iraq. Around 100 thousand inhabitants live in this area. As shown in Table 1, some of these fruits are grown locally, while others are from different countries. Except for date fruit, which was produced in the south of Iraq, other local fruits were grown nearby Koya city.

The edible portions of fruits are retained for analysis after being washed with tap water to remove soil and other unwanted material (Al-Masri, et al., 2004). The resulting fresh mass was then calculated by weighing and carefully distributed for initial drying after being cut into little pieces with a knife. Fruit samples were dried in an electric oven at about 80 °C until they reached a constant weight (Al-Absi, et al., 2015). Samples were crushed and processed into fine powder in the laboratory using an automatic grinder with an adjustable squeezer. The duration of the crushing process has been selected to guarantee homogeneous sample particle size. The components were crushed and then sieved through a sieve with a 35-mesh screen (the hole diameter is 0.5 mm). The same level from each sample was packed in a Marinelli beaker (one-liter size) for both gamma spectrometers. The net weights of samples were determined from the difference between the weights of sample-filled and empty beakers (Pourimani, Noori and Madadi, 2015; Al-Absi, et al., 2015). The codes for the samples were SF for fruits, each followed by a number ranging from SF1 for the date to SF24 for Olive fruit samples (Table I).

Table I presents fruit samples by common name, scientific name, family name, numbering codes, and their countries location. In addition, the dry to fresh weight percentage ratio (DFWR), and the water content percentage. The fresh weight of samples is the weight required to produce 1 kg of dried sample.

B. Gamma detection systems

Two gamma spectrometers were used to measure the activity concentration of ^{40}K in the fruits, the first using HPGe detector and the second using a thallium-activated sodium iodide scintillation (NaI) detector (Salih, Hussein and Sedeeq, 2019; Najam, Tawfiq and Kitha, 2015).

HPGe detector

The HPGe detector is a coaxial, p-type device manufactured by Princeton Gamma Tech (PGT). The diameter and length of the crystal are 70.6 and 70.7 mm, respectively. At the ^{60}Co gamma energy of 1332 keV, the detector resolution is approximately 1.96 keV. The detector's relative efficiency is 73% and has a peak-to-Compton ratio of 75/1. This performance is due to the crystal size, which has an active volume of 265 cm^3 (Ahmad, Almuhsin, and Hamad, 2021). The system is housed within a 10.1 cm high-performance copper/tin lined lead shield Kolga Model A340 cylindrical lead shield. This shield prevents background radioactivity. For about 10 h, the samples were placed over the detector. An empty sealed beaker was counted in the same way and with the same geometry as the samples to determine the background radiation in the surrounding environment around the detector. The counting time for each sample was 36000 s. During operation, the detector was cooled down by liquid nitrogen at 77 K to reduce leakage current and thermal noise prevention, and its warm-up sensor is coupled to the high voltage detector bias supply. The operating voltage was 4000V.

The spectra were analyzed with a Thermo Scientific System 8000 multi-channel analyzer and PGT's Quantum Gold 2001 computer software. The activity concentration of ^{40}K was directly determined using a 1460.5 keV (10.67%) gamma-ray line after measurements and correction of the net peak area of gamma rays of measured samples by background subtraction (Azeez, Ahmad, and Mansour, 2018).

Sodium iodide scintillation NaI(Tl) detector

The scintillator and the photomultiplier tube that make up the NaI (Tl) detector are optically connected. The sodium iodide crystal [NaI(Tl)] employed in the current work is a thallium-activated scintillator. With a 2 × 2-inch active area, the NaI(Tl) scintillation detector offers an energy resolution of 8% at the 662 keV Cs-137 line and 5.91% at the 1173 keV Co-60 line. This study made use of the MAC-CASSY

TABLE 1
SAMPLE IDENTIFICATION AND THE WATER CONTENT PERCENTAGE FOR SAMPLES

Samples	Family name	Location	Scientific name	Sample code	Fresh samples (kg)	DFWR %	Water %
Date	Drupe	Iraq	Phoenix dactylifera	SF1	1.28	78.1	21.9
Avocado	Tropical	Lebanon	Persea americana	SF2	4.48	22.3	77.7*
Banana	Tropical	Africa	Musa sepientum	SF3	3.97	25.1	74.9
Apricot	Drupe	Iraq	Prunus armeniaca	SF4	9.68	10.3	89.7
Pomegranate	Tropical	Iraq	Punica granatum	SF5	4.8	20.8	79.2*
Mulberry	Berries	Iraq	Morus alba	SF6	5.0	20.0	80.0
Fig	Tropical	Iraq	Ficus carica	SF7	3.99	25.0	75.0
Peach	Drupe	Iran	Prunus persica	SF8	8.2	12.2	87.8
Melon	Melons	Iraq	Cantalupensis	SF9	13.4	7.4	92.6
Kiwifruit	Tropical	Iran	Actinidia deliciosa	SF10	6.35	15.7	84.3
Strawberry	Berries	Iran	Fragaria ananassa	SF11	14.1	7.1	92.9
Grape	Berries	Iraq	Vitis vinifera	SF12	10.8	9.2	90.8
Mango	Drupe	Tanzania	Mangifera	SF13	6.9	14.4	85.6
Orange	Citrus	Turkey	Citrus sinensis	SF14	7.25	13.8	86.2
Plum	Drupe	Iran	Prunus domestica	SF15	9.6	10.4	89.6
Quince	Pome	Turkey	Cydonia oblonga	SF16	7.1	14.1	85.9
Lemon	Citrus	Turkey	Citrus limon	SF17	7.44	13.4	86.6
Grapefruit	Citrus	Turkey	Citrus paradisi	SF18	8.95	11.2	88.8
Pineapple	Tropical	Costa Rica	Ananas comosus	SF19	10.45	9.5	90.5
Pear	Pome	Iraq	Pyrus communis	SF20	7.8	12.8	87.2
Green apple	Pome	Iran	Malus domestica	SF21	8.4	11.9	88.1
Apple	Pome	Iraq	Malus pumila	SF22	7.5	13.3	86.7
Watermelon	Melons	Iraq	Citrus vulgaris	SF23	22.44	4.4	95.6
Olive	Drupe	Iraq	Olea europaea	SF24	7.84	12.7	87.3*

*Avocado and olive contain around 10% fat (oil), and Pomegranate contains about 1%, while the others contain<0.5%

computer-assisted science system. The background of the laboratory environment was measured by placing an empty Marinelli beaker on the detector with the same period and operating voltage used to determine the energy spectra of the samples. The operating voltage was 600V. The net sample spectra are calculated by subtracting the background from the sample measurements. A (1.5 cm) lead shield surrounds the detector to lessen the environmental background (Hussain and Hussain, 2011). Particularly at higher energy, NaI (Tl) is more efficient than solid-state detectors at detecting gamma radiation. Because it does not require cooling, the NaI (Tl) detector is easy to utilize in environmental research than a solid-state detector (Htwe and Lwin, 2010). The counting time of each sample was 18000 s.

III. THEORY

A. Activity concentrations

The activity concentrations (A) of potassium ^{40}K, with an energy of 1.46 MeV, were measured in fruit and vegetable samples as follows (Azeez, Mansour, and Ahmad, 2019; Smail, Ahmad and Mansour, 2022).

$$A = \frac{N}{\varepsilon \times I_\gamma \times T \times M} \tag{1}$$

Where N is the net count, ε is the absolute gamma peak efficiency for the detector at a specific gamma-ray energy, and Iγ is the decay intensity of the 1.46 MeV of ^{40}K. T is the counting time of the measurement in seconds and M is

the mass of the sample in kilograms. The relative combined standard deviation σ_A of the activity concentration is given by the formula [2]:

$$\frac{\sigma_A}{A} = \sqrt{\frac{\sigma_N^2}{N} + \frac{\sigma_\varepsilon^2}{\varepsilon} + \frac{\sigma_{I_\gamma}^2}{I_\gamma} + \frac{\sigma_M^2}{M}} \tag{2}$$

Where σ_N is the standard deviation of the N net count rate per second, σ^ε, σ_I^γ, and σ_M are the standard deviations of the ε, I$^\gamma$, and M, respectively (Smail, Ahmad and Mansour, 2022).

B. Energy calibration of the detection systems

The object of energy calibration is to derive a relationship between the peak position in the spectrum and the corresponding gamma-ray energy.

For the HPGe spectrometer, the standard point gamma-ray sources, such as ^{60}Co (peaks 1173.2 and 1332.5 keV), ^{137}Cs (peak 661.7 keV), and ^{226}Ra (peaks 295, 609, and 1120 and 1764 keV) were used to calibrate the energy of the gamma-ray.

Fig. 1 shows the plotted graph for the energy against channel number for the high pure germanium detector.

For the NaI spectrometer, the gamma-ray energy was calibrated by using the standard point gamma-ray sources, such as ^{60}Co (peaks 1173.2 and 1332.5keV), ^{137}Cs (peak 661.7 keV), and ^{22}Na (peaks 511, and 1275 keV).

Fig. 2 shows the plotted graph for the energy against channel number for the scintillation (NaI) detector.

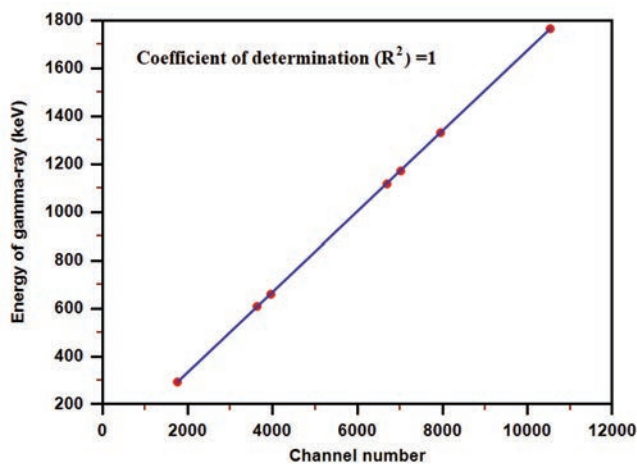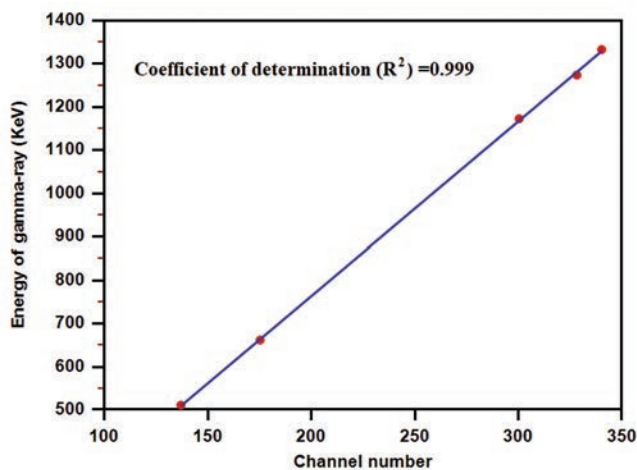Fig. 1. Energy calibration graph for gamma-ray spectrometry system (HPGe).

Fig. 2. Energy calibration graph for gamma-ray spectrometry system (NaI).

C. Efficiency calibration of the detection systems

Four gamma-ray sources, ^{226}Ra, ^{60}Co, ^{22}Na, and ^{137}Cs, were used to calibrate the efficiencies of the two systems. For at least 1 h, the spectra of these sources were counted by each detector. A sample of 713 g of potassium chloride (KCl) was employed as a ^{40}K radioactive source to determine the precise efficiency for the 1460 keV gamma photon of ^{40}K, For the HPGe detection system, the fitted data provide the following polynomial equation:

$$\varepsilon = 0.1158 - 3 \times 10^{-4} E_\gamma + 3 \times 10^{-7} E_\gamma^2 - 1 \times 10^{-10} E_\gamma^3 \tag{3}$$

Fig. 3 shows the plotted graph of the absolute full peak efficiency through the gamma-ray energy for the high pure germanium detector.

While the fitted data for the NaI(Tl) detection system gives the polynomial equation (4) and its plotted graph is shown in Fig. 4:

$$\varepsilon = 0.0291 - 5 \times 10^{-5} E_\gamma + 3 \times 10^{-8} E_\gamma^2 - 7 \times 10^{-12} E_\gamma^3 \tag{4}$$

Where ε is the absolute full peak efficiency of the detector, and E_γ is the energy of the gamma-ray in both equations.
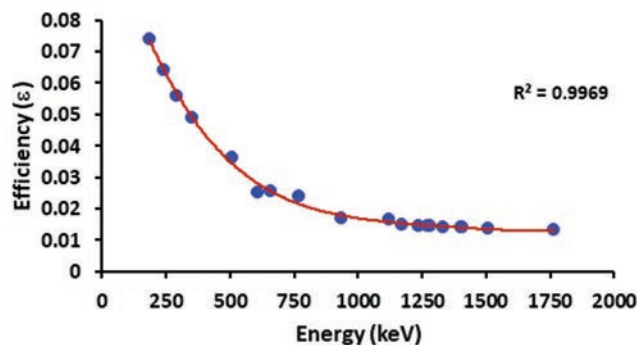
Fig. 3. The efficiency curve of the HPGe gamma spectrometer.
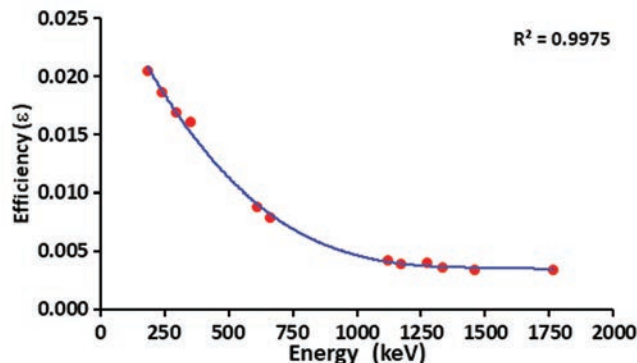
Fig. 4. The efficiency curve of the NaI(Tl) gamma spectrometer.

D. Minimum detectable activity

Currie's derivation is used to determine the minimum detectable activity (MDA) for any given nuclide at a 95% confidence level (L'Annunziata, 2012; Currie, 1968).

$$\varepsilon = 0.0291 - 5 \times 10^{-5} E_\gamma + 3 \times 10^{-8} E_\gamma^2 - 7 \times 10^{-12} E_\gamma^3 \tag{5}$$

Where σ is the standard deviation of the collected background over the energy range of interest during time T. T denotes the collection time (s), and ε is the system efficiency at the specific energy. I_r is the decay intensity of the specified energy peak. M is the mass of the sample (kg). The MDA for a ^{40}K nuclide in the present work is around 3.6 Bq/kg.

E. Total potassium concentration (K)

Activity concentration of ^{40}K, calculated from its gamma emission in the photo peak of 1460 keV, can be used and converted into total potassium concentration in the sample, as follows (Tzortzis and Tsertos, 2004; Tin and Wai, 1993).

$$m = A \times \frac{t_{1/2}}{0.693} \times \frac{M_w}{N_A} x100\% \tag{6}$$

Where m is the percentage (%) fraction of potassium (^{40}K) in the sample, M_w, $t_{1/2}$, and A are the atomic mass (kg mole^{-1}), the half-life of ^{40}K (s), and the measured activity concentration (Bq/kg), respectively, ^{NA} is Avogadro's number (6.022×10^{23} atoms mole^{-1}). The total mass of the potassium in the sample can be determined from the ^{40}K

mass, which is only about 0.012% of the total potassium mass.

F. Daily intakes of ^{40}K radioactivity

The radionuclide intakes can be calculated from the measured concentration of ^{40}K radionuclide in fresh fruit or vegetable (Bq/kg) and the daily consumption rate (kg/day) as follows.

$$Daily\,intake\,(mBq) = A \times DCR \times f \qquad (7)$$

Where A is the activity concentration of the sample in mBq/kg, *DCR* is the daily consumption rate in kg/day and *f* is the number of days the fruit is consumed in a week divided by 7 (days in a week) (Al-Absi, et al., 2015; Adedokun, et al., 2019).

IV. Results and Discussion

A. Potassium contents

One of the gained results of this study is the estimation of the percentage of water in studied fruits. Table I shows that the minimum water content is in the SF1 date fruit sample at 21.7%, and the maximum is in SF23 the fruit watermelon sample at 95.6%. Of twenty-four types, about 5 have a water content above 90%. The water content of fourteen types of fruits is between 80% and 90%. Four samples are in the range 70–80%, which leaves only one type with water content below 70%. Most of the fruits are grown locally or regional and are regularly consumed.

It is worth noting that there are some of these samples are oily fruits and include a negligible percentage of oil, olives and avocados containing approximately 10%, while pomegranates contain 1%. Some of the rest are as low as 0.5%.

This distribution of fruits regarding their percentage of water is a healthy sign for those who eat fruits regularly and in appropriate quantities, especially children. One of the studies conducted especially for this purpose confirms that regular intake of fruits may relevantly improve hydration status (HS) in children. Dietary interventions to increase fruit intake may be a promising strategy to achieve a positive water balance in this population (Montenegro-Bethancourt, Johner and Remer, 2013).

The activity concentration of potassium ^{40}K in (Bq/kg) for dried 24 fruit samples was studied using HPGe detector and thallium activated sodium iodide scintillation detector NaI(Tl). The results of the two systems are in good agreement, and the differences are within the limit of the statistical errors. In Fig. 5, the scatter diagram with the regression line of the two systems measurements is presented, with the determination coefficient being $R^2 = 0.9955$. Therefore, only the results of the HPGe spectrometer were discussed and treated here to avoid repetitions. However, Tables II and III contain both system's measurements for fruits.

Using gamma spectroscopy to measure the ^{40}K activity concentration in food is a perfect technique to estimate the
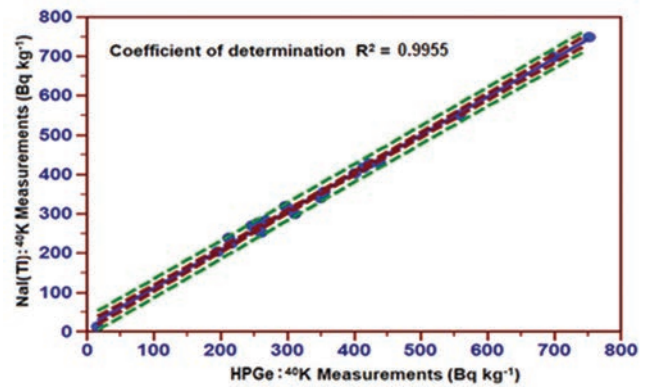

Fig. 5. The scatter diagram with the regression line of the two measurements.

TABLE II
ACTIVITY CONCENTRATION OF ^{40}K (BQ/KG) IN DRIED AND FRESH FRUITS

Fruit samples	Dried fruits		Fresh fruits	
	HPGe	NaI (Tl)	HPGe	NaI (Tl)
Date	194.91±2.71	204.30±1.70	152.27±2.12	159.61±1.33
Avocado	400.85±4.66	402.83±3.29	89.47±1.04	89.91±0.73
Banana	349.45±3.46	340.50±2.73	88.02±0.87	85.77±0.69
Apricot	750.61±4.60	749.63±5.39	77.54±0.47	77.43±0.56
Pomegranate	296.06±3.56	320.48±2.62	61.68±0.74	66.76±0.55
Mulberry	300.76±3.19	313.76±2.53	60.15±0.64	62.75±0.51
Fig	215.39±2.65	225.12±1.86	54.98±0.66	56.42±0.47
Peach	437.85±4.45	429.68±3.46	53.39±0.54	52.40±0.42
Melon	698.11±5.25	698.97±5.31	52.09±0.39	52.16±0.39
Kiwi	255.67±3.15	277.27±2.30	40.26±0.49	43.66±0.36
Strawberry	557.73±5.47	551.49±4.42	39.55±0.38	39.11±0.31
Grape	422.25±4.26	428.30±3.44	39.09±0.39	39.63±0.32
Mango	255.67±3.42	268.38±2.21	37.05±0.49	38.89±0.32
Orange	267.34±3.50	282.28±2.34	36.87±0.48	38.93±0.32
Plum	353.88±3.60	356.19±2.86	36.86±0.37	37.10±0.29
Quince	252.76±3.02	260.26±2.13	35.60±0.42	36.65±0.30
Lemon	258.27±3.05	274.53±2.30	34.71±0.41	36.89±0.31
Grapefruit	310.50±3.36	300.09±2.44	34.69±0.37	33.53±0.27
pineapple	351.48±3.98	349.84±2.86	33.63±0.38	33.47±0.27
Pear	261.12±3.71	253.70±2.12	33.47±0.47	32.52±0.27
Green apple	246.56±3.45	270.37±2.25	29.35±0.41	32.18±0.26
Apple	212.12±3.07	238.85±1.99	28.28±0.41	31.84±0.26
Watermelon	413.28±4.07	417.71±3.34	18.41±0.18	18.61±0.15
Olive	15.64±0.86	13.22±0.11	1.99±0.11	1.68±0.02
Average	336.6±3.6	342.82±2.75	48.72±0.55	49.91±0.4

radiological parameters and an efficient method to estimate the total potassium in food samples, especially those which contain a high ratio of light elements H, C, N, and O, such as the carbohydrates. These light elements are out of the lower limit of detection in the XRF technique. Therefore, the results of the XRF may lead to an overestimation of Potassium content.

The ^{40}K activity concentration in dried and fresh fruit samples is presented in Table II.

The activity concentrations in the fresh fruits can be obtained by considering the activity of 1 kg of the dried fruit sample as the activity of the total weight of the original quantity before drying.

Table II shows that the dried apricot has a maximum activity concentration of ^{40}K, 750.61 ± 11.88 Bq/kg and the minimum is 15.64 ± 1.57 Bq/kg in dried olive. In the fresh fruits, the maximum activity concentration of ^{40}K is 152.75 ± 3.58 Bq/kg in fresh date and the minimum is 1.99 ± 0.20 Bq/kg in fresh olive. It is worth mentioning that the ^{40}K activity concentration in fresh apricot is 77.54 ± 0.47 Bq/kg, which is almost about 10th of its activity as a dry sample. This ratio is related directly to the dry to fresh weight ratio of the apricot sample, and this concept is valid for all samples.

Fig. 6 presents the activity concentration of ^{40}K for dried fruit samples by both detectors. Fig. 7 shows the activity concentration of ^{40}K in fresh fruit for these samples. Both figures emphasize that the two systems give very close results or equal.

TABLE III
TOTAL POTASSIUM CONTENT (K) IN (MG/100G) IN FRESH FRUITS

Sample	Fresh fruits	
	HPGe	NaI (Tl)
Date	489.81	513.41
Avocado	287.81	289.23
Banana	283.14	275.88
Apricot	249.42	249.09
Pomegranate	198.40	214.76
Mulberry	193.49	201.85
Fig	173.65	181.48
Peach	171.75	168.55
Melon	167.58	167.78
Kiwifruit	129.51	140.45
Strawberry	127.23	125.81
Grape	125.76	127.56
Mango	119.19	125.11
Orange	118.61	125.24
Plum	118.57	119.34
Quince	114.51	117.91
Lemon	111.66	118.69
grapefruit	111.59	107.85
pineapple	108.19	107.68
Pear	107.68	104.62
Green apple	94.41	103.53
Apple	90.97	102.44
Watermelon	59.24	59.87
Olive	6.42	5.42

The measure of the total potassium content (K) in these fruits is a key issue in this study since that concentration has a role in the medical concepts of using or encouraging people to consume food with high potassium. People, in general, consume fresh fruits. However, some of these fruits may present in a dried form in markets or as homemade food. Therefore, we are discussing the potassium concentration in fresh fruits.

Table III illustrates the total potassium content of fresh fruits. The maximum total potassium concentration values were obtained in the date fruit, which were 489.81 mg/100 g and 513.41 mg/100 g from HPGe and NaI spectrometers, respectively, while the minimum values were in olive (6.42 mg/100 g and 5.42 mg/100 g from HPGe and NaI spectrometers, respectively).

The data in Table III are presented in Fig. 8 for fresh fruits. Again, the total potassium concentration in fresh fruits reflects the water content in these types of fruits. From Fig. 8, it is clear that date fruit has the richest potassium content of any of the 24 studied samples of fruits with a clear margin of difference.

There was some difficulty in estimating the intake of total potassium and potassium-40. Since we are not testing an organized or designed food program, it is hard to calculate regular potassium intakes by population. Therefore, we assumed that the better solution to this problem is by calculating these intakes in a unit of consumption. The fruits can be counted and consumed in one meal, such as date, apple, and banana, we calculate the intake for the edible part of a single fruit. For those which cannot be counted or consumed by one person in one meal, such as watermelon and pineapple, we chose a proper served weight in one meal.

Table IV shows the intake of potassium ^{40}K (Bq) and the total potassium (K) (mg) in fruits by defining the mass and the frequency of their daily consumption based on the HPGe measurements. Accordingly, the consumed quantity of the preferable food should agree with established adequate intakes (AIs) by the National Academies of Sciences, Engineering, and Medicine (NASEM) for all ages based on the highest median potassium intakes in healthy children and adults (National Academies of Sciences, 2019).
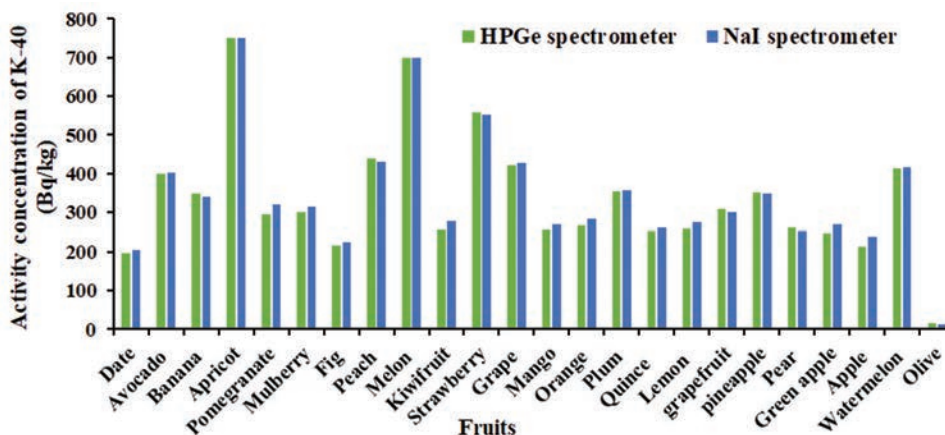

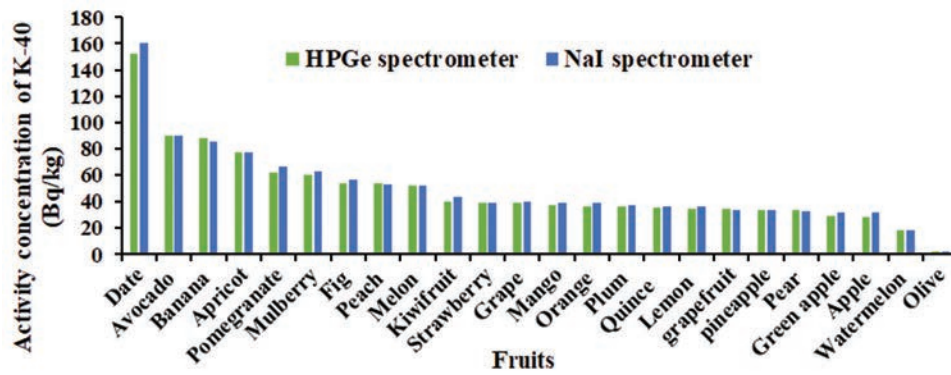Fig. 6. Activity concentration of ^{40}K for dried fruits.

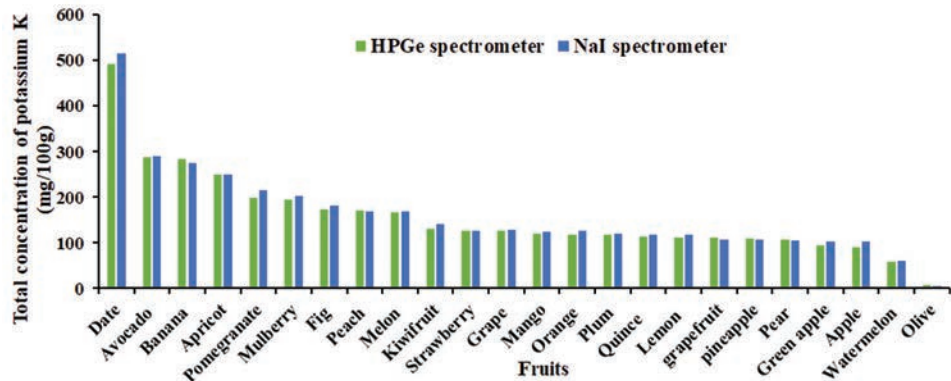Fig. 7. Activity concentration of ^{40}K for fresh fruits.

Fig. 8. Total potassium concentration in fresh fruits.

TABLE IV
INTAKES OF ^{40}K AND TOTAL (K) FROM EATING SPECIFIC QUANTITY
WHICH CAN BE CONSUMED AS A SINGLE UNIT OR PORTION OF
THE FRUITS

Samples	Consumption unit	Mass (kg)	Intake of ^{40}K (Bq)	Intake of total (K) (mg)
Date	1	0.005	0.76	24.49
Avocado	1	0.12	10.74	345.38
Banana	1	0.11	9.68	311.45
Apricot	1	0.025	1.939	62.36
Pomegranate	1	0.2	12.34	396.80
Mulberry	1	0.013	0.78	25.15
Fig	1	0.03	1.62	52.10
Peach	1	0.1	5.34	171.76
Melon	Slice	0.3	15.63	502.74
Kiwifruit	1	0.06	2.42	77.71
Strawberry	1	0.015	0.59	19.09
Grape	1	0.004	0.16	5.03
Mango	1	0.3	11.12	357.57
Orange	1	0.18	6.64	213.50
Plum	1	0.02	0.74	23.71
Quince	1	0.21	7.48	240.48
Lemon	1	0.073	2.53	81.51
Grapefruit	1	0.215	7.46	239.93
Pineapple	Slice	0.3	10.09	324.59
Pear	1	0.14	4.69	150.76
Green apple	1	0.17	4.99	160.51
Apple	1	0.17	4.81	154.66
Watermelon	Slice	0.3	5.53	177.73
Olive	1	0.004	0.01	0.26

B. Comparison with other studies

In Table V, we compare the results of the ^{40}K activity concentration of the present work with some local, regional and international previous measurements and studies, each of these studies has mentioned a few dried or fresh types of the listed samples in the present study. In general, our results agree with some of the referenced data and are higher or lower in some of the other referenced data. Verities and fluctuations in the level of ^{40}K in fruits may neutrally occur due to their production environments regarding the place of growth and the use of fertilizers. Table VI shows the total potassium concentration of the present results compared to published data. Some of the referenced studies here used analytical methods instead of the radioactivity of ^{40}K to estimate the potassium contents.

C. The comparison of the activity levels of the ^{40}K and the potassium concentrations in subgroups of fruits

The activity levels of the ^{40}K and the potassium concentrations in six subgroups of fresh fruits (Fig. 9) were compared using the average values from groups of samples. These classes are classified according to the classification adopted in healthy nutrition systems.

The six classes of fruits were Citrus fruits, berries, drupe, pome, tropical, and melon fruits. Drupe and Tropical fruits generally had the highest level of activity for the ^{40}K and concentration of potassium, while pome fruits showed the

TABLE V
COMPARISON OF POTASSIUM ACTIVITY CONCENTRATION (^{40}K) (BQ/KG) IN
FRUIT SAMPLES FOR THE PRESENT STUDY WITH THE PREVIOUSLY PUBLISHED
LITERATURE

Types of samples		Activity concentration of ^{40}K		
		Present study	Other study	References
Dates	Fresh	152.27±2.12	50.29±2.02	Harb, 2015
Melon	Dry	698.1±12.59	880.28±71.67	Azeez, Mansour,
	Fresh	52.16±0.39	35.49±0.99	and Ahmad, 2019
				Canbazoğlu and
				Mahmut, 2013
Watermelon	Dry	413.3±8.13	300.27±20.07	Azeez, Mansour,
	Fresh	18.61±0.15	34.44±0.88	and Ahmad, 2019
				Canbazoğlu and
				Mahmut, 2013
Orange	Dry	267.3±6.21	279.74±7.70	Priharti and Samat, 2016
Apple	Dry	212.1±5.34	151.49±5.69	Priharti and Samat, 2016
	Fresh	28.28±0.41	26.25±1.24	Harb, 2015
Banana	Dry	349.4±6.76	401.59±5.85	Priharti and Samat, 2016
Mango	Dry	255.7±6.26	171.95±6.61	
Pear	Fresh	33.47±0.47	13.62±1.60	Priharti and Samat, 2016
Quince	Fresh	35.60±0.42	23.01±1.40	Canbazoğlu and
Apricot	Fresh	77.54±0.47	114.90±4.52	Mahmut, 2013
Strawberry	Fresh	39.55±0.38	52.59±2.10	Canbazoğlu and
Fig	Fresh	54.98±0.66	37.25±1.56	Mahmut, 2013
				Harb, 2015
				Harb, 2015
				Harb, 2015

TABLE VI
COMPARISON OF POTASSIUM CONCENTRATION (MG/100 G) IN FRUITS SAMPLES
FOR THE PRESENT STUDY WITH THE PREVIOUSLY PUBLISHED LITERATURE

Fresh fruit samples	Potassium concentration (mg/100 g)		
	Present work	Other study	Reference
Date	489.81	486	Al-Farsi and Lee, 2008
Avocado	287.81	351	Vicente, et al., 2022
Banana	283.14	350	Vicente, et al., 2022
		357	Abt, et al., 2016
Apricot	249.42	259	Vicente, et al., 2022
Pomegranate	198.40	259	Vicente, et al., 2022
Mulberry	193.49	190.96	Liang, et al., 2012
Fig	173.65	232	Vicente, et al., 2022
Peach	171.75	190	Anavi, 2013
Kiwifruit	129.51	280	Cunningham, Milligan and Trevisan, 2001
Strawberry	127.23	170	Vicente, et al., 2022
Grape	125.76	170	Vicente, et al., 2022
Mango	119.19	160	Vicente, et al., 2022
Orange	118.61	150	Vicente, et al., 2022
Plum	118.57	157	Anavi, 2013
Quince	114.51	189	Hegedűs, Papp and Stefanovits-Bányai, 2013
Lemon	111.66	138	Vicente, et al., 2022
Grapefruit	111.59	127	Vicente, et al., 2022
Pineapple	108.19	130	Vicente, et al., 2022
Pear	107.68	119	Anavi, 2013
Apple	90.97	107	Anavi, 2013
Watermelon	59.24	87	Vicente, et al., 2022

lowest levels for ^{40}K and potassium. The major contributor to the potassium level in the drupe fruit is the Date fruit which has the highest level of potassium in fresh fruits in general. Avocados and bananas are the main contributors to potassium levels in fresh tropical fruits.

D. Statistical analysis

Histograms can show the symmetry of data distribution in statistical analysis. If the histogram shapes to the left and right of the vertical line are mirror images of each other, the distribution is symmetrical. The mean and median of a perfectly symmetrical distribution are the same for any data set. However, because each type of fruit is unique in its nature and composition, our data cannot follow such a distribution.

The histograms of the dry fruits data in both detection systems are not symmetrical, as shown in Fig. 10a and b. The distribution is referred to as skewed to the right because it is shifted to the right. In the HPGe detection system, the mean is 336.6 and the median is 298.4. The mean in the NaI(Tl) detection system is 342.8, and the median is 306.9. The Coefficient of Skewness in HPGe is 1.0396 and in NaI(Tl) is 1.006, with rejected normality in both cases.

E. Potassium's health impacts

Potassium is a mineral present in a variety of foods. It is an electrolyte that helps electrical impulses travel throughout the body and aids in several critical biological activities, including blood pressure, proper water balance, muscular
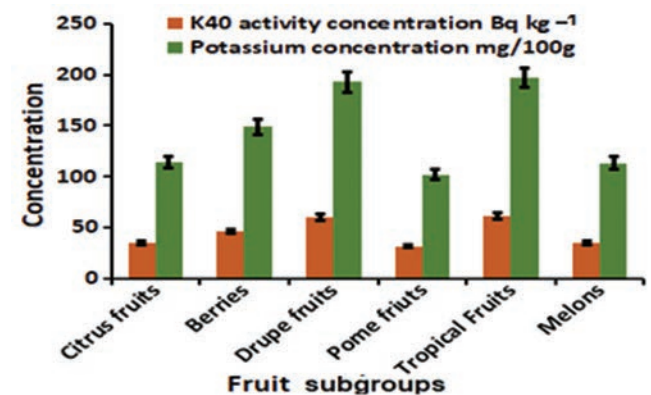

Fig. 9. The comparison of the average activities of ^{40}K and the average concentrations of potassium in six fruit subgroups, with the weighted average of errors of each region (Fresh samples).

contractions, nerve impulses, digestion, heart rhythm, and pH balance (acidity and alkalinity) (Shrimanker and Bhattarai, 2019). Hypokalemia and hyperkalemia are frequent electrolyte diseases induced by potassium intake variations, altered excretion, or transcellular shifts (Viera and Wouk, 2015).

Potassium deficiency can lead to major health concerns. A potassium deficit is characterized by excessive weariness, muscular spasms, weakness, or cramping, irregular heartbeat, constipation, nausea, or vomiting. Kidney illness, heavy diuretic usage, excessive sweating, diarrhea, vomiting, magnesium insufficiency, and antibiotic treatment, such as
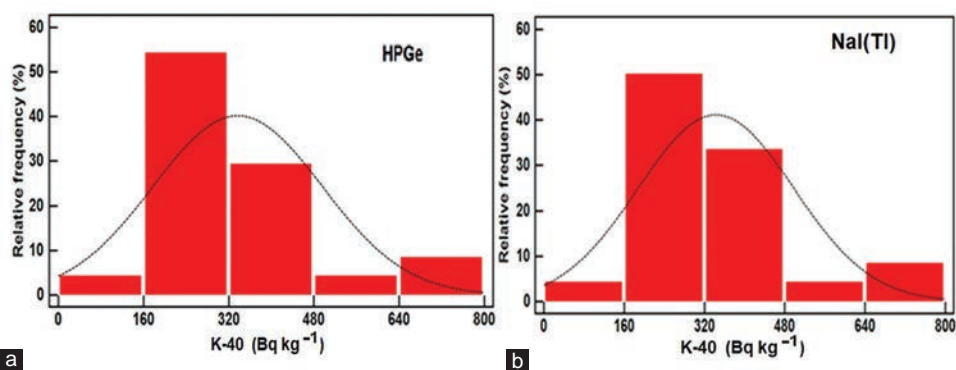
Fig. 10. The Histograms of the data distributions in the two detecting systems. (a) HPGe (b) NaI(Tl) in dry fruits.

carbenicillin and penicillin, can all cause potassium deficit or hypokalemia (low potassium) (Mujais and Katz, 1992; Kardalas, et al., 2018).

Potassium is not produced by the human body on its own. As a result, consuming a balanced diet of potassium-rich foods and drinks is crucial. The use of foods rather than supplements is advised by current efforts to increase potassium consumption. Potassium supplements are not often advised for patients, partly due to concerns about hyperkalemia (high potassium) (Weaver, 2013). An overabundance of potassium is unusual in those who follow a well-balanced diet. If you take too many potassium supplements, you may have overdoses include: renal illness, overtraining, cocaine use, potassium-saving diuretics, chemotherapy, and diabetes. The majority of people get enough potassium through a well-balanced diet. Potassium supplements may be prescribed by a doctor for those who have low potassium levels (Saxena, 1989: Maxwell, et al., 2013).

V. Conclusions

In this study, the activity concentrations of ^{40}K in samples for both detectors are linearly proportional and show the validity of both systems. The results reveal that most fruits contain a considerable amount of water, which makes them very beneficial as a part of the healthy food regime to indirectly preserve the hydration status of those who do not regularly drink enough fresh water and replace it with manufactured drinks and juice, such as children. Fruits contain varying amounts of potassium. Such variation can be used in medical prescriptions or advice for the partial replacement of sodium intake for establishing healthy nutrition or for those suffering from severe hypertension. Drupe and tropical fruits have the highest levels of ^{40}K and potassium in the fruits group, while Pome fruits have the lowest. The date fruit, a member of the drupe fruit family, has the highest level of potassium among fresh fruits in general. Avocados and bananas are the main contributors to potassium levels in fresh Tropical fruits.

Conflict of Interest

The authors declare that they have no conflict of interest.

Data Availability

The datasets created and/or analyzed during the current work can be obtained from the corresponding author upon reasonable requests.

References

Abt, I., Garbini, L. and Schulz, O., 2016. Measurement of the potassium content of different food samples with high purity germanium detectors. *Detection*, 4(3), pp.73-85.

Acquah, G.F. and Pooko-Aikins, M., 2013. Foodstuffs and cancer: Analysis of radionuclides and its radiation levels in common Ghanaian maize. *International Journal of Sciences Basic and Applied Research (IJSBAR)*, 12(1), pp.1-7.

Adedokun, M.B., Aweda, M.A., Maleka, P.P., Obed, R.I., Ogungbemi, K.I. and Ibitoye, Z.A., 2019. Natural radioactivity contents in commonly consumed leafy vegetables cultivated through surface water irrigation in Lagos state, Nigeria. *Journal of Radiation Research and Applied Science*, 12(1), pp.147-156.

Ahmad, S.T., Almuhsin, I.A. and Hamad, W.M., 2021. Radon activity concentrations in Jale and Mersaid warm water springs in Koya District, Kurdistan Region-Iraq. J*ournal of Radioanalytical and Nuclear Chemistry*, 328(3), pp.753-768.

Al-Absi, E., Al-Abdullah, T., Shehadeh, H. and Jamal, A.J., 2015. 226Ra, 228Ra, and 40K activity concentration in some vegetables consumed in Jordan, and resultant annual ingestion effective dose. *Radiation Protection and Environment*, 38(1), pp. 29-34.

Al-Farsi, M.A. and Lee, C.Y., 2008. Nutritional and functional properties of dates: A review. *Critical Reviews in Food Science and Nutrition*, 48(10), pp.877-887.

Alharbi, A. and El-Taher, A., 2013. A study on transfer factors of radionuclides from soil to plant. *Life Science Journal*, 10(2), pp.532-539.

Al-Masri, M., Mukallati, H., Al-Hamwi, A., Khalili, H., Hassan, M., Assaf, H., Amin, Y. and Nashawati, A., 2004. Natural radionuclides in Syrian diet and their daily intake. *Journal of Radioanalytical and Nuclear Chemistry*, 260(2), pp.405-412.

Anavi, S., 2013. *Nutrition and Health-the Importance of Potassium*. International

Potash Institute, Switzerland, pp.3.

Arena, V., 1969. A problem on potassium-40 for radiation biology courses. *The American Biology Teacher*, 31(6), pp.379-382.

Aswood, M.S., Jaafar, M.S. and Bauk, S., 2013. Assessment of radionuclide transfer from soil to vegetables in farms from Cameron Highlands and Penang,(Malaysia) using neutron activation analysis. *Applied Physics Research*, 5(5), pp.85.

Azeez, H.H., Ahmad, S.T. and Mansour, H.H., 2018. Assessment of radioactivity levels and radiological-hazard indices in plant fertilizers used in Iraqi Kurdistan Region. *Journal of Radioanalytical and Nuclear Chemistry*, 317(3), pp.1273-1283.

Azeez, H.H., Mansour, H.H. and Ahmad, S.T., 2019. Transfer of natural radioactive nuclides from soil to plant crops. *Applied Radiation and Isotopes*, 147, pp.152-158.

Burnier, M., 2019. Should we eat more potassium to better control blood pressure in hypertension? *Nephrology Dialysis Transplantation*, 34(2), pp.184-193.

Canbazoğlu, C., and Mahmut, D., 2013. Canbazoğlu, Cumhur, and Mahmut Doğru. A preliminary study on 226Ra, 232Th, 40K and 137Cs activity concentrations in vegetables and fruits frequently consumed by inhabitants of Elazığ Region, Turkey. *Journal of Radioanalytical and Nuclear Chemistry*, 295(2), pp.1245-1249.

Cunningham, J., Milligan, G. and Trevisan, L., 2001. *Minerals in Australian Fruits and Vegetables*. Food Standards Australia, New Zealand.

Currie, L.A., 1968. Limits for qualitative detection and quantitative determination. Application to radiochemistry. *Analytical Chemistry*, 40(3), pp.586-593.

Emumejaye, K., 2012. Determination of potassium-40 concentration in some powdered milk samples consumed in Delta State, Nigeria. *Journal of Applied Physics*, 2, pp.8-12.

Escareño-Juárez, E. and Vega-Carrillo, H.R., 2012. 40K measurement to determine the total potassium content. *Revista Mexicana de Física*, 58(3), pp.211-214.

Harb, S., 2015. Natural radioactivity concentration and annual effective dose in selected vegetables and fruits. *Journal of Nuclear and Particle Physics*, 5(3), pp.70-73.

Hegedűs, A., Papp, N. and Stefanovits-Bányai, É., 2013. Review of nutritional value and putative health-effects of quince (*Cydonia oblonga* Mill.) fruit. *International Journal of Horticultural Science*, 19(3-4), pp.29-32.

Hoeling, B., Reed, D. and Siegel, P.B., 1999. Going bananas in the radiation laboratory. *American Journal of Physics*, 67(5), pp.440-442.

Houston, M. C., 2011. The importance of potassium in managing hypertension. *Current Hypertension Reports*, 13(4), pp.309-317.

Htwe, K.M., and Lwin, T.Z., 2010. Determination of activite of 40K in chemical sample. *Journal Kyaukse Arts and Science*, 2(1), pp.1-12. Available from: https://www.file:///C:/Users/HFC/Downloads/DETERMINATION%20OF%20 ACTIVITY%20(1).pdf [Last accessed on 2022 Aug 01].

Hussain, R.O. and Hussain, H.H., 2011. Investigation the natural radioactivity in local and imported chemical fertilizers. *Brazilian Archives of Biology and Technology*, 54(4), pp.777-782.

Kardalas, E., Paschou, S.A., Anagnostis, P., Muscogiuri, G., Siasos, G. and Vryonidou, A., 2018. Hypokalemia: A clinical update. *Endocrine Connections*, 7(4), pp.R135-R146.

L'Annunziata, M.F. 2012. *Handbook of Radioactivity Analysis*. 3rd ed. Academic Press, Elsevier Inc., USA, Cambridge.

Leutz, H., Schulz, G. and Wenninger, H., 1965. The decay of potassium-40. *Zeitschrift Für Physik*, 187(2), pp.151-164.

Liang, L., Wu, X., Zhu, M., Zhao, W., Li, F., Zou, Y. and Yang, L., 2012. Chemical composition, nutritional value, and antioxidant activities of eight mulberry cultivars from China. *Pharmacognosy Magazine*, 8(31), pp.215-224.

Maxwell, A.P., Linden, K., O'Donnell, S., Hamilton, P.K. and McVeigh, G.E., 2013. Management of hyperkalaemia. *The journal of the Royal College of Physicians of Edinburgh*, 43(3), pp.246-251.

Montenegro-Bethancourt, G., Johner, S.A., and Remer, T., 2013. Contribution of fruit and vegetable intake to hydration status in schoolchildren. *The American Journal of Clinical Nutrition*, 98(4), pp.1103-1112.

Mujais, S.K. and Katz, A.I., 1992. Potassium deficiency. In: *The Kidney: Physiology and Pathophysiology*. Vol. 2. Raven Press, New York, pp.2249-2278.

Najam, L.A., Tawfiq, N.F. and Kitha, F.H., 2015. Measuring radioactivity level in various types of rice using NaI (Tl) detector. *American Journal of Engineering Research*, 4(3), pp.126-132.

National Academies of Sciences., 2019. *Dietary Reference Intakes for Sodium and Potassium*. The National Academics Press, Washington, DC.

Pawloski, J., 2015. *Hypotensive Effects of Potassium and Magnesium*. Arizona State University, Phoenix.

Perez, V. and Chang, E.T., 2014. Sodium-to-potassium ratio and blood pressure, hypertension, and related factors. *Advances in Nutrition*, 5(6), pp.712-741.

Peterson, R., 1996. *Experimental γ Ray Spectroscopy and Investigations of Environmental Radioactivity*. Spectrum Techniques, Tennessee.

Pourimani, R., Noori, M. and Madadi, M., 2015. Radioactivity concentrations in eight medicinal and edible plant species from Shazand, Iran. *International Journal of Ecosystem*, 5(1), pp.22-29.

Priharti, W. and Samat, S.B., 2016. Radiological risk assessment from the intake of vegetables and Fruits in Malaysia. *Malaysian Journal of Analytical Sciences*, 20(6), pp.1247-1253.

Salih, N.F., Hussein, Z.A. and Sedeeq, S.Z., 2019. Environmental radioactivity levels in agricultural soil and wheat grains collected from wheat-farming lands of Koya district, Kurdistan region-Iraq. *Radiation Protection and Environment*, 42(4), pp.128-137.

Santos Jr. J.A., Cardoso J.J., Silva C.M., Silveira S.V. and Amaral R.D., 2005. Analysis of the 40K levels in soil using gamma spectrometry. *Brazilian Archives of Biology and Technology*, 48, pp.221-228.

Sarayegord, A.N., Abbasisiar, F., Abd, A.P. and Ghaisinezhad, M., 2009. Determination of 40K concentration in milk samples consumed in Tehran-Iran and estimation of its annual effective dose. *Iranian Journal of Radiation Research*, 7(3), pp.159-164.

Saxena, K., 1989. Clinical features and management of poisoning due to potassium chloride. *Medical Toxicology and Adverse Drug Experience*, 4(6), pp.429-443.

Shrimanker I., Bhattarai S., 2019. *Electrolytes*. StatPearls Publishing, Treasure Island, FL.

Smail, J.M., Ahmad, S.T. and Mansour, H.H., 2022. Estimation of the natural radioactivity levels in the soil along the Little Zab River, Kurdistan Region in Iraq. *Journal of Radioanalytical and Nuclear Chemistry*, 331(1), pp.119-128.

Steddon, S., Chesser, A., Ashman, N. and Cunningham, J., 2014. *Oxford Handbook of Nephrology and Hypertension*. Oxford University Press, Oxford.

Tin, M.K. and Wai, Z.O., 1993. *Determination of Potassium Content in Myanmar's Clays by High Resolution Gamma Spectrometry and EDXRF Spectrometry*. International Atomic Energy Agency, Vienna.

Tolstykh, E.I., Degteva, M.O., Bougrov, N.G. and Napier, B.A, 2016. Body potassium content and radiation dose from 40K for the urals population (Russia). *PloS One*, 11(4), pp.0154266.

Tzortzis, M. and Tsertos, H., 2004. Determination of thorium, uranium and potassium elemental concentrations in surface soils in Cyprus. *Journal of Environmental Radioactivity*, 77(3), pp.325-338.

Vaessen, H.A.M. and van de Kamp, C.G., 1989. Sodium and potassium assay of foods and biological substrates by atomic absorption spectroscopy (AAS). *Pure and Applied Chemistry*, 61(1), pp.113-120.

Vicente, A.R., Manganaris, G.A., Darre, M., Ortiz, C.M., Sozzi, G.O., and Crisosto, C.H., 2022. Compositional determinants of fruit and vegetable quality and nutritional value. In *Postharvest Handling*: A Systems Approach. Academic

Press, United States, pp.565-619.

Viera, A.J. and Wouk, N., 2015. Potassium disorders: Hypokalemia and hyperkalemia. *American Family Physician*, 92(6), pp. 487-495.

Weaver, C.M. 2013. Potassium and Health. *Advances in Nutrition*, 4(3), pp.368S-377S.

Classification of Different Shoulder Girdle Motions for Prosthesis Control Using a Time-Domain Feature Extraction Technique

Huda M. Radha[1], Alia K. Abdul Hassan[2] and Ali H. Al-Timemy[3]

[1]Department of Computer Science, College of Science, University of Baghdad,
Baghdad, Iraq

[2]Department of Computer Science, College of Science, University of Technology,
Baghdad, Iraq

[3]Department of Biomedical Engineering, Al-Khwarizmi College of Engineering, University of Baghdad,
Baghdad, Iraq

Abstract—**The upper limb amputation exerts a significant burden on the amputee, limiting their ability to perform everyday activities, and degrading their quality of life. Amputee patients' quality of life can be improved if they have natural control over their prosthetic hands. Among the biological signals, most commonly used to predict upper limb motor intentions, surface electromyography (sEMG), and axial acceleration sensor signals are essential components of shoulder-level upper limb prosthetic hand control systems. In this work, a pattern recognition system is proposed to create a plan for categorizing high-level upper limb prostheses in seven various types of shoulder girdle motions. Thus, combining seven feature groups, which are root mean square, four-order autoregressive, wavelength, slope sign change, zero crossing (ZC), mean absolute value, and cardinality. In this article, the time-domain features were first extracted from the EMG and acceleration signals. Then, the spectral regression (SR) and principal component analysis dimensionality reduction methods are employed to identify the most salient features, which are then passed to the linear discriminant analysis (LDA) classifier. EMG and axial acceleration signal datasets from six intact-limbed and four amputee participants exhibited an average classification error of 15.68 % based on SR dimensionality reduction using the LDA classifier.**

Index Terms—**Biosignal analysis, Dimensionality reduction, LDA classifier, Time domain**

I. Introduction

The amputation of an upper limb is a common problem that affects people worldwide. The root causes are many and include vascular disorders such as trauma from accidents, diabetes, and regional conflict-associated traumatic amputations (Choo, Kim and Chang, 2022). The amputee's arm is incomplete if the shoulder has been disarticulated or the upper limb has been amputated below the elbow (Jang, et al., 2011). This means that the amputee cannot use the prosthetic arm because of a lack of muscle power. High-level amputations can be replaced with a body-powered prosthetic hand powered by mechanical cables or a very modern (i.e., TMR) surgical procedure requiring advanced intervention (Craelius, 2021).

Only a few studies have determined that a shoulder girdle electric prosthesis is the newest biological alternative to a limb. For this process to be successful, signals must be transmitted based on the absence of remanent and the shape of the scapula. The focus should therefore be on finding a way to solve the motion detection problem so that adequate prostheses function well as part of the control framework, as the task of the electromagnetic prosthesis is to send the necessary operational signals to detect shoulder motions (Sharba, 2020, Nsugbe et al., 2022).

Rivela, et al., 2015, employed a pattern recognition (PR) system to identify eight healthy people's shoulder motions into nine different groups. The data were segmented using a window length of L = 500 ms and an increment of I = 62 ms as a starting point. Their analysis used waveform length (WL) and root mean square (RMS) as features. The linear discriminant analysis (LDA) was used as a classifier in this research, and the percentage of correctly classified patterns was 92.1 %. Jiang, et al., 2020, examined the ability to build EMG signals from 12 muscles for the upper arm motion pattern identification. In this study, different muscles near the shoulder were shown to produce varied grips. The cross-subjects convolution neural network (CNN) model was employed in this research, and the obtained accuracy was 79.64% in recognizing motion patterns.

Analysis of the data collecting sequence used by Sharba, Wali and Timemy, 2020, has been conducted in another study carried out by Sharba, Wali and Timemy, 2020. This utilized

dataset can be employed for various purposes, including enhancing the operation of prosthetic hands. In recent years, there has been much interest in detecting the intention to move the upper limbs. Forearm muscle activity recordings were used in pattern recognition techniques to identify hand and wrist movements. However, it is impossible to infer the coordinated motion of the body from those signals alone. As a result, the actions of the prosthesis may appear unnatural when viewed as part of the total body, and a dynamic connection between the user and the prosthesis is inconceivable.

A higher-level amputee cannot use these systems because they rely only on contractions of the forearm muscle to generate energy (Li, et al., 2021). However, the reason behind using the time domain analysis for feature extraction is that the time series is a collection of data points that typically include a series of measurements taken at distinct time intervals. A time series analysis statistical technique is designed to examine such measures for feature extraction, model identification, parameter estimation, model validation, and forecasting. A time series model or some statistical aspects of time series data can be recovered as the key damage-sensitive features using the effective feature extraction technique known as time series analysis. In time series analysis, choosing the right model class is crucial. The kind and nature of the time series data and the availability or lack thereof of input or excitation data all affect how this process works (Pulliam et al., 2011). Time-invariant linear models are the best options for feature extraction when vibration measurements are linear and stationary. The most important and fundamental problem in time series modeling is choosing an appropriate and precise order that enables the target model to produce uncorrelated residuals (Entezami, 2021).

The main contributions of the present research are summarized below:

1. For each subject, we combined the seven feature parameters, which are the RMS, four-order autoregressive (AR), WL, slope sign change (SSC), zero crossing (ZC), mean absolute value (MAV), and cardinality to extract 60-dimensional feature vectors. This was accomplished by combining the seven feature groups mentioned previously.
2. The effects of SR and PCA dimensionality reduction algorithms in error analysis testing are then investigated.

The remainder of the work is structured as follows. The following section describes the proposed work plan in detail, and the theoretical background of time-domain feature extraction principles and dimensionality reduction is examined. In Section III, the experimental data and analysis are presented. Finally, Section IV concludes the work and gives suggestions for future investigations.

II. The Proposed Work Scheme

The proposed work scheme for classifying different shoulder girdle motions based on time-domain feature extraction contains signals pre-processing (cross-validation

and segmentation) stage, time-domain features extraction and dimensionality reduction, and then classification stages. First, as indicated in Fig. 1, the cross-validation of training and testing data is applied to pre-process the provided signals, followed by segmentation with overlap window size. After that, time-domain features are conducted. Finally, the LDA classifier is employed to detect seven more shoulder girdle motions for prosthesis control after the SR and PCA dimension reduction presentations. The classes of motion are shown in Fig. 2. The steps are described in more detail in the following subsections.

A. Signal pre-processing

This section outlines the fundamental ideas that underlie the specific data collection techniques employed by Sharba, Wali and Timemy, 2020. The information was received from six participants with intact limbs (Fig. 3a) and four participants with amputees (Fig. 3b). Seven movements were selected for the classification of shoulder girdle motions: Elevation, depression, protraction, retraction, upward rotation, downward rotation, and rest. The extracted data were recorded from five EMG channels. A 3-axis accelerometer sensor (ADXL335) was also mounted on top of the shoulder (Fig. 4) to provide three acceleration signals (Fig. 5).

Cross-validation

Cross-validation was utilized to exclude one trial from this study. To train the classifier in each fold throughout the optimization procedure, eight trials were used. The remaining trials were utilized to evaluate the classifier and calculate the classification error rate. This procedure was done 8 times (eight runs) to calculate the average error rates for the eight runs.
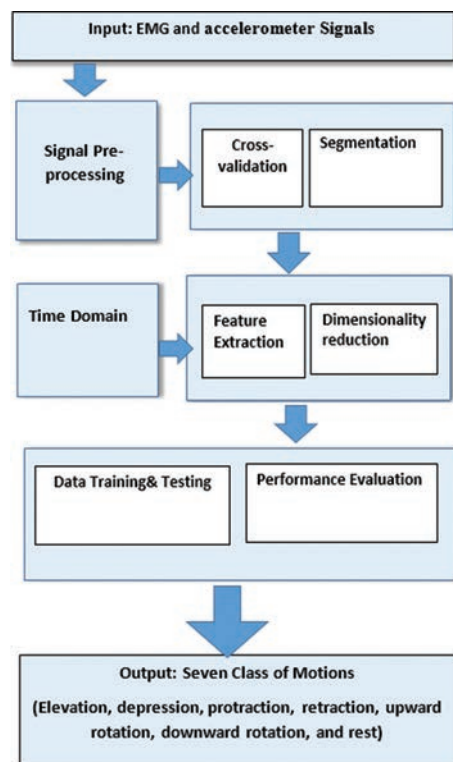

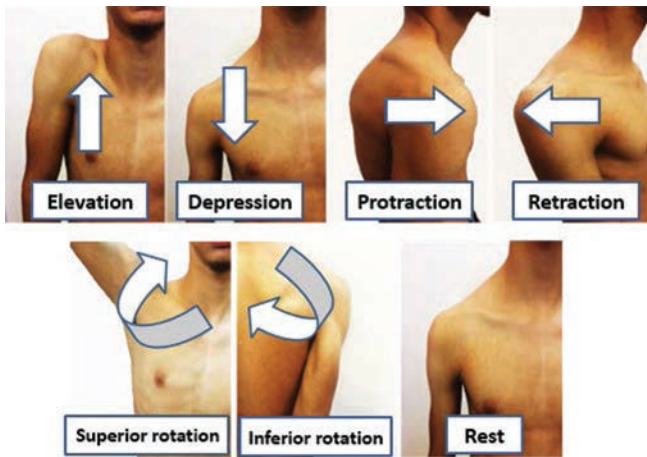Fig. 1. The general design of the proposed work scheme.

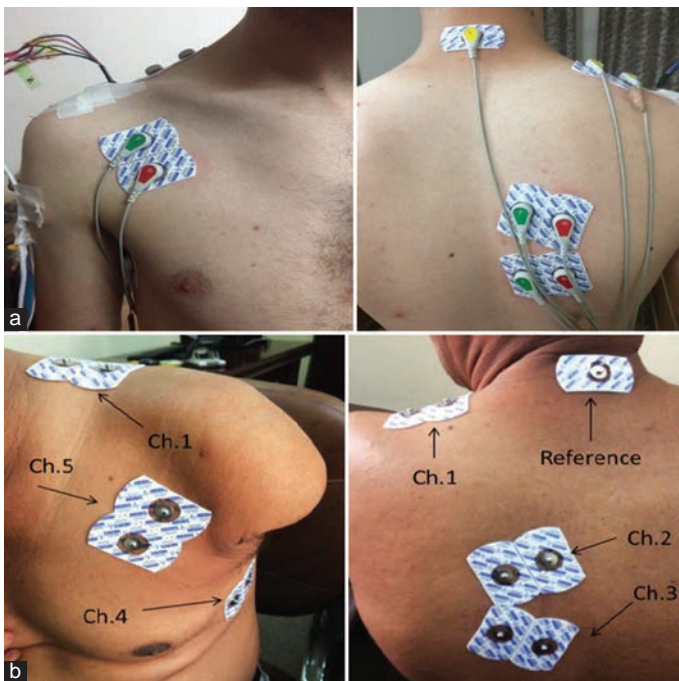Fig. 2. Classes of shoulder girdle motions examples (Sharba, Wali and Timemy, 2020).

Fig. 3. (a) An example of an intact-limbed subject's specific electrode positions. (b) An example of an amputee subject's specific electrode positions.

Fig. 4. Examples of EMG data (5 channels) in the pre-processing condition.

Segmentation

The raw signals are sampled at a rate of 1 kHz. The data were also segmented using an overlapped segmentation approach with a window size of 150 ms and a window increment of 50 ms; Fig. 3 displays the EMG signals from five channels. Fig. 4 depicts the three acceleration signals channels.

B. Time-domain-based feature extraction methods

The time-domain (TD) features are commonly utilized in classification studies and are the most advantageous. The main advantage of TD features extraction is that they are easily extracted and produce excellent results compared to other approaches such as frequency- domain (FD) and time-

FD (TFD) features. The previous research demonstrated the use of TD, particularly in terms of its speed, ease of implementation, and absence of any required transformation (Padfield, 2022). However, the nature of the available signals determines whether or not the time or frequency features should be used. Examining the temporal features of the signal is ineffective in some situations. Thus, researchers must look at it from a new perspective (Boashash, Khan and Ben-Jabeur, 2015). The TDs fundamental problem is that features are formed from the signal's stationary properties. There may be large differences in the components when dealing with
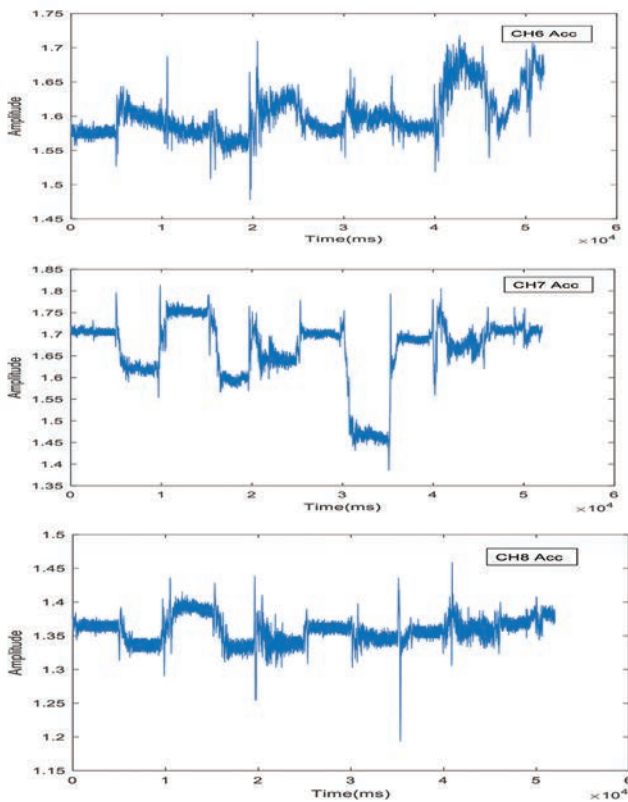
Fig. 5. Examples of 3-axis accelerometer sensor data (three channels) in the pre-processing condition

non-stationary signals like EMG that is typically collected during dynamic movements (Samuel, et al., 2018).

Since the characteristics of the TD have evolved entirely based on the amplitude of the EMG, it is particularly susceptible to noise captured during data collection. Differentiating between stratigraphic movements requires temporal and spectral information. Furthermore, this will be the main criterion for distinguishing between the FD and the TFD in the classification. However, not all of these qualities were used at the same time by a large number of researchers. To achieve efficient classification, important features were selected (Rampichini, et al., 2020). The time-domain features used in this study are as follows:

RMS

The RMS has been described as a Gaussian process analogous to the usual technique of muscular contraction. Furthermore, it mimics the standard deviation SD method (Phinyomark, Khushaba and Scheme, 2018). The RMS mathematical model has been represented as:

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} xi^2} \tag{1}$$

Where, xi represents the EMG signal, whereas N represents the sample number of a signal *and RMS* is a ubiquitous feature that utilizes tested seven TD features with the EMG analysis.

MAV

The MAV feature is ubiquitous and widely exploited by EMG signal researchers. The integrated EMG (IEMG) value

is derived from the corrected EMG moving average. This characteristic has been known by numerous different names, including ARV (i.e., average rectified value) (Phinyomark, Khushaba and Scheme, 2018). The formula for calculating MAV is as follows:

$$MAV = \frac{1}{N}\sum_{i=1}^{N}|xi| \tag{2}$$

Where, xi represents the signal of the EMG, whereas N represents the signal's sample number.

ZC

In temporal domain analysis, the word "ZC" refers to something linked to frequency. It is a measure of the spectral components that occur when the number of EMG magnitudes is more than the level of zero amplitude (Phinyomark, Khushaba and Scheme, 2018). To avoid low-voltage fluctuations or background noise, the threshold condition must be achieved, and the mathematical definition of this condition is as follows:

$$ZC = \sum_{i=1}^{n-1}[sgn(xi \times xi+1)\cap|xi - xi+1| \ge threshold] \tag{3}$$

$$sgn(x) = \begin{cases} 1, & if\ x \ge threshold \\ 0, & otherwise \end{cases}$$

Meanwhile, one characteristic measure the upward ZC split by the peaks (NP) (Al-Timemy, et al., 2015). Only their spectral moments can be used to examine this feature. The following equation can be used to express the corresponding quality as follows:

$$IF = \frac{ZC}{NP} = \frac{SM2}{\sqrt{SM0 \times SM4}} \tag{4}$$

WL

WL can be defined as an EMG complexity measure that describes the total summation of fluctuations throughout every signal segment. This attribute is also referred to as the wavelength, and it is referred to as the total value of absolute derivative signals (WAVE) (Rampichini, et al., 2020). The equation for calculating WL is expressed as:

$$WL = \log\left(\frac{\sum_{i=0}^{n-1}|\Delta x|}{\sum_{i=0}^{n-1}|\Delta 2x|}\right) \tag{5}$$

SSC

The SSC is a ZC trait with a recognizable character, which is how one may characterize it. Calculations are made in which differences in the slope sign are used to indicate information regarding signal frequency. Within their threshold function, positive and negative slope changes have been counted three times in sequence. As a direct consequence of this, background noise in the EMG will not be present (Toledo-Pérez, et al., 2019). The corresponding mathematical expression for this property is as follows:

$$SSC = \sum_{i=2}^{n-1}[f[(x_i - x_{i-1}) \times (x_i - x_{i+1})]] \tag{6}$$

$$f(x) = \begin{cases} 1, & if\ x \ge threshod \\ 0, & otherwise \end{cases}$$

The threshold parameter of this feature should have a value between 50 and 100 mV, as recommended (Tkach, Huang and Kuiken, 2010). Nevertheless, it may be different if the instrument's gain value and the background noise amount are not adjusted to the same level (Badr and Abdul-Hassan, 2021).

Auto-regressive coefficient

When the peak location of the EMG signal is known, the AR feature can be developed on a statistical method using spectral information. It is a prediction model in which the EMG signal is described as a linear combination of the previous samples x_{i-p} and white noise w_i (Phinyomark, Khushaba and Scheme, 2018). Several classification techniques make use of the AR coefficient as a feature vector. The following is the basic AR model:

$$xi = \sum_{p=1}^{p} a_p x_{i-p} + w_i \qquad (7)$$

Where, P represents the AR order at a certain autoregressive coefficient a_p between the fourth-order (*AR4*) and sixth-order (*AR6*) (Chen, Luo and Hua, 2021), research works have been proposed for ideal AR to utilize in the analysis of the EMG.

Cardinality feature

Cardinality is a recently proposed and potentially useful feature. It is denoted by counting the number of components in a group of items while excluding all comparable objects from the elements in that collection (Barton, 2020).

Compared to other commonly used individual features in the literature, cardinality was shown to be one feature that can achieve high levels of accuracy despite the variations in the sampling frequency, window segments, and number and type of movement classes (Barton, 2020). As a result, cardinality has become a fundamental focus of future research. Cardinality was shown to be one feature that can achieve high levels of accuracy despite the variations in the sampling frequency (Praciano, et al., 2021).

As was indicated earlier, researchers have made extensive use of all six of the TD features that were chosen for this study. The previous research, such as that conducted by Samuel, et al., 2018, and Phinyomark, Khushaba and Scheme, 2018, has proposed and used the coupling of TD features with AR. They showed that the feature could be utilized to improve EMG signal classification, which is a significant achievement.

Compared with any FD and TFD hand movement detection system, these features provide extremely high classification accuracy (Karheily, et al., 2022). This was the primary reason behind this research's decision to use the TD features described previously in the EMG data acquired for this investigation.

C. Dimensionality reduction methods

In several numbers of data processing disciplines, such as data mining, machine learning, information retrieval, and pattern recognition, dimensionality reduction has been a significant challenge. The performance of supervised machine learning algorithms degrades when they are presented with multiple features that are not essential for making accurate predictions of the desired output (i.e., prediction accuracy) (Zeng, et al., 2009). One of the most critical aspects of information discovery, machine learning, pattern recognition, and computer vision is isolating a few distinctive and useful characteristics. It is common practice to approach the resolution of this problem by employing methods that involve dimensionality reduction (Sarker, 2021). In this paper, we propose the two most essential methods for dimensionality reduction: Spectral regression (SR) and principal component analysis (PCA). Both of these algorithms are known for their ability to reduce the number of dimensions.

SR

Training sample embedding results are provided by conventional manifold learning techniques such as the locally linear embedding, Laplacian Eigenmap, and isomap. The out-of-sample extension is a challenge; therefore, SR develops a regression model that can avoid the Eigen decomposition of dense matrices, which resolves the challenge of learning and embedding functions (Dong, 2021).

In real-world applications, the resulting data representations are frequently highly dimensional. Practical algorithms generally behave poorly when faced with many unnecessary features. Finding a way to merge them in a lower-dimensional unified space may, therefore, be beneficial for some tasks, such as those involving pattern recognition and regression problems (Jia, et al., 2022). However, DR techniques, including unsupervised, supervised, and semi-supervised methods, were frequently used in many information processing sectors despite the varying assumptions regarding the distribution of the data or the availability of data labeling (Adadi, 2021). Regression- and spectral graph-based SR avoids the Eigen decomposition of dense matrices and operates more effectively at a faster learning rate. The conditions can also be supervised, unsupervised, or semi-supervised (Thudumu, et al., 2020).

Traditional spectral dimensionality reduction strategies need Eigen decomposition of the dense matrices, which has a high computational cost in terms of time and memory, to find an embedding function that minimizes the objective function. The SR algorithm uses the least squares method to determine the ideal projection direction rather than computing the features' density matrix, enabling it to learn substantially more quickly. A G affinity graph with labeled and unlabeled points was developed to examine the complexity of the underlying geometry and learn responses from given data. The embedding function is realized utilizing these responses and standard regression (Nanga, et al., 2021).

PCA

PCA is a mathematical method that can be defined as a technique that simultaneously reduces the dimensionality of data while keeping the majority of its variance 1. It accomplishes this reduction by identifying the major components or directions contributing to maximum data fluctuation. Using a limited number of components, each sample can be characterized by a small number of integers

TABLE I
RESULTS OF TESTING ERROR FOR PRINCIPAL COMPONENT ANALYSIS WITH $No_{Features}$=35 AND LINEAR DISCRIMINANT ANALYSIS CLASSIFIER FOR SIX INTACT LIMBED

Fold No	Intact-limb 1 (%)	Intact-limb 2 (%)	Intact-limb 3 (%)	Intact-limb 4 (%)	Intact-limb 5 (%)	Intact-limb 6 (%)
Fold 1	13.92	21.24	22.34	14.02	17.83	9.51
Fold 2	18.13	13.92	19.73	19.93	14.12	10.52
Fold 3	12.02	11.42	15.73	12.32	28.95	12.72
Fold 4	12.42	16.63	15.83	15.03	15.23	11.02
Fold 5	12.82	11.82	20.24	12.32	18.43	11.22
Fold 6	15.23	7.71	19.93	14.02	22.74	9.81
Fold 7	22.34	10.32	14.32	14.12	22.44	16.83
Fold 8	18.83	15.83	14.72	14.82	27.95	12.92
Mean	15.71	13.61	17.86	14.57	20.96	11.82

TABLE II
RESULTS OF AVERAGE TESTING ERROR FOR SPECTRAL REGRESSION AND LINEAR DISCRIMINANT ANALYSIS CLASSIFIER FOR SIX INTACT LIMBED

Fold No	Intact-limb 1 (%)	Intact-limb 2 (%)	Intact-limb 3 (%)	Intact-limb 4 (%)	Intact-limb 5 (%)	Intact-limb 6 (%)
Fold 1	12.22	16.53	17.83	12.42	14.72	8.31
Fold 2	16.73	11.42	19.23	13.72	12.02	8.61
Fold 3	12.12	8.91	10.82	9.21	24.94	10.42
Fold 4	8.31	15.93	11.32	12.02	13.12	10.02
Fold 5	12.82	9.61	14.02	9.11	16.13	9.31
Fold 6	13.92	7.11	15.23	12.52	19.23	8.11
Fold 7	18.63	8.81	14.02	9.41	17.93	15.60
Fold 8	18.73	13.02	12.82	10.92	21.44	9.91
Mean	14.19	11.42	14.41	11.17	17.44	10.04

instead of the thousands of values associated with the tens of thousands of variables (Jolliffe and Cadima, 2016). On the other hand, PCA provides an orthogonal transformation that converts samples with linearly associated features into data with correlated variables. The main components are new features with fewer or equivalent variables to the starting ones. Since the PCA is an unsupervised approach, the data label information is not included. Ordinarily dispersed data have self-contained principal components (Groth, et al., 2013). The PCA is an easy nonparametric method for extracting the most important information from a collection of redundant or noisy data, and this is why it should be used. The benefits of using them extend far beyond image analysis and data compression to include pattern identification and visualization, as well as the prediction and regression of time series (Nanga, et al., 2021). According to Khalid, Khalil, and Nasreen, 2014, PCA has a few drawbacks, which are as follows:

1. It suggests that the relationships between the variables are linear.
2. It can only be interpreted if all variables are scaled numerically.
3. It lacks a probabilistic model framework that has been regarded as crucial in certain contexts, such as Bayesian decision-making and mixed modeling.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental setup

Data were initially segmented using an overlapping segmentation approach with a 150 ms window and a 50 ms

TABLE III
RESULTS OF TESTING ERROR FOR PRINCIPAL COMPONENT ANALYSIS WITH $No_{Features}$=35 AND LINEAR DISCRIMINANT ANALYSIS CLASSIFIER FOR FOUR AMPUTEES

Fold No	Amputee 1	Amputee 2	Amputee 3	Amputee 4
Fold 1	29.65	22.64	24.45	9.91
Fold 2	20.54	16.23	24.44	15.3
Fold 3	22.24	21.14	26.75	18.23
Fold 4	22.34	17.43	21.04	13.12
Fold 5	18.93	22.34	25.25	18.73
Fold 6	15.83	16.43	28.25	13.72
Fold 7	20.74	20.64	19.13	11.02
Fold 8	28.15	17.13	22.84	12.52
Mean	22.30	19.24	24.01	14.06

TABLE IV
RESULTS OF AVERAGE TESTING ERROR FOR SPECTRAL REGRESSION AND LINEAR DISCRIMINANT ANALYSIS CLASSIFIER FOR FOUR AMPUTEES

Fold No	Amputee 1	Amputee 2	Amputee 3	Amputee 4
Fold 1	27.35	17.43	20.44	9.11
Fold 2	18.33	14.92	25.15	17.7
Fold 3	20.74	19.63	28.45	15.03
Fold 4	20.84	14.62	23.04	11.02
Fold 5	17.63	18.53	23.54	15.83
Fold 6	15.73	14.22	25.55	12.32
Fold 7	18.63	19.93	17.43	10.42
Fold 8	25.65	19.13	21.94	12.52
Mean	20.61	17.30	23.19	12.99

window increment. In each of the folds, classifiers were trained on eight trails. Classes and error rates for classifiers were calculated using individual trails. This method has been repeated eight times. After completing the eight runs, the average error rate for those eight trials was computed.
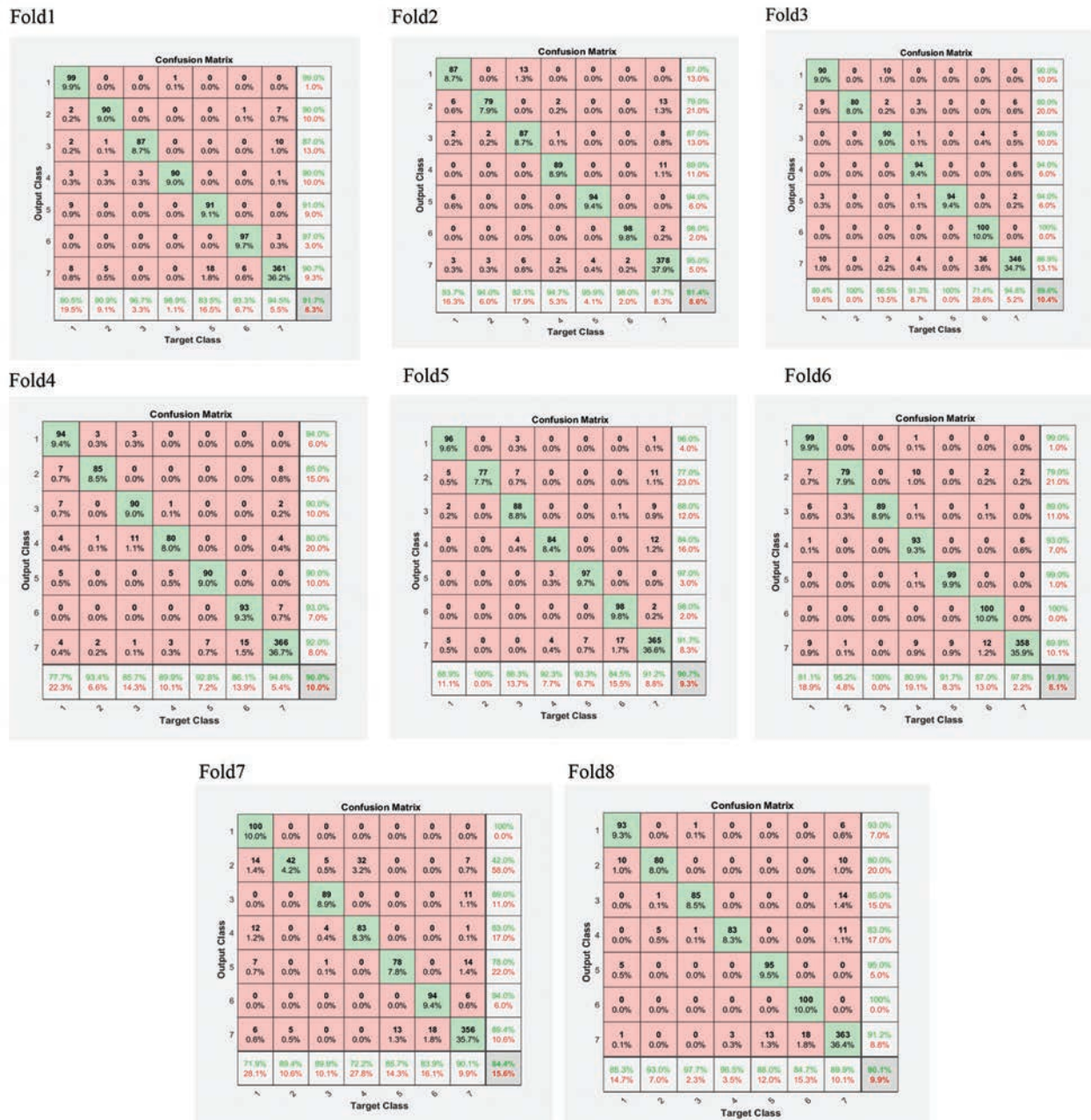
Fig. 6. Confusion matrices of accuracy for each fold of intact-limb 6.

TABLE V
THE OBTAINED AVERAGE TESTING ERROR USING SPECTRAL REGRESSION AND PRINCIPAL COMPONENT ANALYSIS WITH THE LINEAR DISCRIMINANT ANALYSIS CLASSIFIER FOR SIX INTACT LIMBED

Dimensionality reduction	Intact-limb 1 (%)	Intact-limb 2 (%)	Intact-limb 3 (%)	Intact-limb 4 (%)	Intact-limb 5 (%)	Intact-limb 6 (%)	Mean (%)
SR	14.19	11.42	14.41	11.17	17.44	10.04	13.11
PCA	15.71	13.61	17.86	14.57	20.96	11.82	15.75

SR: Spectral regression, PCA: Principal component analysis

TABLE VI
THE OBTAINED AVERAGE TESTING ERROR USING SPECTRAL REGRESSION AND PRINCIPAL COMPONENT ANALYSIS WITH THE LINEAR DISCRIMINANT ANALYSIS CLASSIFIER FOR FOUR AMPUTEES

Dimensionality reduction	Amputee 1	Amputee 2	Amputee 3	Amputee 4	Mean (%)
SR	20.61	17.30	23.19	12.99	18.52
PCA	22.30	19.24	24.01	14.06	19.90

SR: Spectral regression, PCA: Principal component analysis

B. Testing error based on PCA features reduction

The first experiment is conducted using the group of TD features, which are RMS, 4th Autoregressive, WL, SSC, ZC, MAV, and cardinality. In addition, eight different values for each fold (cross-validation) have been tested. The obtained average testing error is calculated for every intact-limb

TABLE VII
COMPARATIVE ANALYSIS OF THE BIOSIGNALS TO CLASSIFY SHOULDER GIRDLE MOTIONS TO OTHER STATE-OF-THE-ART

Authors	Method	Channels	Classifier	Classification accuracy (%)
X. Li et al., Li et al., 2017	Time-domain for feature extraction	10 channels EMG and 10 channels EEG	LDA	84.2
Nsugbe and Al‐Timemy, 2022	Time-frequency domain	5 channels EMG and 3 channels Acc	LDA	70
The proposed scheme	Time-domain for feature extraction	5 channels EMG and 3 channels Acc	LDA	84.9

EMG: Electromyography, LDA: Linear discriminant analysis, EEG: Electroencephalography

subject value-based PCA (dimensionality reduction) with the LDA classifier, as shown in Table I.

Again, the second experiment is conducted using the group of TD features such as RMS, 4th Autoregressive, WL, SSC, ZC, MAV, and cardinality. Every subject value is used to compute the attained average testing error. In addition, the cross-validation method has examined eight distinct values for each fold. As shown in Table II, the LDA classifier and the confusion matrices, as shown in (Fig. 6) for each fold of subject 6, as examples, the obtained average testing error is determined for each subject value based on SR.

The third experiment is done with a set of TD features, such as RMS, 4th Autoregressive, WL, SSC, ZC, MAV, and cardinality. Cross-validation has also been used to test eight different values for each fold. For every amputee's subject value-based PCA (dimensionality reduction) with the LDA classifier, the average testing error is calculated, as shown in Table III.

On the other hand, in the fourth trial, we use a variety of TD features, including RMS, 4th Autoregressive, WL, SSC, ZC, MAV, and cardinality. There have been eight alternative values for each fold tested using cross-validation. Table IV displays the average testing error for each amputee when using SR with the LDA classifier based on the subject values of the amputees.

The final experiment has been carried out to compare the effect of the used dimensionality reduction methods (SR and PCA) on the accomplished testing error for the resulting models with the LDA classifier. Tables V and VI list obtained results for the same values utilized in the previous two experimentations.

As shown in Table V, the dimensionality reduction method (SR or PCA) can affect the accuracy of the classifier (LDA). The value of classification error decreases when using SR with the LDA classifier, which is 13.11% for six intact limbs. The average testing error for four amputees based on the same dimension reduction (SR with the LDA classifier) is 18.52%, as given in Table VI. This exemplifies the nature of SR dimensionality reduction, in which identifying the optimal feature extraction may work effectively with a large feature dimension.

Comparatively, PCA dimensionality reduction performs less well, with an average test error of 15.75% for six intact limbs. However, the average testing error for four amputees with the same dimension increased to 19.90%. As shown in Tables V and VI, the classification error based on SR decreases as long as the size of feature vectors for all subjects. This is the outcome of the dimensionality reduction method, which searches for patterns among gathered features.

The complexity of the work increases with the number of features.

Classification error results showed a slight difference of 4–5% between amputees and intact-limbed participants (Tables V and VI). This could be related to factors such as getting older, how long it has been since an amputation, or even doing shoulder girdle motions. It was also shown that some amputees did not use the shoulder girdle muscles and did not engage in any sort of exercise to train these muscles. However, it is important to note that SR dimensionality reduction outperformed PCA for both limb-intact and limb-amputated participants.

C. Comparative analysis of results

Table VII contains a list of the most common methods for classifying different shoulder girdle motions for controlling prosthetics in a comparative examination of EMG and/or acceleration signals collected from the upper limb. The results of the proposed paper were found to have low test errors than previous works.

IV. CONCLUSION

This study used EMG and accelerometer signals to categorize seven distinct types of shoulder girdle movements performed by high-level upper limb amputees. The results will narrow the focus of dimensionality reduction in the EMG and extract the accelerometer signal feature, which will help determine whether shoulder girdle movements are appropriate as non-invasive and intuitive control signals for upper limb amputees using PR systems. The experimental results showed that SR dimensionality reduction with the LDA classifier was facilitated by extracting regular patterns from biosignals. Experimental results showed that the proposed PR system could identify seven shoulder girdle motions with a classification error of 13.11% for intact-limbed subjects and 18.52% for amputees of SR dimensionality reduction with LDA, and 15.76%–19.90% for PCA dimensionality reduction with the same classifier. Methods from linear algebra that specializes in reducing dimensionality, such as matrix factorization, can be applied in the future studies to decompose a dataset matrix into its component parts. In addition, features can be extracted from biosignals using the TFD to highlight fundamental patterns.

References

Adadi, A., 2021. A survey on data-efficient algorithms in big data era. *Journal of Big Data*, 8, pp.1-54.

Al-Timemy, A.H., Khushaba, R.N., Bugmann, G. and Escudero, J., 2016. Improving the performance against force variation of EMG controlled multifunctional upper-limb prostheses for transradial Amputees. *IEEE Transactions on Neural Systems and Rehabilitation Engineeri*ng, 24(6), pp.650-661.

Barton, N., 2020. Absence perception and the philosophy of zero. *Synthes*e, 197, pp.3823-3850.

Boashash, B., Khan, N.A. and Ben-Jabeur, T., 2015. Time-frequency features for pattern recognition using high-resolution TFDs: A tutorial review. *Digital Signal Processing*, 40(C), pp.1-30.

Praciano, F.D.B., Amora, P.R., Abreu, I.C., Pereira, F.L. and Machado, J.C., 2021. Robust cardinality: A novel approach for cardinality prediction in SQL queries. *Journal of the Brazilian Computer Society*, 27, pp.1-24.

Chen, S., Luo, Z. and Hua, T., 2021. Research on AR-AKF model denoising of the EMG Signal. *Computational and Mathematical Methods in Medicine*, 2021(11), pp.1-10.

Craelius, W., 2021. *Prosthetic Designs for Restoring Human Limb Function*. Springer, Berlin.

Choo, Y.J., Kim, D.H. and Chang, M.C., 2022. Amputation stump management: A narrative review. *World J Clin Cases*, 10(13), pp.3981-3988.

Dong, Y., 2021. A brief review of linear sufficient dimension reduction through optimization. *Journal of Statistical Planning and Inference*, 211, pp.154-161.

Entezami, A., 2021. Feature extraction in time domain for stationary data. In: *Structural Health Monitoring by Time Series Analysis and Statistical Distance Measures*. Springer, Berlin.

Groth, D., Hartmann, S., Klie, S. and Selbig, J., 2013. Principal components analysis. *Computational Toxicology*. Springer, Berlin.

Jang, C.H., Yang, H.S., Yang, H.E., Lee, S.Y., Kwon, J.W., Yun, B.D., Choi, J.Y., Kim, S.N. and Jeong, H.W., 2011. A survey on activities of daily living and occupations of upper extremity amputees. *Annals of Rehabilitation Medicine*, 35(6), pp.907-921.

Jia, W., Sun, M., Lian, J. and Hou, S., 2022. Feature dimensionality reduction: A review. *Complex amd Intelligent Systems*, 8, pp.1-31.

Jiang, Y., Chen, C., Zhang, X., Chen, C., Zhou, Y., Ni, G., Muh, S. and Lemos, S., 2020. Shoulder muscle activation pattern recognition based on sEMG and machine learning algorithms. *Computer Methods and Programs in Biomedicine*, 197, pp.105721.

Jolliffe, I.T. and Cadima, J., 2016. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A Mathematical, Physical and Engineering Sciences*, 374, pp.20150202.

Karheily, S., Moukadem, A., Courbot, J.B. and Abdeslam, D.O., 2022. sEMG time-frequency features for hand movements classification. *Expert Systems with Applications*, 210, pp.118282.

Khalid, S., Khalil, T. and Nasreen, S., 2014. A survey of feature selection and feature extraction techniques in machine learning. In: *2014 Science and Information Conference*. IEEE, Piscataway, pp.372-378.

Li, Q., Liu, Y., Zhu, J., Chen, Z., Liu, L., Yang, S., Zhu, G., Zhu, B., Li, J. and Jin, R., 2021. Upper-limb motion recognition based on hybrid feature selection: Algorithm development and validation. *JMIR mHealth and uHealth*, 9(9), pp.e24402.

Nanga, S., Bawah, A.T., Acquaye, B.A., Billa, M.I., Baeta, F.D., Odai, N.A., Obeng, S.K. and Nsiah, A.D., 2021. Review of dimension reduction methods. *Journal of Data Analysis and Information Processing*, 9(3), pp.189-231.

Nsugbe, E. and Al-Timemy, A.H., 2022. Shoulder girdle recognition using electrophysiological and low frequency anatomical contraction signals for prosthesis control. *CAAI Transactions on Intelligence Technolog*y, 7(1), pp.81-94.

Padfield, N., 2022. Effective EEG Analysis for Advanced AI-driven Motor Imagery BCI Systems.

Phinyomark, A., Khushaba, R.N. and Scheme, E., 2018. Feature extraction and selection for myoelectric control based on wearable EMG sensors. *Sensors*, 18(5), pp.1615.

Pulliam, C.L., Lambrecht, J.M. and Kirsch, R.F. 2011. EMG-based neural network control of transhumeral prostheses. *Journal of rehabilitation research and developmen*t, 48(6), pp.739-754.

Rampichini, S., Vieira, T.M., Castiglioni, P. and Merati, G., 2020. Complexity analysis of surface electromyography for assessing the myoelectric manifestation of muscle fatigue: A review. *Entropy* (*Basel*), 22(5), pp.529.

Rivela, D., Scannella, A., Pavan, E.E., Frigo, C.A., Belluco, P. and Gini, G. 2015. Processing of surface EMG through pattern recognition techniques aimed at classifying shoulder joint movements. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2015, pp.2107-2110.

Samuel, O.W., Zhou, H., Li, X., Wang, H., Zhang, H., Sangaiah, A.K. and Li, G., 2018. Pattern recognition of electromyography signals based on novel time domain features for amputees' limb motion classification. *Computers and Electrical Engineering*, 67, pp.646-655.

Sarker, I.H., 2021. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), pp.160.

Sharba, G.K., Wali, M.K. and Timemy, A.H.A., 2020. Wavelet-based feature extraction technique for classification of different shoulder girdle motions for high-level upper limb amputees. *International Journal of Medical Engineering and Informatics*, 12(6), 609.

Thudumu, S., Branch, P., Jin, J. and Singh, J.J., 2020. A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, 7, pp.1-30.

Tkach, D., Huang, H. and Kuiken, T.A., 2010. Study of stability of time-domain features for electromyographic pattern recognition. *Journal of Neuroengineering and Rehabilitation*, 7, pp.1-13.

Toledo-Pérez, D.C., Rodríguez-Reséndiz, J., Gómez-Loenzo, R.A. and Jauregui-Correa, J., 2019. Support vector machine-based EMG signal classification techniques: A review. *Applied Sciences*, 9(20), 4402.

Zeng, Y., Yang, Z., Cao, W. and Xia, C., 2009. Hand-motion patterns recognition based on mechanomyographic signal analysis. In: *2009 International Conference on Future Biomedical Information Engineering (FBI*E). Institute of Electrical and Electronics Engineers, Piscataway, pp.21-24.

Investigation of Bacterial Persistence and Filaments Formation in Clinical *Klebsiella pneumoniae*: First Report from Iraq

Sarah N. Aziz and Mohammed F. Al Marjani

Department of Biology, College of Science, Mustansiriyah University,
Baghdad, Iraq

Abstract—Bacterial persistence is recognized as a major cause of antibiotic therapy failure, causing biofilms, and chronic intractable infections. The emergence of persisters in *Klebsiella pneumoniae* isolates has become a worldwide public health concern. The goal of the present study is to investigate the formation of persister cells beside filaments in Iraqi *K. pneumoniae* isolates. A total of fifty clinical *K. pneumoniae* isolates were collected from different clinical specimens and identified using the genotypic identification by using specific primer (rpoB gene) from housekeeping genes. Persister cells investigation is performed by exposure of stationary phase *K. pneumoniae* isolates to a high concentration of ciprofloxacin (×10 MIC) and counting the number of viable persister cells by CFU counts. Bacterial filament formation is detected and measured by light microscope scanning electron microscope. The results show the ability of these pathogenic bacteria to form persister cells to survive the bactericidal antibiotics and to cause chronic infection. Furthermore, persistent isolates have the ability to change in shape and size extensively, about 4 times increase in cell length than their normal length. These phenomena are possibly the initial stages of bacterial resistance prevalence.

Index Terms—Filamentation, Klebsiella pneumoniae, Persistence, Scanning electron microscope, Survivor cells.

I. Introduction

Antibiotic therapy failure is known as a global threat in modern medicine that is commonly attributed to resistance. Many studies have uncovered many genetic-molecular mechanisms of bacterial resistance that reduced the effective concentration of antibiotics (Blair, et al., 2015; Khazaal, et al., 2020). However, this is only part of the story. It has long been realized that the formation of persisters in the bacterial population allows a small subpopulation of bacterial cells to survive lethal doses of bactericidal antibiotics. In 1944, medical doctor Bigger noticed the capacity of a small number of bacterial cells in *Staphylococcus aureus* to tolerate and survive a high concentration of antibiotics (Podlesek, et al., 2016). Genetically, persister cells have the same ranges of minimal inhibitory concentration (MIC) as the susceptible cells, but they survive the lethal concentrations of bactericidal antibiotics (Brauner, et al., 2016). Persister cells are dormant, multidrug-tolerant, and non-heritable phenotypic variations of bacteria that differ from resistant populations (Helaine and Kugelberg, 2014). When the antibiotic stress is removed, persisters switch back to a growing state and generate a new population that is genetically identical to the wild-type cell (Aziz, et al., 2021). Persisters may be a middle stage for developing bacterial resistance due to their ability to undergo cellular division during persistence. In addition, persisters can accelerate mutagenesis and horizontal gene transfer due to their stress response mechanisms (Windels, et al., 2019). Therefore, persisters may play a role in the development of multi-drug resistant bacteria (Windels, et al., 2019). *Klebsiella pneumoniae*, belonging to the Enterobacteriaceae family, is an opportunistic bacterium with a capsule that can cause serious hospital-acquired infections and community-acquired infections, such as urinary tract infections, bloodstream infections, pneumonia, pyogenic liver abscesses, and endogenous endophthalmitis (Aziz, et al., 2019). At present, with the recent emergence and dissemination of antibiotic-resistant strains, *K. pneumoniae* have gained notoriety as an infectious agent due to a rise in the number of severe infections and the increasing lack of effective treatments so, it has become a major health concern (Da Silva, et al., 2019). The failure of antibiotics against *K. pneumoniae* has been widely studied, bacterial persistence may play a role in treatment failure (Navon-Venezia, et al., 2017). However, it still few studies about the persistence *K. pneumoniae* (Li, et al., 2018). Furthermore, there are no studies about persistence formation in Iraq. Therefore, to provide more information about *K. pneumoniae* persistence, the goal of the present study was to investigate the formation of persisters cells beside filaments formation in Iraqi clinical *K. pneumoniae* isolates.

II. Materials and Methods

A. Bacterial Isolates, Identification, and Growth Conditions

A number of fifty clinical *K. pneumoniae* isolates were obtained from Teaching Laboratories in Medical city/Baghdad. The identification of the isolates was done using their growth characteristics in culture media and confirmed using and Vitek-2 system (BioMe´rieux, France). Conventional PCR was performed for the genotypic identification of *K. pneumoniae* isolates using specific primers for *rpoB* gene. The sequence of the primers and PCR cycling conditions are listed in Table I. The isolates were routinely grown in Luria Bertani (LB) broth at 37°C. for a 24 h.

B. The MIC Determination

The MIC values of ciprofloxacin antibiotic were investigated using E-test gradient strips (bioMérieux/France) on a Mueller-Hinton agar medium. The MIC was measured and the lowest concentration of ciprofloxacin antibiotic non-permissive for visible growth was reported to be the MIC. *Escherichia coli* ATCC 25922 was used as the negative control strain.

C. Phenotypic Detection of Persister Cells Formation

With few modifications, persister cells formation was carried out according to (Chung, Wi and Ko, 2017) by exposing exponential phase bacteria (optical density OD600 = 0.5) to a high concentration of ciprofloxacin (10 MIC), then incubated at 37°C for 16 h in LB media. The bacterial cells of *K. pneumoniae* were washed with phosphate-buffer saline and then spread on (LB) agar medium for 24 h. After the incubation, persister cells numbers in the agar medium were determined by colony-forming units. Furthermore, few colonies from each bacterial sample were re-inoculated on (LB) broth medium for 24 h and an antibacterial sensitivity test was performed, to confirm their identity as persister cells by non-changing of MIC values.

D. Molecular Detection of hipB and mazE Genes in Persister Cells

PCR reaction mixture for *hipB* and *mazE* genes, which are related to persistence state, was set up for each gene alone in a final volume of 25 μl. It was composed of 2.5 of Green master mix ×2, 5.5 μL of nuclease-free water (Promega, USA), 2 μL of primers (1 μL forward and 1 μL reverse), and 5 μL template DNA in PCR Eppendorf tubes. Negative control was used with all PCR experiments, which consisted of all the materials but without the template DNA.

The mixture was vortexed and briefly centrifuged to move the contents to the bottom of the tubes before being placed in a thermocycler PCR. Table I shows the primer sequence and the size products of the PCR.

E. Detection of Filament Formation

Detection of filament formation was done using microtiter wells plate according to (Buijs, et al., 2008) with few modifications. Briefly, eight sub-MIC concentrations of ciprofloxacin antibiotic were prepared (5, 10, 20, 30, 40, 50, 100, and 120 mg/L). A negative control concentration that did not have any antibiotic also was used as a reference for *K. pneumoniae* isolates length. The ciprofloxacin antibiotic solution was diluted in distilled water then 20 μL of these dilutions were added microtiter wells plate followed by adding 180 μL of *K. pneumoniae* suspension with gentle mixing then incubated for 4 h at 37°C. After that, slide smear and Gram stain were done for each well to examine by light microscope. Following Gram's stain, determine the filament size which is the most characteristic often used to help identify filament formation bacteria, persistence *K. pneumoniae* isolate was detected. For determining the filament sizes, microscopy and scanning electron microscope were used.

F. Statistical Analysis

The data results of this study were analyzed using GraphPad Prism 8 software and Microsoft Excel 2013 for each biological replicate. The level of probability at $P \leq 0.05$, which used to identify a significant difference.

III. Results and Discussion

A. Bacterial Identification and Growth Conditions

The isolates of *K. pneumoniae* were collected from different clinical samples including 20 isolates from urine, 15 isolates from burns swab, 6 isolates from stool, 5 isolates from blood, and 4 isolates from sputum. The identification of *K. pneumoniae* isolate was confirmed by phenotypic identification using the Vitek 2 system (BioMe´rieux, France), then re-confirmed by gene sequence analysis of *rpoB* gene (β-subunit of RNA polymerase). Fig. 1 shows the amplification of *rpoB* gene in the gel electrophoresis for detecting *K. pneumoniae* isolates.

B. Phenotypic Detection of Persister Cells Formation

Out of 50 *K. pneumoniae* isolates, 2/50 isolates were persister cell formation. Ciprofloxacin dose-dependent killing

TABLE I
Primers Used in the Current Study for PCR Amplification

Activity	Name of gene	Primers 5 ---- 3	Size products	References
Housekeeping gene	*rpoB*	F: GGC GAA ATG GCW GAG AAC R: GAG TCT TCG AAG TTG TAA	1056 bp	(Kareem, et al., 2021)
Toxin-antitoxin system genes	*hipB*	F: AGCCCAACGCAATTGGCGAATGCA R: CTGTTCTGTTGATTCTGGCGAGGC	225 bp	(Hemati, et al., 2014)
	mazE	F: ATGATCCACAGTAGCGTAAAGCGT R: TTACCAGACTTCCTTATCTTTCGG	249 bp	(Hemati, et al., 2014)

(MIC = 0.031 µg/mL; sensitive) of *K. pneumoniae* test isolates. With 0.31 µg/mL ciprofloxacin (10-fold MIC), the majority of the population was efficiently killed, followed by a near plateau with minimal decrease in survival up to 40 g/mL ciprofloxacin concentration.

The percentage of survival at 30 µg/mL ciprofloxacin was 1.3% of the initial cell count, and it was about 1.5% at 40 µg/mL ciprofloxacin (1000-fold MIC).

Because they survived the ciprofloxacin treatment, the survivors along the plateau of the antibiotic dose-dependent killing curve could be considered persister cells. To ensure the persistence state, MIC of these survivor cells against ciprofloxacin remained unchanged when compared to the test organism's wild-type cell population. Furthermore, no colonies appeared after an incubation period of 24 h at 37°C on LB agar plates containing 30 µg/mL ciprofloxacin that were pre-inoculated with survival cells when compared to their growth on LB agar plates alone without added ciprofloxacin. Our study is resembling for many recent studies revealed that *K. pneumoniae* are the pioneer intricate bacteria that can persist and resist in the presence of distinct antibiotics, including fluoroquinolones (Lee, et al., 2019; Abokhalil, et al., 2020; Kareem, et al., 2021). However, in Iraq, there is a need to focus on bacterial persistence, especially with opportunistic bacteria like *K. pneumoniae*.

C. Molecular Detection of hipB and mazE Genes in Persister Cells

The results showed that the same two *K. pneumoniae* isolates which were able phenotypically to form persister cells were harboring both *hipB* and *mazE* genes as shown
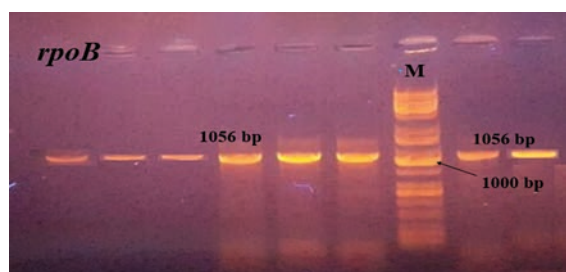
Fig. 1. Gel electrophoresis (1% agarose, 70 volt for 50 min) of *rpoB* gene (1056 bp). Lane M 1000bp DNA Ladder, the other lanes bands are that the other lines are the positive results of *rpoB*.
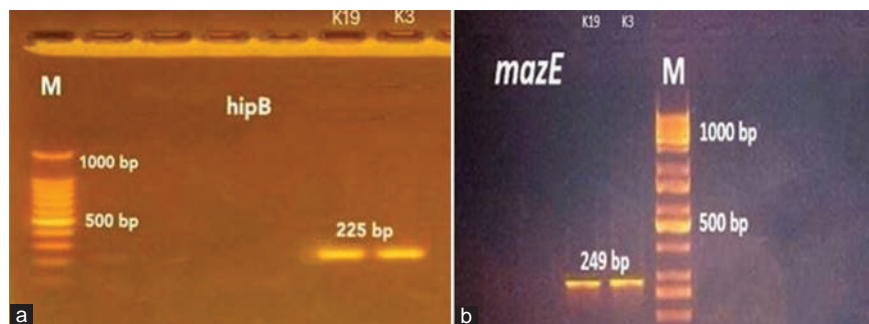
in Fig. 2. These results were supported by many other results such as a recent study by (Fraikin, Goormaghtigh and Van Melderen, 2020). Furthermore, in their report, (Wood and Songs, 2020) suggested that TA system can be a realistic choice for both inductions of persistence and their elimination. The current study supported by many international studies that indicated the wide distribution of TAs type II in *K. pneumoniae*, especially a high level of *hipAB*, *mazEF*, *ccdAB*, and *relEB* genes (Coskun, et al., 2018; Horesh, et al., 2020). When activated, TA stress-responsive elements cause reversible bacterial growth indictment, making them typical candidates for persistence drivers. The role of TA operons in persistence, on the other hand, is a hotly debated topic. When Moyed and Bertrand discovered a high-persistence mutation in HipAB, they established the first connection between persisters and TA modules (Gollan, et al., 2019). HipB encodes an auto repressor of *hipBA* transcription (Black, Irwin and Moyed, 1994; Harms, et al., 2018). HipA overproduction inhibited cell growth by reducing translation, DNA replication, and transcription, as well as significantly increasing tolerance to bactericidal antibiotics (Korch, Henderson and Hill, 2003).

When exposed to bactericidal antibiotics, bacterial persisters form through stress responses that stimulate ppGpp synthesis, leading to MazE antitoxin degradation and MazF toxin activation. MazF inhibits RNA and protein synthesis effectively blocking cell elongation and protecting persister cells from death by lysis. Ciprofloxacin generates DNA breaks which, if not repaired through the SOS response path-way, quickly kill dividing bacteria that are actively synthesizing DNA. By inhibiting DNA synthesis, MazF blocks cell division and protects persisters from ciprofloxacin. Bactericidal antibiotics were reported to stimulate ROS production that could lead to mutant cells, but these conclusions were challenged by subsequent studies (Gollan, et al., 2019).

D. Detection of Filament Formation

Results of exposing persistent *K. pneumoniae* isolates to various concentrations from the ciprofloxacin sub-MIC showed that cell filament formed at 0.25, 0.19, and 0.125 µg/mL concentrations so, the bacterial isolate was shape shifter into a filamentous, Fig. 3. Our results showed that persistent *K. pneumoniae* was changed in
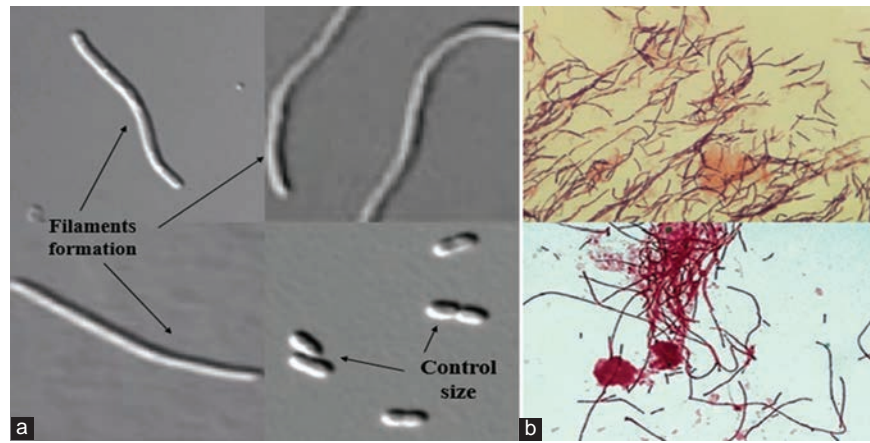

Fig. 2. Gel electrophoresis (1% agarose, 70 volt for 50 min) of (a) *hipB* (225 bp), (b) *mazE* (249 bp) genes.

Fig. 3. (a) Scanning electron microscope, (b) light microscope, ×100 magnification, filament of *Klebsiella pneumoniae* isolate.

shape and size extensively (filamentation) associated with antibiotics therapy and persistence formation. Persistent *K. pneumoniae* isolates exhibited cell elongation, about 4 times increase in cell length when examined under a scanning electron microscope. A recent study suggested that Filaments Formation by persistent bacteria, during antibiotics exposure or under other stress conditions, is possibly the initial stage of emergence of the bacterial resistance (Joseph, et al., 2018).

IV. Conclusion

This study reports the formation of persistence *K. pneumoniae*. The isolation of persister cells can be done with antibiotic treatment (ciprofloxacin). The slow growth rate after resuscitation indicated that metabolisms of persistent bacteria had slowed. This emphasizes the significance of researching the underlying mechanisms contributory to the persistence state, which should be the focus of our future perspectives.

References

Abokhalil, R.N., Elkhatib, W.F., Aboulwafa, M.M. and Hassouna, N.A., 2020. Persisters of *Klebsiella pneumoniae* and *Proteus mirabilis*: A common phenomenon and different behavior profiles. *Current Microbiology*, 77(7), pp.123-1244.

Aziz, S.N., Al Marjani, M.F., Rheima, A.M. and Al Kadmy, I.M., 2021. Antibacterial, antibiofilm, and antipersister cells formation of green synthesis silver nanoparticles and graphene nanosheets against *Klebsiella pneumoniae*. *Reviews in Medical Microbiology*, 33(1), pp.56-63.

Aziz, S.N., Al-Sallami, K.J., Abd, S.Y., Al-Musawi, A.M.A., Mohammed, M.A., Abed, A.M.H. and Mohammed, S.Q., 2019. Improving the antibacterial activity by the combination of zirconium oxide nanoparticles (ZrO₂) and ceftazidime against *Klebsiella pneumoniae*. *Global Journal of Public Health Medicine*, 1(1), pp.16-20.

Black, D.S., Irwin, B. and Moyed, H.S. 1994. Autoregulation of hip, an operon that affects lethality due to inhibition of peptidoglycan or DNA synthesis. *Journal of Bacteriology*, 176(13), pp.4081-4091.

Blair, J.M., Webber, M.A., Baylay, A.J., Ogbolu, D.O. and Piddock, L.J. 2015. Molecular mechanisms of antibiotic resistance. *Nature Reviews Microbiology*, 13(1), pp.42-51.

Brauner, A., Fridman, O., Gefen, O. and Balaban, N.Q. 2016. Distinguishing between resistance, tolerance and persistence to antibiotic treatment. *Nature Reviews Microbiology*, 14(5), pp.320-330.

Buijs, J., Dofferhoff, A.M., Mouton, J.W., Wagenvoort, J.H.T. and Van Der Meer, J.W.M., 2008. Concentration-dependency of β-lactam-induced filament formation in Gram-negative bacteria. *Clinical Microbiology and Infection*, 14(4), pp.344-349.

Chung, E.S., Wi, Y.M. and Ko, K.S., 2017. Variation in formation of persister cells against colistin in *Acinetobacter baumannii* isolates and its relationship with treatment failure. *Journal of Antimicrobial Chemotherapy*, 72(7), pp.2133-2135.

Coskun, U.S.S., Cicek, A.C., Kilinc, C., Guckan, R., Dagcioglu, Y., Demir, O. and Sandalli, C., 2018. Effect of *mazEF*, *higBA* and *relBE* toxin-antitoxin systems on antibiotic resistance in *Pseudomonas aeruginosa* and *Staphylococcus isolates*. *Malawi Medical Journal*, 30(2), pp.67-72.

Da Silva, Y., Ferrari, R., Marin, V.A. and Junior, C.A.C., 2019. A global overview of β-lactam resistance genes in *Klebsiella pneumoniae*. *The Open Infectious Diseases Journal*, 11(1), pp.22-34.

Fraikin, N., Goormaghtigh, F. and Van Melderen, L., 2020. Type II toxin-antitoxin systems: Evolution and revolutions. *Journal of Bacteriology*, 202(7), pp.e00763-19.

Gollan, B., Grabe, G., Michaux, C. and Helaine, S., 2019. Bacterial persisters and infection: Past, present, and progressing. *Annual Review of Microbiology*, 73, pp.359-385.

Harms, A., Brodersen, D.E., Mitarai, N. and Gerdes, K., 2018. Toxins, targets, and triggers: An overview of toxin-antitoxin biology. *Molecular Cell*, 70(5), pp.768-784.

Helaine, S. and Kugelberg, E., 2014. Bacterial persisters: Formation, eradication, and experimental systems. *Trends in Microbiology*, 22(7), pp.417-424.

Hemati, S., Azizi-Jalilian, F., Pakzad, I., Taherikalani, M., Maleki, A., Karimi, S., Monjezei, A., Mahdavi, Z., Fadavi, M.R., Sayehmiri, K. and Sadeghifard, N., 2014. The correlation between the presence of quorum sensing, toxin-antitoxin system genes and MIC values with ability of biofilm formation in clinical isolates of *Pseudomonas aeruginosa*. *Iranian Journal of Microbiology*, 6(3), p.133-139.

Horesh, G., Fino, C., Harms, A., Dorman, M.J., Parts, L., Gerdes, K., Heinz, E. and Thomson, N.R., 2020. Type II and Type IV toxin-antitoxin systems show different evolutionary patterns in the global *Klebsiella pneumoniae* population.

Nucleic Acids Research, 48(8), pp.4357-4370.

Joseph, J., Sharma, S., and Dave, V.P., 2018. Filamentous gram-negative bacteria masquerading as actinomycetes in infectious endophthalmitis: A review of three cases. *Journal of Ophthalmic Inflammation and Infection*, 8(1), pp.1-6.

Kareem, S.M., Al-Kadmy, I.M., Kazaal, S.S., Ali, A.N.M., Aziz, S.N., Makharita, R.R., Algammal, A.M., Al-Rejaie, S., Behl, T., Batiha, G.E.S. and El-Mokhtar, M.A., 2021. Detection of *gyra* and *parc* mutations and prevalence of plasmid-mediated quinolone resistance genes in *Klebsiella pneumoniae*. *Infection and Drug Resistance*, 14, p.555.

Khazaal, S.S., Al-Kadmy, I.M., and Aziz, S.N., 2020. Mechanism of pathogenesis in multidrug resistant *Acinetobacter baumannii* isolated from intensive care unit. *Gene Reports*, 18, p.100557.

Korch, S.B., Henderson, T.A. and Hill, T.M., 2003. Characterization of the hipA7 allele of *Escherichia coli* and evidence that high persistence is governed by (p) ppGpp synthesis. *Molecular Microbiology*, 50(4), pp.1199-1213.

Lee, J.S., Choi, J.Y., Chung, E.S., Peck, K.R., and Ko, K.S., 2019. Variation in the formation of persister cells against meropenem in *Klebsiella* pneumoniae

bacteremia and analysis of its clinical features. *Diagnostic Microbiology and infectious disease*, 95(3), p.114853.

Li, Y., Zhang, L., Zhou, Y., Zhang, Z., and Zhang, X., 2018. Survival of bactericidal antibiotic treatment by tolerant persister cells of Klebsiella pneumoniae. *Journal of Medical Microbiology*, 67(3), 273-281.

Navon-Venezia, S., Kondratyeva, K., and Carattoli, A., 2017. *Klebsiella* pneumoniae: A major worldwide source and shuttle for antibiotic resistance. *FEMS Microbiology Reviews*, 41(3), pp.252-275.

Podlesek, Z., Butala, M., Šakanović, A., and Žgur-Bertok, D., 2016. Antibiotic induced bacterial lysis provides a reservoir of persisters. *Antonie Van Leeuwenhoek*, 109(4), pp.523-528.

Windels, E.M., Michiels, J.E., Fauvart, M., Wenseleers, T., Van den Bergh, B., and Michiels, J., 2019. Bacterial persistence promotes the evolution of antibiotic resistance by increasing survival and mutation rates. *The ISME Journal*, 13(5), pp.1239-1251.

Wood, T.K. and Song, S., 2020. Forming and waking dormant cells: The ppGpp ribosome dimerization persister model. *Biofilm*, 2, p.100018.

Medicinal Plants Traditionally Used in the Management of COVID-19 in Kurdistan Region of Iraq

Mahmoud D. Abdulrahman[1], Fattma Z. Mohammed[2], Saber W. Hamad[1,3], Harmand A. Hama[1], and Abubakar A. Lema[4]

[1]Biology Education Department, Faculty of Education, Tishk International University, Erbil,
Kurdistan Region – F.R. Iraq

[2]Department of Nursing, College of Nursing, Hawler Medical University, Erbil,
Kurdistan Region – F.R. Iraq

[3]Department of Field Crops and Medicinal Plants, College of Agricultural Engineering Sciences, Salahaddin University-Erbil,
Kurdistan Region – F.R. Iraq

[4]Department of Biological Sciences, College of Natural and Applied Sciences, Al-Qalam University Katsina,
Katsina State, Nigeria

Abstract—**Coronaviruses are infectious respiratory tract illnesses, but they can also affect the digestive tract and infect both humans and animals. The new coronavirus results in complicated health problems all over the world. The most urgent concern of all researchers around the world has been the treatment of the virus. The following study aimed to use quantitative ethnobotany to help scientist in addressing the deadly virus. Expert sampling method was adopted with the aid of an in-depth interview guide. Thirty-nine respondents were interviewed. Eighty-one medicinal plant species from 35 families were documented. Males 25 (64.1%) constitute the greater percentage of the total respondents. Majority of the respondents had formal education. Eighty-one medicinal plant species from 35 families were documented. Leaves are the most utilized 25.8 followed by seed 17.7 and fruits 12.1%, respectively. Relative frequency of citation ranged from 0.5 to 0.9, whereas the FL value ranged from 0.4 to 0.85, revealing how effective the documented plant species are in the management of COVID-19 in the region. A greater amount of research into documented medicinal plants is warranted because of the high likelihood that they contain many active ingredients.**

Index Terms—**Coronaviruses, Expert sampling method, Iraqi Kurdistan, Medicinal plant, Relative frequency of citation.**

I. Introduction

Man has long been using plants to treat ailments. Since the dawn of time, man has been enthralled by the knowledge and application of traditional medicinal plants, which has been passed down from generation to generation (Dogara, et al., 2022). Plants have long been known for their therapeutic properties, and people all over the world have traditionally employed them to cure a variety of diseases. Plant-based medicines are now commonly considered as the safest and most effective way to combat infectious diseases (Merouane, et al., 2022). Plant chemical compositions have a wide range of medical applications (Abdulrahman, et al., 2022). The most urgent concern of all researchers around the world has been the treatment of the virus. The dearth of effective vaccines against this devastating viral illness has prompted experts to look for natural remedies that could aid in the fight against the viral pandemic (Lim, et al., 2021). China has been using herbal traditional medicines since the start of the COVID-19 outbreak (Khadka, et al., 2021). Traditional medicines, on the other hand, were shown to help 90% of the 214 patients they were given. Healthy people were protected from SARS-CoV-2 infection, whereas patients with mild or severe symptoms saw improvements in their health after using specific traditional herbal remedies (Benarba and Pandiella, 2020). The pandemic has prompted researchers from a wide range of disciplines to investigate the virus's origin, structure, causes, diagnostic techniques, and therapeutic alternatives. The following study aimed to use quantitative ethnobotany to help scientist unaddressing the deadly virus.

II. Materials and Methods

A. Sampling and Interview sessions

In this study, non-random probability approach and expert sampling methods were used. Traditional medical practitioners and elderly folks with traditional plant knowledge are interviewed. An in-depth questionnaire served as a guide for the interview.

B. Data Collection Procedure

Direct interviews with local people were done in the Kurdistan Region of Iraq from January 2021 to June 2022

to gather data for this study. The responders' verbal consent was sorted. The significance of the study was communicated to them. Each respondent was visited two to three times to ensure that the data were accurate. If there was a discrepancy between the information provided previously and the information received during subsequent trips to a particular plant, it was deemed unreliable and dismissed. Data were gathered using communicable dialects within the area, as per the traditional inquiry approach.

C. Plant Collection and Herbarium Specimen Deposition

Species of plants were collected or purchased from the herbalist during the interview. Plants with different names were avoided. A licensed botanist (Dogara, et al., 2022) from Tishk International University, Faculty of Education, Department of Biology, identified the collected plant specimens. Letters plants species were confirmed in the Salahaddin University herbarium. Plant names were verified according to http://www.worldfloraonline.org/.http://www.worldfloraonline.org/.

D. Data Analysis

Based on the below information, the study used a simple descriptive analysis of the ethnobotanical data to calculate the frequencies and percentages:
1. Demographic information of the participants
2. Documentation of the reported plants, parts of plants used, preparation methods, prescription, administration methods, and toxicity of reported plants
3. Symptoms of COVID-19
4. Quantitative analysis was computed based on the following:
   a) Relative frequency of citation (RFC): = Fc/N, where Fc is the number of people who mentioned a particular plant species and N is the overall number of respondents interviewed (Mahmoud and Abba, 2021).
   b) Fidelity level (FL): = Ns/N × 100. Where Ns = Total number of respondents who indicated they employed a specific plant species to treat a specific condition and N = Total number of informants who mentioned the plant species during the interview (Mahmoud and Abba, 2021).

III. Results and Discussion

A. Demographic Profile of the Participants

The COVID-19 pandemic has engulfed the entire globe; people are dying by hundreds every day without access to effective treatment, and it is impossible to bring this global health crisis to an end without it (Adhikari, et al., 2021). There have been several different experiments, but none of them have given promising results (Khadka, et al., 2021). COVID-19 has been the subject of numerous hoaxes on social media, including the use of medicinal plant products to prevent or treat the disease. To prevent erroneous knowledge from spreading, ethnobotanists should connect with local people and record the therapeutic plants used. Participants are the most important part of any ethnobotanical study. Their age, gender, education level, occupation, religion,

etc., provide insight into the survey and facilitate placing the data provided in its proper social context for analysis and interpretation (Abdulrahman, et al., 2022). Traditional practitioners in the research area were discovered, as they play an important part in the primary healthcare systems of the local people. Table 1 shows the demographic profile of the respondents, males 25 (64.1%) constitute the greater percentage of the total respondents, whereas females were represented by 14 (35.9%) respondents. The gender difference might be explained by the fact that male knowledge holders in communities are more comfortable to talk than female knowledge holders who faced cultural restrictions. Females were forbidden from conversing or discussing with stranger males. Due to these factors, fewer women participated in the documentation. Similar finding was also reported (Chinsembu, Hijarunguru and Mbangu, 2015; Kankara, et al., 2022). The survey also revealed that most of the respondents are members of the higher age group; the age range of 50 and above is shown in Fig. 1. This is an indication that there is a wide gap of ethnomedicinal knowledge between the elderly and the younger generation. This, however, poses a serious threat to the indigenous knowledge because it may eventually be lost following the demise of the older generation (Abdulrahman, et al., 2018; Kankara, et al., 2015). Cultural changes brought about by modernization have

TABLE 1
DEMOGRAPHIC INFORMATION OF THE INFORMANTS

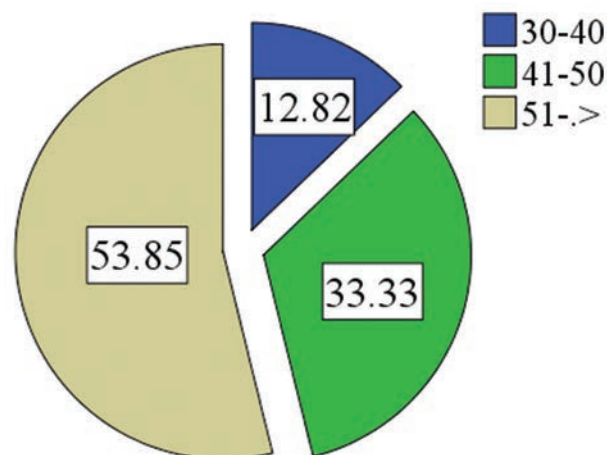| Parameters	Frequency	Percentage (%)
Gender		
Male	25	64.1
Female	14	35.9
Education		
None	15	38.5
Basic	8	20.5
Secondary	9	23.1
Tertiary	7	17.9
Experience		
5–10	9	23.1
11–20	19	48.7
61–80	11	28.2

Fig. 1. Experience of the respondents.

contributed to the younger generation's rejection of traditional values. Majority of the respondents had formal education (Table 1). Formal education has been cited as a major element in the decline of traditional knowledge. The study revealed that 64.1% of the respondents are traditional medical practitioners and 35.9% are indigenous people with traditional knowledge of using medicinal plants (Fig. 2). This revealed how reliable the information documented in the following studies are.

B. Documented Medicinal Plants

Natural products have been depended on for thousands of years due to their strong efficacy and safety records (Adhikari, et al., 2021). People and plants interact deeply based on their requirements. People used medicinal herbs to fight pandemics in the past (Arora, et al., 2010; Mukhtar, et al., 2008), and their use as a means of combating COVID-19 in the current world may have increased. Medicinal plants have been used by the Kurdish people since ancient times, and they play an important role in traditional medical practices (Kayfi and Abdulrahman, 2021). The current investigation was carried out to catalogue the various plant species that have been utilized for both culinary and medicinal purposes. Eighty-one medicinal plant species from 35 families were identified in this study (Table 2). The results showed that Lamiaceae was the taxonomic family with the most utilized plants (19.75%), followed by Apiaceae (9.9%), Fabaceae (8.6%), Asteraceae (7.4%), Lauraceae, Myrtaceae, Malvaceae, and Rutaceae (3.7% each), Anacardiaceae, Amaryllidaceae, Zingiberaceae, Burseraceae, and Moraceae (2.5% each), and finally, all remaining families (1.2% each; Table 2 and Fig. 3). The results of the study are not in line with the ethnobotanical studies carried out in Choman, Kurdistan, where they reported Asteraceae as the most abundant family in the area (Kayfi and Abdulrahman, 2021). As the informant mentioned during fieldwork, this family's widespread distribution is related to its members' resistance to drought and other environmental stresses. The vast bulk of them are imported from nearby nations. Many different illnesses, including diabetes, common cold, fever, cancer, ulcers, and body immune boosters, were

treated using the plants that were documented, as reported by the respondents (Table 2).

C. Parts of the Plant Method of Preparation, Administration, and Duration of treatment

Utilizing plant aerial parts are quite beneficial. Despite being well known for their medical properties, these plants' metabolic make-up is unknown to traditional practitioner (Vasquez, et al., 2013). Traditional healers use various plant parts despite the absence of genuine understanding about the contents of those parts. Among all plant parts, medicinal plants' leaves are utilized most frequently (25.8%) than any other portion, followed by seed (17.7%), flowers (12.9%), and fruits (12.1%), Fig. 4. Previously, similar parts of the plant were also reported (Abdulrahman, et al., 2018; Mahmoud and Abba, 2021; Mahmoud, Labaran and Yunusa, 2020). Whereas some studies have indicated various portions of the plant, these results are consistent with research done in other parts of the world. The fact that leaves are used may also be due to their greater availability and abundance in nature compared to other plant parts. In addition, they have
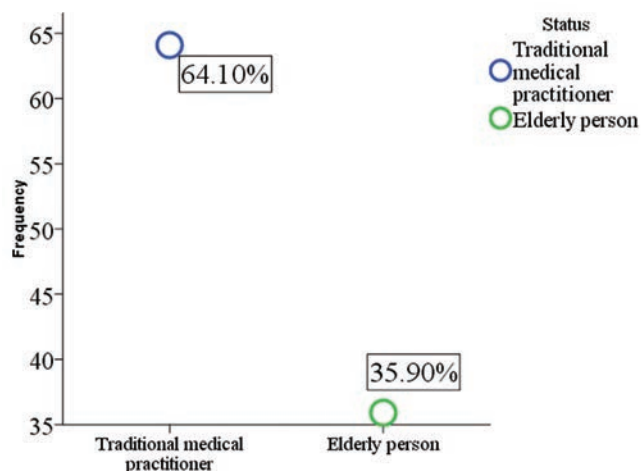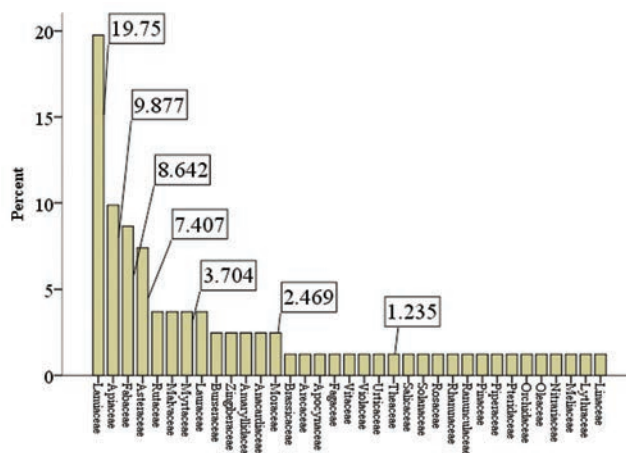

Fig. 3. Family distribution of the documented families.
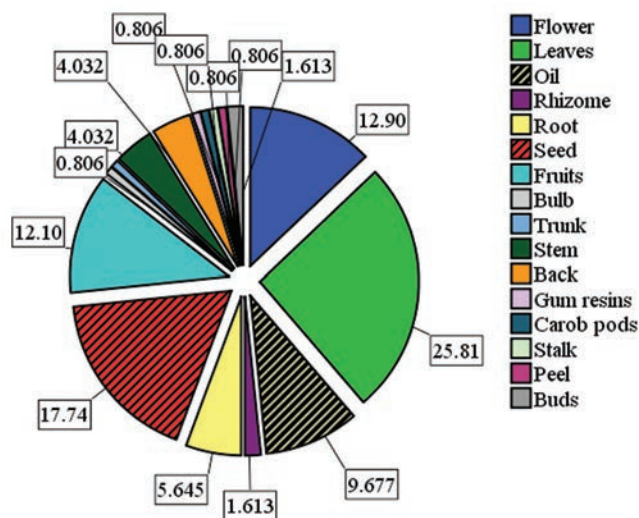

Fig. 2. Status of the respondents.

Fig. 4. Different parts of the plant used in the area for the management of COVID-19.

TABLE 2
Plants with the Recipe used for the Management of COVID-19 Complications in Kurdish

S/N	Family	Scientific name	Vernacular name	Part used	Preparation	Administration
1	Apiaceae	*Cuminum cyminum* L.	zîrre kemun	Seeds	Add teaspoon of cumin seed or 1/2 teaspoons of cumin powder into a boiled cup of water, add 1 teaspoon of fresh ginger or 1/2 teaspoons of ginger powder, heat the mixture for 5 min, leave it to rest for 5–7 min, then strain and get rid of the seeds	Oral (drink 3 times a day after each meal)
2	Apiaceae	*Carum carvi* L.	kerrawye (zeytî kerrawye)	Seeds and oil	1. Grind and place 1 tablespoon of seed in cattle and add 2 cups of water, boil it for 5 min, and leave it to rest for 15 min (if you let it to rest for 1 h is better or overnight will be optimal) after it is done, strain, and get rid of the seeds. You can add honey or a few leaves of mint or 1/4 teaspoons of mint powder for more benefit and taste. 2. Mix a few drops (1–2 drops) of caraway oil with a carrier oil (warmed sweet almond, coconut, avocado oils, or any vegetable oil), and apply directly to the neck, throat, and chest. Massage the area gently for better absorption.	Oral (drink it at evening before you go to bed. keep the rest in the refrigerator for the 2nd day). Dermal (Do this at the evening before bedtime).
3	Apiaceae	*Ridolfia segetum* (L.) Moris	toyi şwît (toyi şbît)	Seeds	Boiled the seed, then strain, and get rid of the seeds	Oral (2–3 times/day) for 2 weeks
4	Apiaceae	*Ammi visnaga* (L.) Lam.	xultan (toyi xultan)	Seed	1. Boiled water with tablespoon of seeds and cinnamon bark. 2. Place 1 tablespoon of seeds crushed and tied up in a small cloth bundle that is used for inhalation: for nasal congestion. Use a similar bundle, place it near the pillow while sleeping; for cough, drink hot water after chewing little	Oral (for 3 days 3 times a day)
5	Apiaceae	*Petroselinum crispum* (Mill.) Fuss	mu'dh nus	Leaves & stems	1. Add parsley into your food and salads. 2. Boil a cup of parsley leaves and stem in 1 L of water for 10–15 min, let it to seep and strain, and then drink a cup after each meal.	Oral
6	Apiaceae	*Pimpinella anisum* L.	yanison	Seeds	1. Place 1 teaspoon of Anise seed in one cup, water, boil for 2 min, let it to rest for 5 min, then strain and drink it. 2. Chew 1 teaspoon of anise seed for 5 min or grind the seeds and put it in a glass of warm water and drink it.	Oral
7	Apiaceae	*Foeniculum vulgare* Mill.	raziyane (şumerr yan ħbe ħlu zeytî)	Seeds	Taking 1 tablespoon of crushed fennel seeds in a glass cup, then pouring boiling water over it immediately, covering for 10 min, then strain and drink the extract.	Oral
8	Apiaceae	*Coriandrum sativum* L.	gjinîj (kezberre)	Leaves and seeds	1. Boiling 1/2 small cup of fresh leaves of coriander into 500 mL of boiled water for 30 min or 2. One tablespoon of powder coriander with boiled water for 10 min, then add 1 lemon and 1/2 teaspoon salt.	Oral (3 times a day after each meal, and the last cup you need to drink it before going to sleep)
9	Anacardiaceae	*Pistacia eurycarpa* Yalt.	talle binîşit yan binîşite tall	Trunk	1. Chew a small piece (like a chewing gum), this will enhance the secretion of saliva during coronavirus and solve the dry mouth and nose. 2. Grind and add 1 tsp to 1 cup warm water, drink 1 cup/day.	Oral
10	Anacardiaceae	*Rhus coriaria* L.	smaq	Seeds	1. Grind or use whole; mix 2 teaspoons of fresh or dry seed into 1 cup of warm water, leave it for 20 min. 2. Add sumac to your food and salads, it is good anti-inflammatory, fight coronavirus, and raise up the immunity of the body to resist the viruses.	Oral (1 time/day for 7 days)
11	Amaryllidaceae	*Allium cepa* L.	piyaz	bulb/ leaves	1. Add onion to your daily food and salad 2. Boil 1 chopped onion in 1 cup of water for 10 min. Leave it to cool a little and strain, then add a tablespoon of honey to it. 3. Roast or grill the onion for 5 min when the outer peel is burnt and is cooked	Oral (2/times a day).

(Contd...)

TABLE 2
(Continued)

S/N	Family	Scientific name	Vernacular name	Part used	Preparation	Administration
12	Amaryllidaceae	*Allium sativum* L.	sîr (yarrawhî sîr)	Fruits, Oil	1. Place 12 garlic cloves and 1 cup of honey in a jar to make fermented garlic honey. 2. Add garlic into your foods 3. N.B: The above mixture has immune boosting properties and fights the inflammations in the body. .	Oral (to an empty stomach) 1 time/day, then eat after 30-45 min.
13	Apocynaceae	*Hoodia gordonii* (Masson) Sweet ex Decne.	xuba	Seeds and flowers (oil)	1. Put 1 drop of oil in each nose twice for 40 days. 2. Take 1 teaspoon of oil twice/day.	Oral
14	Arecaceae	*Phoenix dactylifera* L.	xurma	Fruit	1. The patient needs to eat 3 or 5 dates (odd numbers only) every day. Dates are a very good remedy for coughing, phlegm, and bronchitis and inflammation during COVID-19. Note: Do not use this recipe if diabetics.	Oral (1 time/day for 7–10 days)
15	Burseraceae	*Commiphora myrrha* (Nees) Engl.	binêşite tall	Bark	1. Put some of the gum and put in warm water then used for gargling 2. Take 1 g of gum	Oral (3 times/day)
16	Burseraceae	*Boswellia sacra* Flueck.	dar bun (darî binêşitî kurdî)	Gum resins	1. Crush 1 tablespoon of frankincense, 1 teaspoon of fenugreek seed, 1 teaspoon of black seeds, 1/2 teaspoon of mastic and place them in a small pan, add 2 cups of water, put it on a low heat for 5 min, 2. Place 1/2 tablespoon of frankincense in a cup of cold water for 12 h, then stir, filter and drink.	Oral (1/day on empty stomach)
17	Brassicaceae	*Brassica nigra* (L.) K. Koch	xirtele (xerdel)	Seeds	3. Eat 5–10 seeds with 1–2 cups of water	Oral (1/7 days)
18	Asteraceae	*Bellis perennis* L.	gulle beybun	Flower	Boiling and filtrate or boiling with clove and filtration	Oral
19	Asteraceae	*Artemisia vulgaris* L.	şîĥ	Seed	Put in hot water then filtration	Oral
20	Asteraceae	*Saussurea costus* (Falc.) Lipsch.	qst hndî (qst hndî)	Root	1. Take 1 teaspoon of grinded dried roots, add it to 1 cup of warm water, then add 1 teaspoon of honey to get rid of the bitter taste. 2. Use thin slices of the roots of these herbs to prepare hot baths.	Oral (2/day)
21	Asteraceae	*Taraxacum fontanum* Hand. -Mazz.	tallîşk tallk	Roots, leaves, stems	Place 1 tablespoon of leaves and roots of fresh or dry dandelion in a boiled cup of water, let it boil for 3–5 min, then leave it for 10 min, then filter it.	Oral (1 morning and evening).
22	Asteraceae	*Chamaemelum nobile* (L.) All.	gulle ĥacîle (beybune rumanî)	flower	1. Put 1 teaspoon of chamomile in 1 cup of warm boiled water, leave it for 10–15 min, strain and drink. It twice a day. 2. Mix 1 teaspoon of chamomile with 1/2 teaspoon of ginger in 1 cup warm water (boiled before), let it seep for 5 min, strain and drink twice	Oral (1 day, morning and evening before 1 h of sleeping)
23	Asteraceae	*Matricaria chamomilla tzvelevii* Pobed.	gullu ĥlcîlu (bubunî ullmanî)	Flower	Place 1 tablespoon of powered flowers in a cup of boiled water or cold water, for 2–3 min.	Oral (2/day)
24	Fabaceae	*Cercis siliquastrum* L.	gulle wenewşe	Flower	Boil 1 cup of flowers with 2 litter of water	Oral (1/2 cup/day, until better for 5 days
25	Fabaceae	*Trigonella foenum*-graecum L.	şmillî yan ĥlbu (şîmbuluk)	Seeds	1. Add 1 teaspoon of Fenugreek powder to 2 cups of water and boiled for 1–2 min, then filtered and drink in multiple doses, 2. One tablespoon every h for 5 days or till feel better 3. Mix 2 tablespoon of fenugreek powder with 2 tablespoon of olive oil, take 1/2 tablespoon till feel better.	Oral (4 times a day)
26	Fabaceae	*Senna alexandrina* Mill.	sinemkî (zît)	Leaves, fruits, oil	1. Place 1 tablespoon of dried leaves or pods in hot water and boil it for 10 min, let it to seep for 3–5 min, strain and drink it. 2. Soak 7 leaves at night, and drink in the morning once every 4 nights,	Oral
27	Fabaceae	*Glycyrrhiza glabra* L.	mêkuk yan rhge buluk	Root	Place 1 teaspoon of root and put it in a cup of boiling water and leave for ten min, then drain and drink it. N.B: High pressure patients should not use it as it raises blood pressure.	Oral (3/day/day)

(Contd...)

TABLE 2
(Continued)

S/N	Family	Scientific name	Vernacular name	Part used	Preparation	Administration
28	Fabaceae	*Ceratonia siliqua* L.	doşawî xurinub	Carob pods	Take 1 tablespoon of carob extract twice a day after meals	Oral (twice a day)
29	Fabaceae	*Acacia senegal* (L.) Willd.	smẍî 'errebî	Bark	1. Chew the gum for 10 min five times daily for 7 days 2. Add 2 teaspoon of Arabic gum with 1 glass of water, and then drink it. 3. Add Arabic gum powder to food during cooking. N.B: Arabic gum relieving cough and sore throat, promote better digestion, anti-inflammatory, and enhance more secretion of saliva during dry mouth problem during COVID-19. N.B: Drink plenty of water after drinking it, as T contains a high amount of fiber.	Oral
30	Fagaceae	*Quercus cerris* L.	burruwî kurdî	Fruit	Put 5–6 oaks in a pot boiled in 4 cups (1 L) of water, let it seep and strain, then gargle and wash the mouth and throat.	Oral (2/morning and evening)
31	Lamiaceae	*Mentha piperita* L.	nu'nayi bîburî	Leaves, flower, oil	1. Add up to 7–10 drops of oil to 4 cups of boiling water; cover your head with a towel and inhale the steam through your nose. 2. Add 2 or 3 drops of oil to a cotton ball; breathe deeply for direct inhalation. (Apply this for 10 days till you feel better). 3. Add 1/2 cups of chopped peppermints and fresh leaves combined with 1 teaspoon of basil or cilantro to be added to your salad and food on a daily basis till cure.	Oral
32	Lamiaceae	*Glechoma hederacea* L.	lawlaw	Leaves	Take a medium spoonful of crushed leaves and putt it in a cup, leave it for 15 min, then drink.	Oral (1–2 cups/day, lunch, after dinner or when going to bed)
33	Lamiaceae	*Ocimum tenuiflorum* L.	şarreyẖan	Leaves, flowers	Boil leaves (1 cup) with 1/2 tablespoon of clove in 2 cups of water, when the water gets reduced to half, strain the liquid, and drink immediately.	
34	Lamiaceae	*Thymus vulgaris* L.	catire û zeytî catire	Stems, leaves, oil	1. Add thyme to your daily food and salads 2. Fill a teaspoon of the powered leaves to fill a cup of boiled water and leave for ten min, then drink a cup after each meal.	Oral
35	Lamiaceae	*Zataria multiflora* Boiss.	ze'ter (catire şîrazî)	Leaves, flower	Add 1 teaspoon of dried flower and leaves into 1 cup of warm water for 5 min. N.B: This mixture is good for calming coronavirus patients	Oral (2/day for 15 days).
36	Lamiaceae	*Lavandula latifolia* Medik.	zeytî gulle xezîm yan gulle erxewan	Flower, oil	Apply few drops of Lavender essential oil to the ears cotton stick and rub the nose, massaging it gently (the skin inside nose holes) do not put it inside the nose. N.B: Lavender oil is relaxing and giving the ability to have a good sleep and reduce stress and anxiety, depression, and insomnia during the coronavirus)	Dermal
37	Lamiaceae	*Thymus serpyllum* L.	(xuzamî yan lavînder)	Leaves, flowers	1. Add 1 teaspoon of dried thyme leaves into 1 cup of warm water, leave it for 15 min, and drink once a day before or after meal. 2. Gaggle 1 teaspoon of soaked thyme in warm water for 30 min, keep it for a few min if you can, then spit out. 3. Place 1 teaspoon dried leaves into the cup of boiled water then leave it to rest for 15 min, strain add honey or lemon for more benefit.	Oral
38	Lamiaceae	*Salvia officinalis* L.	giya me yrh mî (mwîrh mî yan gullu mwîrh mî)	Leaves, flowers	1. Place teaspoon of dried sage (2 teaspoons of fresh sage) and 1 cup of water in a small pan to cover, heat or boil it for 2–3 min, then let it cool down and seep for 10 min but with putting a cover, then seep and add 1 teaspoon of honey. 2. Add 1 tablespoon of dried sage leaves into 1 cup of boiled water, leave it for 5-10 min then seep to get rid of the leaves, and drink it once a day	Oral (1 time/day)

(Contd...)

TABLE 2
(Continued)

S/N	Family	Scientific name	Vernacular name	Part used	Preparation	Administration
39	Lamiaceae	*Clinopodium menthifolium* (Host) Stace	pung	Leaves	1. Make a tea from fresh or dried leaves. 2. Grind boil 1 tablespoon with 1 cup water.	Oral (1/day for 2 days)
40	Lamiaceae	*Mentha spicata* L.	awî ne'na	Leaves, oil	1. Wash fresh leaves or use dry leaves to make tea or use in food as much as you want. 2. Apply a few drops of mint oil to the ears cotton stick and rub the mixture onto the nose, massaging it gently for 5–7 days	Oral
41	Lamiaceae	*Rosmarinus officinalis* L.	rozimarî	Leaves	Add 1 teaspoon of dried leaves onto 1 cup of warm water, let it rest for 10–15 min, strain, and then add teaspoon of honey.	Oral (1 cup/day) at early morning before breakfast to 1 h.
42	Lamiaceae	*Ocimum basilicum* L.	reyḧan (şarreyḧan)	Leave, flower	Place 2–3 teaspoons of dried or 5 teaspoons of fresh leaves in a cup of boiled water and let it to rest for 5–6 min, strain, and drink.	Oral (once a day)
43	Lamiaceae	*Origanum majorana* L.	merde guş (merize)	Leaves	Place equal portion of 1 teaspoon of each of marjoram, melissa, and linden leaves into a cup of boiled water, drink it whereas it is warm. N.B: The above mixture will respiratory problems during coronavirus infection.	Oral (3 to 5 times a day)
44	Lamiaceae	*Salvia hispanica* L.	toyi şiya (bzur alşiya)	Seed	1. Place 1 teaspoon of seed into 1 glass of water, cover it and leave it for 20–30 min, then drink the jelly mixture to an empty stomach. 2. Add seed into your salad 3. Soak 1 teaspoon of seed in 1 cup of water, after 20 min, squeeze 1/2 lemon lime into it for more benefit and flavor. N.B: Chia seeds help in boosting immunity and transporting oxygen around the body, as well as reduce phlegm and inflammation during coronavirus infection.	Oral (1/day) most favorable time is on early morning on empty stomach before breakfast with 30–45 min
45	Lamiaceae	*Rosmarinus officinalis* L.	klîl alcbil	Leaves and oil	1. Add 1 teaspoon of dried flowers into 1 cup of warm water, leave it for 10–15 min strain and drink it. 2. Add a few drops of oil to a few drops cinnamon oil, then rub and massage the mixture on the chest and throat and neck. 3. Dilute drops of oil in boiled water about 500 mL^{-1} L of water and inhale the steam N.B: rosemary oil or tea is good for clearing out your lungs during COVID-19.	Oral (once per day) Dermal (Morning and before sleeping)
46	Lamiaceae	*Melissa officinalis* L.	giyalîmo (gîrawe û zeyt)	Leaves, oil	1. Place 1/2 of teaspoon of dried lemon balm herb in hot water. Steep and drink up to 4 times daily. 2. Take 1 capsule (300–500 mg) of dried lemon balm, 2–3 times daily. 3. Use a few drops of dried lemon balm oil on the skin of chest, neck, and stomach and massage gently (2 times, morning and before sleep)	Oral, Dermal
47	Lauraceae	*Laurus nobilis* L.	gwallî bwîyi (urq alẍar)	Leaves	Add 3–5 leaves with 1 teaspoon of ginger and 2 big Cinnamon sticks with 1 cardamom pod in chicken broth, eat every day till cure	Oral (once a day)
48	Lauraceae	*Cinnamomum verum* J. Presl	darçîn	Bark	1. Add 1 teaspoon of Cinnamon back powder to 1 big cup of boiled water, mix them well. Let the water simmer for 2–3 min and add 1 teaspoon of honey and drink it right away. 2. Mix 1 big teaspoon of cinnamon back powder with 2 teaspoons of honey to make a paste or added to a warm cup of water. N.B: Eating cinnamon with honey will fight coronaviruses, flu and inflammation.	Oral (1–2 on empty stomach)
49	Lauraceae	*Cinnamomum camphora* (L.) J. Presl	kafur	leaves and shoots (oil)	1. Put 3 drops of camphor oil with 2 drops of peppermint oil in a bowl, then put boiling water over it. Inhale the steam rising from the mixture of oils with boiling water. 2. Put in the palm of your hand 2 drops of camphor oil with 3–4 mL of apricot oil, 3 drops of pine oil, and add two drops of tea tree	Oral

(Contd...)

Table 2
(Continued)

S/N	Family	Scientific name	Vernacular name	Part used	Preparation	Administration
					essential oil, massage the chest and back area against the lungs 2–3 times a day, for 4–6 days. N.B: Using the above oil for massaging will improve breathing and clean the lungs. N.B: The above remedy is not recommended for people suffering from asthma.	
50	Linaceae	*Linum usitatissimum* L.	toyi kutan	Seed	Mix 1/2 tablespoon of flaxseeds, 2 teaspoons of sesame seeds, and 1 tablespoon of honey with a small amount of salt and combine them properly and consume this mixture daily before bed.	Oral (daily before bed) for 5 days
51	Lythraceae	*Punica granatum* L.	hh nar û pullkî hh nar	Peel	1. Make pomegranate juice everyday so as your body fights the viruses and resist infection when immunity is raised up. 2. Boil 1 cup of pomegranate peel in 1 L of water for 15 min, when it cools down, try to drink 1 cup per day or gargling the mouth and throat with it.	Oral
52	Moraceae	*Ficus sycomorus* L.	hencîrî kêwî	Fruits, leaves	Place 10 fresh or 3–5 dried fig leaves in a small pot containing 1 L of boiled water for 15 min and then remove the leaves and drink the tea.	Oral (1/day)
53	Moraceae	*Morus alba* L.	doşawî tû	Fruits	Eat the fruits early morning because it gives energy.	Oral
54	Myrtaceae	*Eucalyptus globulus* Labill.	(zeytî kalîptos) zeytî kafûr	Leaves/oil	Add a few drops of the oil in boiling water. Then cover the head with a towel and breathe in the steam.	Oral (once a day for 5–7 days)
55	Myrtaceae	*Myrtus communis* L.	murtk	Fruit, leaves, and branches	Take 1 tablespoon of the leaves and boil with a cup of water for 2 min, then cool filter.	Oral (1/day).
56	Myrtaceae	*Syzygium aromaticum* (L.) Merr. & L.M. Perry	mêxek	flower bud	1. Mix 1/2 teaspoon of clover powder with warm water and gaggle with it 2. Chew some cloves raw 3. Mix it with hot water and drink it early morning.	Oral
57	Meliaceae	*Azadirachta indica* A. Juss.	tusbîħ (şcrh alinîm yan sbħbħ)	Leaves, seeds	Boiled and take 2 tablespoon 3 times daily and birth with it twice	Oral, Dermal
58	Malvaceae	*Alcea kurdica* Alef.	gule hîro	Flower	Put 3–5 flowers in 2 cups of hot water and covered for 10–15 min, then strain and drink 1/2 cups.	Oral
59	Malvaceae	*Hibiscus sabdariffa* L.	kucerat (gulle kerkedî, çayi tirş yan gulle xene)	Flowers, leaves	1. Place 1/2 teaspoon flower or powder into 1 cup of warm water, shake well, leave it for 10 min to rest, strain or used without strain. 2. Place 1 cup of leaves/flower into 2 cups of water, add 1 teaspoon sugar, and 1 teaspoon of rose water, let it boil for 5 min, seep for 10 min, and strain.	Oral (1 cup at morning and a cup at evening)
60	Malvaceae	*Sphaeralcea angustifolia* (Cav.) G. Don	xuzamî (lavênder gulle xezêm û zeytî)	Flower	1. Inhale essential oils by sniffing the bottle directly or adding a few drops to a cotton ball. 2. Massage body of the COVID-19 patient with lavender oil every night until cure. N.B: Lavender oil can help the coronavirus patient to sleep better and speed the recovery process from viral and other infections.	Oral
61	Nitrariaceae	*Peganum harmala* L.	ħermell û zihytî ħerrmall	Seeds, bark, and root	One teaspoon of powdered seed and put it in a cup of warm water and drink it once until cure every day.	Oral 1/day/day and steaming once a day (at night before sleeping)
62	Oleaceae	*Olea europaea* L.	gelayi zeytun û zihytî zeytun	Leaves & seeds (oil)	1. Put olive oil into your food and salads. 2. Take 1 tablespoon of olive oil at early morning to an empty stomach. 3. Use a few drops of olive oil for massaging the chest, neck, back, and stomach. N.B: Olive mentioned in Verses of Holy Quran and Muhammad (PBUH) said "Eat olive oil and massage it over your body since it is a holy (Mubarak) tree". He also stated that olive oil cures 70 diseases.	Oral/tropical (once a day)
63	Orchidaceae	*Masdevallia molossus* Rchb.f.	doşawî tû	Fruits	Eat fruits at early morning on an empty stomach, give the patient power and energy for whole day	Oral (1/day).

(Contd...)

TABLE 2
(Continued)

S/N	Family	Scientific name	Vernacular name	Part used	Preparation	Administration
64	Pteridaceae	*Adiantum capillus-junonis* Rupr.	bîberî rhiş	Leaves, stalk	Mix 2 cups of fresh leaves or 1 cup of dry leaves with 4 cups of water, boil for 5 min, and leave it to rest for 3 h. Then, strain add 1 cup of honey.	Oral (take 1 tablespoon three times per day)
65	Piperaceae	*Piper nigrum* L.	bîberî reş	Seeds and fruits	1. Mix 1/2 teaspoon of black pepper and 1/2 of ginger (chopped), boil in 2 cups of water, then let it to rest for 5 min, shake well, and add 2 small teaspoons of honey. 2. Mix 1/2 of black pepper with 1/2 of turmeric to make a paste and eat it twice before meal. 3. Add black pepper to your food and salads.	Oral (twice/day)
66	Pinaceae	*Pinus sylvestris* L.	sinewber yan zeytî sinewber	Buds, leaves, oil	1. Soaked 1 cup of pine tree buds (ground buds) and place it in a liter of cold water and leave it for 3 h, filter and drink the extract. 2. Mix a few drops of oil with 1 tablespoon of coconut or almond oil and apply it directly to the chest, back, and neck. 3. Use 3 drops of oil for inhalation.	Oral (1 cup morning and evening)
67	Rutaceae	*Citrus limon* (L.) Osbeck	lîmo (ḥamz)	Fruits	1. Place 1 tablespoon of fresh Lemon juice to 1/4 teaspoons of turmeric powder in a glass of warm water, then add 3 teaspoons of honey. 2. Mix 1/2 of ginger (fresh or powder) with 2 tablespoons of fresh lemon juice in a cup or glass of warm water, add 1 teaspoon of honey. 3. Add lemon to your food and salads. 4. Use dried lemon peel for making a tea, to have it twice a day.	Oral
68	Rutaceae	*Citrus reticulata* Blanco	lalengî	Fruits	Mix 1 cup of cow yogurt with 1/2 cup honey with 3 fresh mandarin oranges and drink before meal twice a day.	Oral
69	Rutaceae	*Citrus aurantiifolia* (Christm.) Swingle	lîmo besirh	Fruits	Boil 1 fruit with 1 cup of water, leave it for 15–20 min. Strain and drink it two times a day after meal for 3 days. drink i	Oral.
70	Ranunculaceae	*Nigella sativa* L.	reşke (bereke) (zeytî bereke)	Seeds, oil	1. Mix 3 drops of black seed oil with 1 teaspoon of honey in a cup of warm water (You can add two tablespoons of apple cider vinegar). 2. Heat few drops of black seed oil and massage the chest, back and throat properly 3. Eat 1 teaspoon of black seed on 1 teaspoon of honey with empty stomach at morning	Oral, Dermal
71	Rhamnaceae	*Rhamnus alaternus* L.	sîdr yan wereqet sîdr	Bud and leaves	Boiled tablespoon of dried buckthorn or leaves and drink it twice a day for treating infection in your body. N.B: Rhamnus (buckthorn) treats sweating problems, helps in relaxation, solving sleeping problems & insomnia. It treats the psychological problems such as anxiety and tension during COVID-19 infection by the calming the nerve of the patient.	Oral
72	Rosaceae	*Malus pumila* Mill.	sirkeyi sêw (xellî sêw)	Fruit	Boil 3 cups of water, add 1 cinnamon stick, 1 peeled and smashed clove garlic, then add 1 tablespoon apple cider vinegar, finally place 1 tablespoon honey and stirring all ingredients.	Oral (drink a cup 3 times a day (before the meal, if you have a problem with your stomach then drink a cup of it after meals)
73	Solanaceae	*Capsicum annuum* L.	bîberî surî tîj	Fruits	Bring one cup of warm water, put a dash of cayenne pepper (1/8 of teaspoon), then add 1 teaspoon of apple cider vinegar, ¼ teaspoon of ginger, then add ¼ of turmeric (either fresh or powder), to it, you can add ½ lemon and 1 clove of smashed garlic, but it optional.	Oral (drink three times a day after each meal)
74	Salicaceae	*Salix alba* L.	darebî (şorre bî yan çnar)	Bark, Leaves	Boil bark in 4 cups of water for 30 min. N.B: The above recipe reduces the pain fever during COVID-19.	Oral (take 1 cup/3 times a day)
75	Theaceae	*Camellia sinensis* (L.) Kuntze	çayi kesk yan sewz	Leaves	Boil water and add 1/tablespoon of green tea leaves, then add 2 slices of fresh ginger, 1/2 lemon slice. Afterward, add 1/2 tablespoons of honey.	Oral (2/day 30 min before breakfast and dinner)

(Contd...)

Table 2
(Continued)

S/N	Family	Scientific name	Vernacular name	Part used	Preparation	Administration
76	Urticaceae	*Urtica dioica* subsp. *afghanica* Chrtek	ruwekî gezgeze (xerekeçuze) zeytî gezn	Leaves, root, oil	1. Make a tea of boiled leaves or roots for 15 min. 2. Heat several drops of nettle oil and massage your body, chest, back, muscles, and joints before bedtime to reduce the pain and give good sleep during coronavirus infection. N.B; The above recipe will open the pulmonary airways. The oil has analgesic and anti-inflammatory properties, so the nettle and the oil extracted from it becomes an effective treatment for joint pain and various body infections	Oral (once a day)
77	Violaceae	*Viola odorata* L.	gulle wenewşe	Flowers, root	Grind 1 teaspoon of the dried leaves and roots together and put in a cup of boiling water and left before using for 3–5 h, then drink 1 cup twice a day till feel better.	Oral (2 times/day)
78	Vitaceae	*Vitis vinifera* L.	mêwjî tirê reş	Fruits	Mix dates+raisins+dried figs as following: Make a syrup consisting of 1 cup of dried dates, 1 cup grams of black raisins, and 1 cup of dried figs. Place this mixture in a pot, then add a liter (or 5 cups) of water, leave it on the fire to boil until the contents of the pot soften, then it is eaten by dividing into three after each meal during the day. N.B: This treatment can be used for cough, taking out phlegm, and lung and breathing problems during coronavirus infection. Moreover, the grape is mentioned in Holy Quran. *This recipe should not be used by diabetics	Oral (3 times/day)
79	Zingiberaceae	*Curcuma longa* L.	zerdeçû	Rhizomes	1. Add 1 tablespoon of turmeric powder to a cup of milk, stirring and boiling it, then drink once a day. 2. Mix the powder of turmeric with honey and make a paste.	Oral (3/day on an empty stomach)
80	Zingiberaceae	*Zingiber officinale* Roscoe	zencefîl	Rhizome and Root	1. Boiling (Zingiber+Thyme+clove+saffron) then filtered it and drink the extract. 2. Mix 2 teaspoons of honey with 1/2 teaspoons of ginger powder or 1 teaspoon of fresh ginger and 1/2 teaspoons of turmeric with make a paste and eat it early morning before breakfast. 3. Boil 2 cups of water, put 1 tablespoon of fresh ginger or 3 teaspoons of ginger powder, add 3 teaspoons of fresh turmeric or 2 teaspoons of turmeric powder, mix them together, then squeeze 1/2 lemon, then add 1 tablespoon of apple cider, afterward add 1 teaspoon of honey.	Oral (twice a day)
81	Zingiberaceae	*Elettaria cardamomum* (L.) Maton	hîl	The seeds, pods and the oil from the Seeds	1. Put 2–3 seed or one pod of cardamom in a boiled cup of water and honey in it, leave it for 10 min then drink it after taking out the cardamom seed/pods. 2. Mix 1/2 teaspoon of cardamom powder with a cup of boiled water and drink it with a bit of sugar or 1 teaspoon of honey. 4. Chewing a cardamom pod or seed for 30 min	Oral

S/N=Serial number

been said to be more effective in the past since secondary metabolites are formed largely before being transferred to other areas of the plant (Kankara, et al., 2015). Respondents in the study reported using a variety of techniques to prepare the medicinal plants (Table 2). Water, honey, yogurt, vinegar, and many other substances were used as a diluent in the preparation of different recipes of medicinal plant species (Table 2). The results revealed oral treatment as the most popular form of administration (92.9%, Table 2 and Fig. 5). However, the complexity of the disease explains why a combination of approaches and oral treatments proved the most effective. The finding of the study is in line with other studies in the region and other parts of the world (Achour, et al., 2022; Megersa and Woldetsadik, 2022). According to the study's findings, respondents in the study region said that they prepared traditional herbal medicine using one, two, three, or more plant species, respectively (Table 2).

D. Quantitative Ethnobotany

A significant quantitative parameter describing the relative value of medicinal plant species in the management of COVID-19 in the research area. The RFC calculated for the

TABLE 3
QUANTITATIVE INDICES

S/N	Scientific name	RFC%	FL%
1	*Cuminum cyminum* L.	0.85	0.85
2	*Carum carvi* L.	0.6	0.7
3	*Ridolfia segetum* (L.) Moris	0.6	0.7
4	*Ammi visnaga* (L.) Lam.	0.9	0.82
5	*Petroselinum crispum* (Mill.) Fuss	0.7	0.4
6	*Pimpinella anisum* L.	0.85	0. 85
7	*Foeniculum vulgare* Mill.	0.7	0.4
8	*Coriandrum sativum* L.	0.5	0.6
9	*Pistacia eurycarpa* Yalt.	0.85	0. 85
10	*Rhus coriaria* L.	0.7	0.4
11	*Allium cepa* L.	0.6	0.7
12	*Allium sativum* L.	0.5	0.6
13	*Hoodia gordonii* (Masson) Sweet ex Decne.	0.8	0.7
14	*Phoenix dactylifera* L.	0.7	0.4
15	*Commiphora myrrha* (Nees) Engl.	0.9	0.82
16	*Boswellia sacra* Flueck.	0.8	0.5
17	*Brassica nigra* (L.) K. Koch	0.5	0.6
18	*Bellis perennis* L.	0.7	0.4
19	*Artemisia vulgaris* L.	0.6	0.7
20	*Saussurea costus* (Falc.) Lipsch.	0.9	0.82
21	*Taraxacum fontanum* Hand-Mazz.	0.5	0.6
22	*Chamaemelum nobile* (L.) All.	0.9	0.82
23	*Matricaria chamomilla tzvelevii* Pobed.	0.7	0.4
24	*Cercis siliquastrum* L.	0.5	0.6
25	*Trigonella foenum*-graecum L.	0.7	0.4
26	*Senna alexandrina* Mill.	0.8	0.5
27	*Glycyrrhiza glabra* L.	0.9	0.82
28	*Ceratonia siliqua* L.	0.5	0.6
29	*Acacia senegal* (L.) Willd.	0.9	0.82
30	*Quercus cerris* L.	0.7	0.4
31	*Mentha piperita* L.	0.8	0.5
32	*Glechoma hederacea* L.	0.5	0.6
33	*Ocimum tenuiflorum* L.	0.9	0.82
34	*Thymus vulgaris* L.	0.5	0.6
35	*Zataria multiflora* Boiss.	0.7	0.4
36	*Lavandula latifolia* Medik.	0.5	0.6
37	*Thymus serpyllum* L.	0.9	0.82
38	*Salvia officinalis* L.	0.5	0.6
39	*Clinopodium menthifolium* (Host) Stace	0.8	0.5
40	*Mentha spicata* L.	0.8	0.4
41	*Rosmarinus officinalis* L.	0.5	0.7
42	*Ocimum basilicum* L.	0.5	0.7
43	*Origanum majorana* L	0.9	0.82
44	*Salvia hispanica* L.	0.7	0.5
45	*Rosmarinus officinalis* L.	0.5	0.7
46	*Melissa officinalis* L.	0.6	0.7
47	*Laurus nobilis* L	0.8	0.5
48	*Cinnamomum verum* J.Presl	0.5	0.7
49	*Cinnamomum camphora* (L.) J.Presl	0.9	0.82
50	*Linum usitatissimum* L.	0.8	0.4
51	*Punica granatum* L.	0.7	0.5
52	*Ficus sycomorus* L.	0.6	0.7
53	*Morus alba* L.	0.8	0.4
54	*Eucalyptus globulus* Labill.	0.8	0.5
55	*Myrtus communis* L.	0.5	0.7
56	*Syzygium aromaticum* (L.) Merr. & L.M. Perry	0.8	0.4
57	*Azadirachta indica* A.Juss.	0.5	0.7
58	*Alcea kurdica* Alef.	0.6	0.7
59	*Hibiscus sabdariffa* L.	0.85	0. 85
60	*Sphaeralcea angustifolia* (Cav.) G.Don	0.5	0.6

(Contd...)

TABLE 3
(CONTINUED)

| S/N	Scientific name	RFC%	FL%
61	*Peganum harmala* L.	0.7	0.5
62	*Olea europaea* L.	0.8	0.5
63	*Masdevallia molossus* Rchb.f.	0.5	0.6
64	*Adiantum capillus-junonis* Rupr.	0.7	0.5
65	*Piper nigrum* L.	0.85	0. 85
66	*Pinus sylvestris* L.	0.7	0.5
67	*Citrus limon* (L.) Osbeck	0.6	0.7
68	*Citrus reticulata* Blanco	0.7	0.5
69	*Citrus aurantiifolia* (Christm.) Swingle	0.85	0. 85
70	*Nigella sativa* L.	0.7	0.5
71	*Rhamnus alaternus* L.	0.6	0.7
72	*Malus pumila* Mill.	0.85	0. 85
73	*Capsicum annuum* L.	0.9	0.82
74	*Salix alba* L.	0.5	0.6
75	*Camellia sinensis* (L.) Kuntze	0.9	0.5
76	*Urtica dioica* subsp. *afghanica* Chrtek	0.6	0.7
77	*Viola odorata* L.	0.9	0.82
78	*Vitis vinifera* L.	0.6	0.7
79	*Curcuma longa* L.	0.85	0. 85
80	*Zingiber officinale* Roscoe	0.6	0.7
81	*Elettaria cardamomum* (L.) Maton	0.9	0.82

S/N=Serial Number
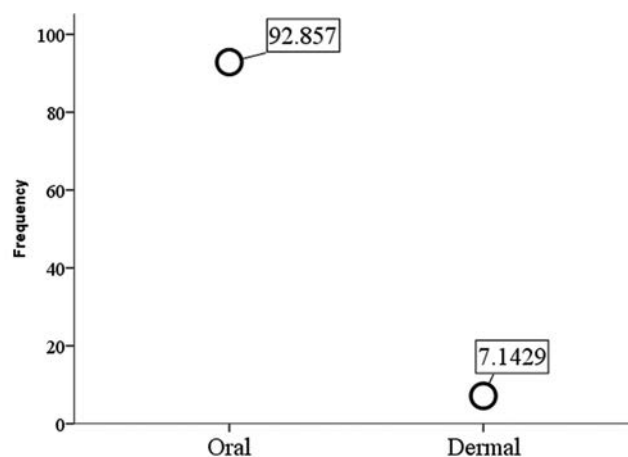
Fig. 5. Methods of administration of medicinal plants used in the area for the management of COVID-19.

different species ranged from 0.5 to 0.9, suggesting that the respondent made regular use of these plants. Whereas the FL value ranged from 0.4 to 0.85, revealing how effective the documented plant species are in the management of COVID-19 in the region. Consequently, high levels of RFC and FL in medicinal plants are suggestive of therapeutic efficacy and make them promising candidates for future, more in-depth studies (Kayfi and Abdulrahman, 2021; Mahmoud, et al., 2020).

E. Symptoms Considered by the Respondents

The primary signs and symptoms regarded by traditional herbalists as evidence of virus infection include breathing issues, obstructions of the airways brought on by mucus or phlegm, dizziness, fatigue, headache, fever, shortness of breath, insomnia, nausea, occasional vomiting, and diarrhea, cloudy

thinking, dementia, chest pain, digestive issues, sore throat, runny nose, muscle pain, chills, and eye redness (conjunctivitis). Since physical health has been harmed by the viral infection, patients' psychological conditions tend to be unstable, and they frequently experience anxiety, stress, and sadness.

IV. Conclusions and Recommendations

This is the first study of its sort to document the medicinal application to the treatment of COVID-19 among the Kurdish people of Kurdistan, Iraq. Eighty-one plant species were identified with RFC and FL values that were employed for conventional COVID-19 management. Traditional practitioners and elders of the area are the keepers of the region's traditional knowledge. It is suggested in this study that these plants may be the subject of additional research, including phytochemical, toxicological, and clinical studies so that we can have a better understanding of the safety and efficacy of their dosages. The indigenous knowledge of medicinal plants and their use must be protected and preserved, and this can only be done if people are made aware of the importance of these plants. In addition, it's important to inspire the local population to work on conservation.

References

Abdulrahman, M.D., Ali, A.M., Fatihah, H., Khandaker, M.M., and Mat, N. (2018). Traditional medicinal knowledge of Malays in Terengganu, Peninsular Malaysia. *Malayan Nature Journal*, 70(3), pp.349-364.

Abdulrahman, M.D., Bradosty, S.W., Hamad, S.W., Ibrahim, M.T., Lema, A.A., Sunusi, N. and Wada, N. (2022). Traditional methods for treatment and management of measles in Northern Nigeria: Medicinal plants and their molecular docking. *Ethnobotany Research and Applications*, 23, pp.1-18.

Abdulrahman, M.D., Zakariya, A.M., Hama, H.A., Hamad, S.W., Al-Rawi, S.S., Bradosty, S.W. and Ibrahim, A.H. (2022). Ethnopharmacology, biological evaluation, and chemical composition of *Ziziphus spina-christi* (L.) Desf.: A review. *Advances in Pharmacological and Pharmaceutical Sciences*, 2022, p.4495688.

Achour, S., Chebaibi, M., Essabouni, H., Bourhia, M., Ouahmane, L., Mohammad Salamatullah, A. and Giesy, J.P. (2022). Ethnobotanical study of medicinal plants used as therapeutic agents to manage diseases of humans. *Evidence-Based Complementary and Alternative Medicine*, 2022, p.4104772.

Adhikari, B., Marasini, B.P., Rayamajhee, B., Bhattarai, B.R., Lamichhane, G., Khadayat, K. and Parajuli, N. (2021). Potential roles of medicinal plants for the treatment of viral diseases focusing on COVID-19: A review. *Phytotheraphy Research*, 35(3), pp.1298-1312.

Arora, R., Chawla, R., Marwah, R., Arora, P., Sharma, R., Kaushik, V. and Tripathi, R. (2010). Potential of complementary and alternative medicine in preventive management of novel H1N1 flu (Swine flu) pandemic: Thwarting potential disasters in the bud. *Evidence-Based Complementary and Alternative Medicine*, 2011, p.586506.

Benarba, B. and Pandiella, A. (2020). Medicinal plants as sources of active molecules against COVID-19. *Frontiers in Pharmacology*, 11, p.1189.

Chinsembu, K., Hijarunguru, A. and Mbangu, A. (2015). Ethnomedicinal plants used by traditional healers in the management of HIV/AIDS opportunistic diseases in Rundu, Kavango East Region, Namibia. *South African Journal of Botany,* 100, pp.33-42.

Dogara, A.M., Hamad, S.W., Hama, H.A., Bradosty, S.W., Kayfi, S., Al-Rawi, S.S. and Lema, A.A. (2022). Biological evaluation of *Garcinia kola* Heckel. *Advances in Pharmacological and Pharmaceutical Sciences*, 2022, p.3837965.

Kankara, S.S., Ibrahim, M.H., Mustafa, M. and Go, R. (2015). Ethnobotanical survey of medicinal plants used for traditional maternal healthcare in Katsina state, Nigeria. *South African Journal of Botany*, 97, pp.165-175.

Kankara, S.S., Nuhu, A.I., Bindawa, K.A., Haruna, M.R., Bello, A. and Abubakar, I.B. (2022). Indigenous traditional knowledge of medicinal plants used for the management of HIV/AIDS opportunistic infections in Katsina State, Nigeria. *Ethnobotany Research and Applications*, 23, pp.1-17.

Kayfi, S. and Abdulrahman, M.D. (2021). Ethnopharmacology of plants in choman, the kurdistan region of Iraq. *Applied Biological Research,* 23(4), pp.322-330.

Khadka, D., Dhamala, M.K., Li, F., Aryal, P.C., Magar, P.R., Bhatta, S. and Shi, S. (2021). The use of medicinal plants to prevent COVID-19 in Nepal. *Journal of Ethnobiology and Ethnomedicine,* 17(1), pp.1-17.

Lim, X.Y., Teh, B.P. and Tan, T.Y.C. (2021). Medicinal plants in COVID-19: Potential and limitations. *Frontiers in Pharmacology*, 12, pp.1-8.

Mahmoud, A.D. and Abba, A. (2021). Ethnomedicinal survey of plants used for management of inflammatory diseases in Ringim local government, Jigawa state, Nigeria. *Ethnobotany Research and Applications,* 22, pp.1-27.

Mahmoud, A.D., Fatihah, H.N.N., Khandaker, M.M., Ali, A.M. and Mat, N. (2020). Ethnobotany of *Syzygium polyanthum* (Wight) walp in Terengganu, Peninsular Malaysia. *Journal of Agrobiotechnology*, 11(2), pp.39-47.

Mahmoud, A.D., Labaran, I. and Yunusa, A. (2020). Ethnobotany of medicinal plants with antimalarial potential in Northern Nigeria. *Ethnobotany Research and Applications*, 19, pp.1-8.

Megersa, M. and Woldetsadik, S. (2022). Ethnobotanical study of medicinal plants used by local communities of Damot Woyde district, Wolaita zone, southern Ethiopia. *Nusantara Bioscience,* 14(1), pp.104-116.

Merouane, A., Fellag, S., Touaibia, M. and Beldi, A. (2022). A ethnobotanical survey of medicinal plants consumed during holy month of Ramadan in Chlef region, Algeria. *Ethnobotany Research and Applications,* 23, pp.1-14.

Mukhtar, M., Arshad, M., Ahmad, M., Pomerantz, R.J., Wigdahl, B. and Parveen, Z. (2008). Antiviral potentials of medicinal plants. *Virus Research*, 131(2), pp.111-120.

Vasquez, J., Jiménez, S.L., Gómez, I.C., Rey, J.P., Henao, A.M., Marín, D.M. and Alarcón, J.C. (2013). Snakebites and ethnobotany in the eastern region of Antioquia, Colombia-the traditional use of plants. *Journal of Ethnopharmacology*, 146(2), pp.449-455.

Landfill Site Selection for Solid Waste Using GIS-based Multi-Criteria Spatial Modeling: TaqTaq Sub-district in Iraqi Kurdistan Region as a Case Study

Rostam S. Aziz

Department of Geography, Faculty of Education, Koya University,
Koya KOY45, Kurdistan Region – F.R. Iraq

Abstract—This study gains insight into landfill sites with the observance of all the political, economic and environmental difficulties for the implementing appropriate site measures by adopting a collection of geospatial technique and weighted linear combination (WLC) in TqaTaq sub-district. In the current study, there are several areas determined as appropriate sites for landfill location. In this study, the criteria of distance from the roads, the city center, rivers, surface water, and land use map were used. According to this analysis, only 25.21% of the TaqTaq sub district is suitable for a landfill. Thus, basing on the findings, 20.93% of the concerned sub-district is regarded as least adequate site for this mission, whereas only 3.25% of the area is regarded as moderate suitable. Thus, this study has found out that 1.03% area is the most suitable. The majority of suitable area was located in the North of the Town, where waste production is more than other locations. It should be noted that based on the outcome of this study, the amount of waste produced in the TaqTaq Town for the next 10 years, from 2022 to 2032, is predicted to be about 4080 tons. According to the density calculated for the waste of this area and considering the height of 4 m for the landfill center, in the next 10 years, about 3000 m² of land is required for the landfill location. Since the suitable area found in this research is about 15 hectares.

Index Terms—Geographic information systems, Landfill sitting, Multi-criteria evaluation, Solid waste management, WLC process

I. Introduction

The level of solid waste is daily growing due to the increase of human population and their activities and consumes. Hence, it is regarded as one of the problematic issues in almost all urban areas of the world. These increases in human activities, consumes, and municipal solid waste (MSW), causing a severe menace to both environment and human health. This can be obviously observed in developing countries where tremendous amount of MSW are dumped randomly. Consequently, this randomness in burying solid wastes has negative impact on all the natural issues, especially on water resources that are used for human consumption. Thus, it considerably affects the health of the consumers, especially on those people who have incredible exposure to it (Omang, et al. 2021).

There are many definitions of MSW given by many researchers, (Schübeler, Christen and Wehrle, 1996) states that MSW is regarded as a rejected from of domestic garbage, as well as non-dangerous solid wastes that came from industrial, commercial, health and institutional foundations in addition to the waste productions from other outdoor human activities. States that nearly 80% of the MSW that has been produced all over the world are buried under land. Although the components of land filled MSW are to a great extent different among countries with various cultural and economic frameworks, they are to some extent similar in some organic and non-organic materials. For instant, organic materials often consist of paper, wood, clothes, food, and garden rubbish, whereas non-organic substances include building debris, electronic and mechanical tools, metal, glass, and plastics. Organic material which is naturally putrefy by microbes into carbon dioxide (CO_2) and methane (CH4) to take part with the rate 6–18% in global methane production (Bingemer and Crutzen, 1987).

The most prevalent and dangerous source that intimidate the world environmental health is municipal solid waste. On this basis, it is crucial that engaged systems of waste handling be noticed within the way toward obtaining constant progress. The ingredients of (MSW) handling minimize the waste, reutilizing, recycling, power recovery, incineration, and land filling (Limoodehi, et al. 2017). Although the abovementioned techniques are utilized, and if the right understanding for waste declining is adopted, the availability of a healthy landfill is important to a MSW administrative system. Hence, healthy landfilling is crucial element of MSW management system (Kouloughli and Kanfoud, 2017).

Adequate site nomination of landfills may have a great contribution in decreasing the environment contamination.

Concerning the issues connected to the nomination of appropriate landfill location it can be deduce that making mind up is not easy in this respect, so the speed and accuracy are considerably declined as a result of the absence of standards (Moeinaddini, et al., 2010).

Landfill has become very hard process to proceed, because land is among the most precious and restricted resources that should be exploited cleverly. Appropriate sites should be selected on the basis of technical, economical and socio-environmental aspects and meet the recommended global criteria. The utility of MCE method looks very suitable. Making use of geography information system provides the spatial decision support systems (SDSS) in the process of determining appropriate landfills in terms of health, economy, social, and environment.

In all the urban areas in KRG, the waste proportion has risen suddenly over last decade. Therefore, this study aims at considering TaqTaq sub-district for the foundation of disposal centers for this sub-district. Therefore, to fulfill this purpose some significant parameters in disposal center are introduced (*We can list distance from roads, rivers, surface water the city center, and a map of land use as some of these factors*(. Later, the factor maps of studied area are made ready. There are several map combination processes which are in the form of Waited Linear Combination (WLC), index overlay combination. In this study, WLC considerably was utilized for merging maps. Ultimately, the appropriate sites for the establishing disposal center are nominated.

II. Materials and Methods

A. Study Area

Taq Taq Sub-District is located in Koysinjaq district, Erbil Governorate about 31 Km southwest of Koi Sanjaq city, about 61 Km northeast of Kirkuk, and 85 Km southeast of Erbil, Fig. 1, includes 19 villages. The study area has shared borders with lower zap on the east, Erbil City on west and koysinjaq town on north and Kirkuk City on south. The study area approximately located between 44° 32′ 46" and 44° 17′ 35" eastern longitudes and 36° 2′ 44" and 35° 50′ 49" northern latitudes an area of around 280 km² which constitutes 13.6% of the gross KoiSanjaq district area of the highest point of the field area is located at an elevation of 605 m and the lowest point 300 m above sea level. The rainfall average of 350 mm (Average of rainfall in the study area 1995 – 2018 is taken from TaqTaq Meteorology Department). The main land covers the following types of pastures, fallow lands, agriculture, water, and inhabitant zones. Economic flourishing in the last decade has resulted in a considerable raise in inhabitants and consequently in solid waste production. The population of TaqTaq sub-district is nearly (29078) persons, which equals 26.5% of the total population of Koi Sanjaq district statistics and expected to be (37721) in 2032 account by a population increase proportion is about 2.8% per annum concerning the distribution of population according to rural community and civilized community in the study area. The number of TaqTaq town population is (23812) persons, but the number of its rural population is (5264) persons Table I.
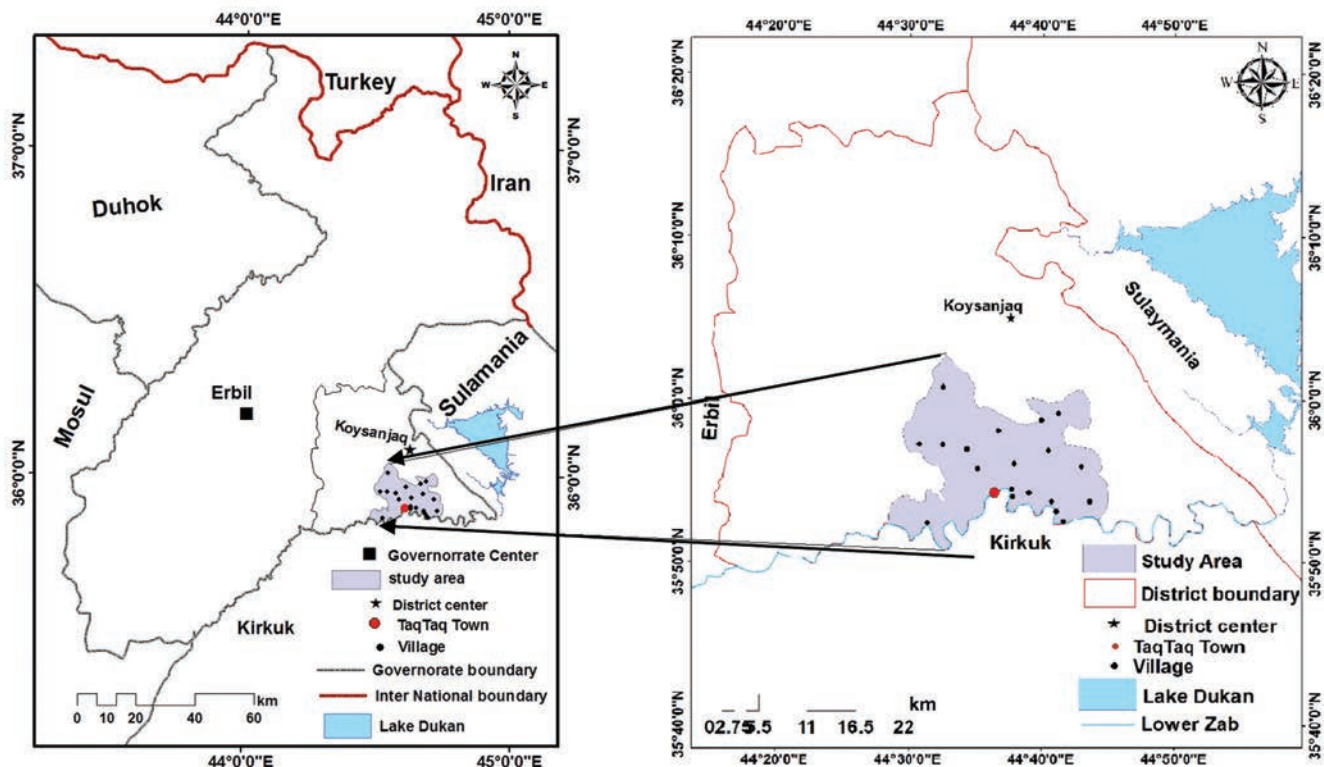

Fig. 1. Location map of the study area.

TABLE I
POPULATION OF TAQTAQ SUB-DISTRICT BY RURAL AND TOWN AREAS (2009–2032)

Sub-distric	2009	2014	2022	2032
TaqTaq Town	17004	19394	23814	31388
Rural	3434	4220	5264	6938

It is rated that solid waste production of TaqTaq is nearly 30 tons/day. The available solid waste landfill locations have been in a constant service for nearly 10 years in a way that the current waste burying places is located 1 km a way in the southeast of TaqTaq Town. In this area, probably 30 tons of waste is the daily produce (Taq Taq municipality, total quantity of MSW Generation [2016–2020]). As a consequence of the absence of suitable measures to cover the waste with soil in a standard way, there is a constant dissemination of parasitic and infectious diseases in the area, and this will be also a source of disturbing of the travellers who make the use of the main road (TaqTaq - Kirkuk) or the secondary roads within 2 km from the site, its bad smell can be felt. Because it is to a great extent close to the residential area and this causes health problems and sometimes social problems for the inhabitants. Thus, the available solid waste landfill site is not adequate. To Protect and support ecosystem in terms of ecology, indicating a suitable location is crucial, and this requires a sort of cooperation and collaboration of other concerned organizations. In this respect, a study to gain insight into the appropriateness of landfill is recommended to be placed within 5–10 km of the TaqTaq Town as it has been approved and addressed in this article.

B. Calculating the Amount of Waste and the Required Land Area

This study aims at selecting an appropriate site landfill as well as figuring out the number of land sites needed for burying rubbish. This can be fulfilled by considering some factors such as garbage production rate, population, and density of compact material. Thus, number of residents, annual production of rubbish, height, and shape of land in which the waste will be buried in are the main issues that should be taken into account. For estimation of inhabitants of TaqTaq, we depend on the latest public demography with the adaptation of the below equation.

$$P_t = P_o \cdot (1+r) \cdot t \qquad (1)$$

In this equation P_t, P_o, and r are the number of the residents of TaqTaq Town in 2009, 2014, and increasing proportion of population, respectively; thus, the number of population has been anticipated as 23814 by a 2.8% a population increase proportion.

Regarding to the amount of rubbish for one person per day is 1.3 kg/day (Data Collection available by Koi Sanjaq municipality 2016–2020) with average density of 430 Kg/m³, where the weight and volume of rubbish will be about 4080 tons and 1186 m³, respectively, during the horizon of 10 years.

American Planning Association (APA) states that the following empirical formula has been offered to determine the required space for sanitary waste disposal (Krizek and Power, 1996).

$$V=R/D \ (1-P/100) + CV \qquad (2)$$

V=Required space (volume) during the year
CV=The volume of required covering soil
D=Average density of waster
R=Rate of per capita (waste) production
P=Percent of waste reduction due to condensation.

In the similar way, the amount of produced waste during (2022–2032) is calculated. Next, the amount of produced waste of each estimated year were added and the divide it by the density of regional waste 430 Kg/m³, the volume of waste for the following 10 years is 4080438 kg which is equivalent to 4080 tons. Based on the density *waste* of TaqTaq Town (0.430 kg/m³) and the formula for calculating the amount of waste in the following 10 years is as follows:

$$P = M/V$$
$$V = m/v = 4080438/430 = 9489 \ m^3$$

With regard to the selected sites which are valley shape –if square –like base of waste disposal sites is taken into account whose side is on the ground level and its height has been confined from the level, of the hole to the high point of the level, we can have estimated the concerned volume, using by covering soil that is 1–4 cm is as follows:

V = R/D (1-P/100) + C V (Eq. 2)
V = Required space (volume) during the year
CV = The volume of required covering soil
D = Average density of waster
R = Rate of per capita (waste) production
P = Percent of waste reduction due to condensation
CV = The volume of required covering soil=9489/4=2372
V = Required space (volume) during the year = 2372 + 9489 = 11861 m³ (the estimated volume of the following 10 years is 11861 m³) = 11861/4 = 2965 m².

C. Analysis Method

The identification of suitable landfill location needs very influential criteria assessment to increase social, economic, environmental, and health costs (Siddiqui, et al., 1996). The methodology makes use of a developed digital GIS database, in which spatial information is provided. Due to the diversity in the measures on which criteria (*distance from the city center, roads, rivers, surface water, and land use map*) are built, it is essential that factors be standardized before combination. All the contributes of input data were given scores. The scores show land composition for siting a

landfill to extend from 0 to 5. A score of 0 points out no constraint, and a score of 5 shows a total constraint. Weight is also employed to these maps in a way that the total weight should be increased up to 100% to make the output map to be meaningful and consistent, and the contribute scores must be selected with the use of a scheme that was similar for each map. In this query, the maps of input data were not paid enough attention, because some factors were more effective than others while choosing appropriate landfill locations. Nevertheless, the significance of each factor may be different from one study area to another based on the local condition of each. The major criteria used in this study are shown in Table II. For further clarification, all the maps are geographic data layers stored in the GIS raster-based with the 30 m × 30 m grid cells. After taking out matrix relative importance and criteria weights, pair-wise comparison consistency should be recognized. The process of consistency index is known as (CR). Consistency ratio illustrates the probability of matrix ratio haphazard producing. The consistency rate should be <0.1; otherwise, there will be a requirement to reassess the relative significance. Whereas the consistency proportion is lower than threshold limit, accounted weights are affected by the criteria (factors) map layers Table II.

TABLE II
CRITERIA WEIGHT AND RANKING

Criteria	Weight	Buffer zone	Ranking
Slope	0.0525	0–10	5
		10–20	4
		20–30	3
		30–40	2
		>40	1
River	0.11	0–1000	1
		1000–1500	2
		1500–2000	3
		2000–2500	4
		>2500	5
Water bodies	0.15	0–250	1
		250–500	4
		500–1000	5
		1000–1500	3
		>2500	2
Land use	0.1125	Water (pond and stream)	0
			0
		Wheat and barley	1
		Orchard	1
		Plow and barley	5
		Forestry	6
		Rangeland	7
Road	0.08875	0–250	1
		250–500	3
		500–1000	5
		1000–1500	4
		>1500	2
Residential	0.21	0–250	1
		250–500	3
		500–1000	5
		1000–1500	4
		>1500	2
Boundary of town	0.25	0–2000	1
		2000–3000	5
		3000–3500	4
		3500–4000	3
		>4000	2

D. Uses of GIS in Waste Management

It is evidence that one of the crucial utility of GIS is the presentation and the analysis of data to aid environmental decision-making. A decision is recognized as a selection among the choices, where the alternatives may be diverse activities, locations, objects, etc. For instance, a concerned establishment may need to determine the best location for dangerous waste stuffs, or probably determine the areas which will be best choice for a new progress. The function of GIS in (SWM) is also observable as many other issues as its planning and operations are incredibly dependent on spatial data. Overall, GIS highly contributes in achieving account data to make collection procedures easier. Therefore, issues such as customer service, analyzing optimal locations, the transfer system across different places in an urban area to the landfill, and vice versa are also performed by GIS. Thus, GIS is a tool that not only saves time and money in the process of selecting appropriate place, but also supplies a digital data base for upcoming consideration of the site (Tomlinson, 1990).

E. Waited Linear Combination Model

Weighted linear combination method (WLC) is the most popular and scientific method in assessing the multi scale evaluation. This method usually draws on the content of weight average. Analyzer or responsible figures should base on "relative importance" weighted directly to the scales. Then through multiplying relative weight in feature value, ultimate amount can be gained for each choice (Zadeh, et al., 2013). After determining the ultimate value for each option, alternatives which have higher value, will be most adequate choice for the intended purpose. A GIS based MCE technique, using WLC analysis, gaining insight into a number of possible choices for the problem related to the process of selection, and also take into account multiple criteria and conflicting objectives. With making use of GIS for site determination, data were taken from various sources, and kept in the GIS system. The data that are utilized in this study are useful for software uses a weighted sum analysis that is act as a WLC analysis (Al-Hanbali, Alsaaideh and Kondoh, 2011). A weighted sum analysis provides the ability to weight and gather multiple inputs to conduct an integrated analysis, that is, it joins together multiple raster inputs and represents multiple factors of different weights or relative importance. It is one of most effective and popular methodologies that are utilized for indicating suitable location in general, and for indicating solid waste disposal locations in specific.

III. RESULTS AND DISCUSSION

Based on the findings derived over conducting this study, an appropriate landfill should be located and constructed to fulfill the required criteria for avoiding pollution of the soil, groundwater, or surface water to guarantee enough collection of filtration. In addition to that, a landfill site must be placed away from residential density for decreasing pollution influence to residential health. Moreover, also we should take the distance between the landfill and the available road access to it into consideration because it is better to be located close

to the existing roads for making sure that the transportation facility is available instead of constructing new roads (Gorsevski, et. al., 2011). Eight appropriate criteria are utilized in this work and an exclusive map is drawn for each suitable criterion, and these result in designing the final composite map that is produced by WLC process. The criteria analysis for landfill site determination is illustrated in paragraphs below.

A. Slope

Land structure is regarded as an essential issue in the landfill selection. Land structure is assessed by its slope degree as it is shown either in percentage or in degrees. Steep slopes are not appropriate for landfill foundation where the construction expenses of excavation usually rise in higher slopes. However, an adequate slope of land is significant in avoiding the leach ate. The slope layer map was achieved from the TaqTaq Sub-district area DEM map drawing on pixel size in percentage. The slope of more than 40% was regarded inappropriate. The slope of <10% with 5 value was regarded as an appropriate selection, Fig. 2a, (Rezaeisabzevar, Bazargan and Zohourian, 2020).
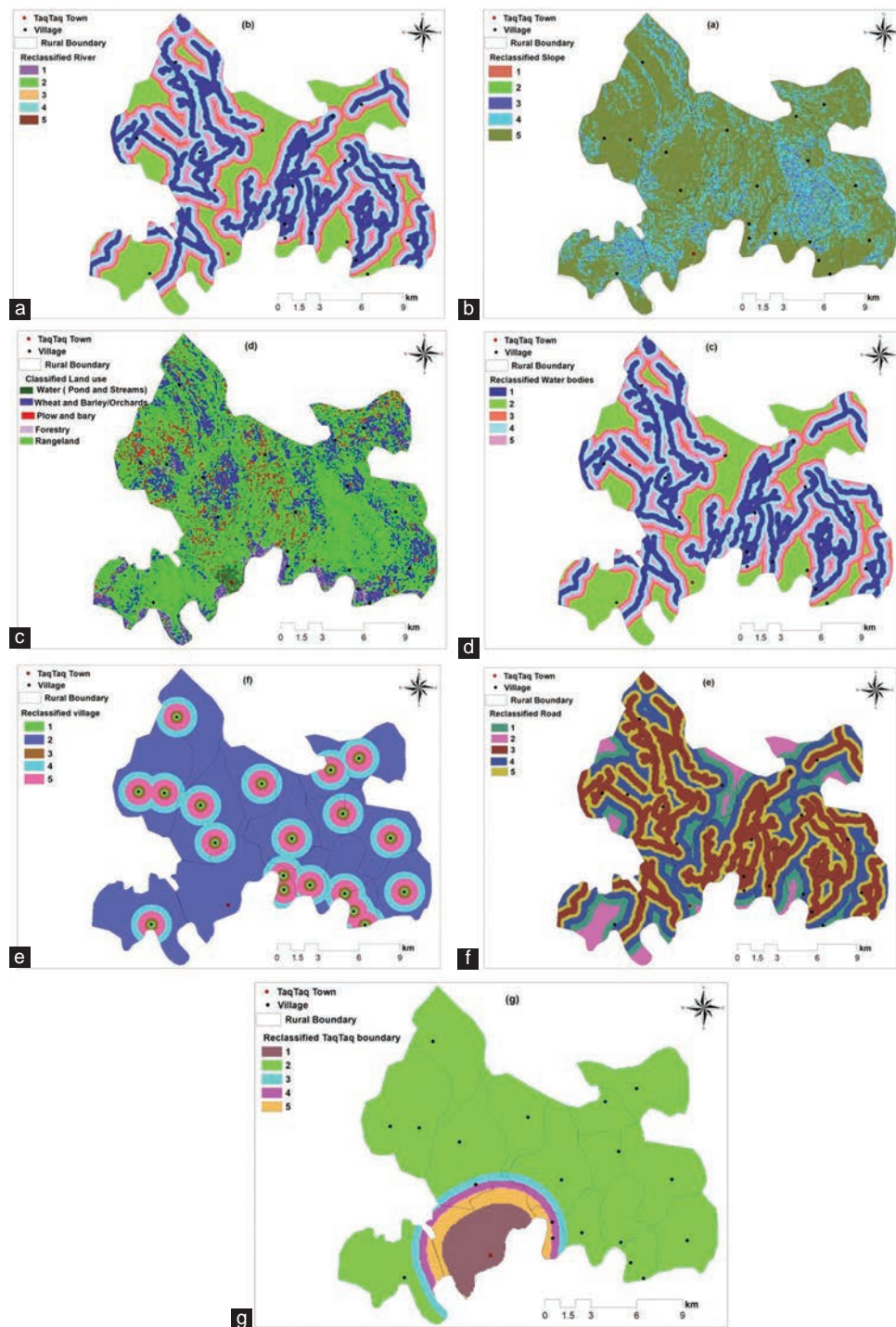
Fig. 2: (a-g) Geospatial data Maps in the study area.

B. Distance to river

Landfill location should not be built close to river because it usually leads to water pollution which can occur such as contaminating the water as a consequence of the arisen of leachate from the landfill site area. More than 1000 m buffer zone was constructed for each of the water body in the study area, Fig. 2b, Rezaeisabzevar, Bazargan and Zohourian, 2020).

C. Water bodies

A landfill must be far enough from surface streams and rivers. Proximity to sub river and stream was a significant criterion to evaluating the landfill site. Landfills emits noxious gases and leach ate that make them inappropriate to be near to water stream as to avoid its health dire consequences and the spread of disease, because these influences can extend to areas up to 250 m distance away from water streams, Fig. 2c.

D. Land use

The LU//LC map was taken from Land Cover Categorization in Rural Area of Taq Taq sub-district (Aziz, Hamadamin, and Omer, 2019). Various sorts of land utility in the study area have been adopted. Thus, values were determined to each land use sort depended on its adequacy level. Areas with water (pond and stream), wheat, barley, and farms were identified with 0 scores whereas areas such as vacant and agricultural lands (rangeland, plow, and barley) were given the scores of 6 and 7 successively, Fig. 2d, (Rezaeisabzevar, Bazargan and Zohourian, 2020).

E. Road network

To avoid constructing special roads for landfill access is affordable. Hence, the assigned locations must be near to the main roads to avoid the interference of solid waste transferring vehicles with the public vehicles. The minimum

pixel value (0 value) assigned to 250 m distance from available roads. Moreover, the distance with more than 1500 m is regarded unsuitable because of demanding more transportation costs, Fig. 2e.

F. Residential areas

Landfill site must be placed far from residential areas; otherwise, it annoys inhabitants who are living close to the area. Therefore, the distance of the selected areas for burying wastes must be between 500 and 1000 m away from the inhabitants in the rural areas. However, if the selected location exceeds more than 1500 m away from rural areas, it is regarded as inadequate, Fig. 2f, (Aziz, Hamadamin, and Omer, 2019; Rezaeisabzevar, Bazargan and Zohourian, 2020).

G. Distance from TaqTaq boundary

The safe distance from urban boundary must be at least 2–5 km. Minimum distances from the residential areas in the study area were assigned as at least 2 km for urban boundary is suitable. However, the distance with more than 4 km is regarded unsuitable, because as much the distance of the landfill is far from town, the costs of transportation will be increased. Although the cost is an important issue, considering environmental and safety issues should be inevitably observed, Fig. 2g.

H. Geospatial Operation for Site Selection

To do the process of site selection of suitable lands of land fill, a method of combining layers by WLC Model was used. Then WLC was adopted for indicating proper site as the best location. Finally, by WLC process, an area has been found. In general, in WLC overlay, the sites with high appropriateness were determined as suitable sites for landfill foundation and suitable sites are illustrated in Fig. 3. According to
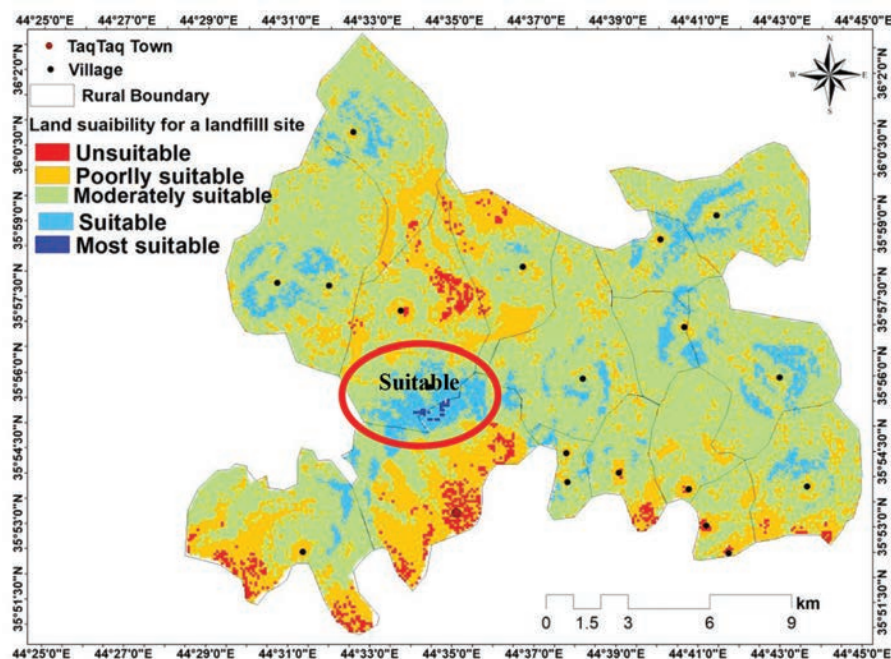
Fig. 3: Landfill site suitability in the study area.

hydrologic and hydrogeology factors, adequate locations have more ideal conditions than other locations due to its distance from surface water with the range of 1500 m and ground water with the range of 300 m. On land use suitable site is placed in dray farmland class and is far away from fertile agricultural zones and farms. Moreover, this study also recommends that the nominated landfill locations should be at least 2000 m. far from TaqTaq Municipal boundary. Finally, landfill sites were obtained for landfill in TaqTaq sub-district including rural area; the most recommended sites are illustrated in Fig. 3.

IV. Conclusion

This study demonstrates both GIS, WLC have been gathered in Model to implement a multi criteria evaluation techniques in determining the adequacy in the selection of right solid waste landfill site. The location appropriateness was determined by employing seven criteria. The achieved results from experts' view state that among sub criteria, water body, groundwater, slope and within social and economic sub criteria, residential areas, land use, and available network roads are significant successively. The extension of some proper locations is more than needed landfill site, so it is possible to organize the related equipment's of recycling in addition to address solid waste better despite decreasing transition costs.

Adopting GIS and multi criteria decision evaluation is intended in the process of site determination problems. GIS is flexible in considering criteria and it is possible to improve this method by considering other influential criteria. A GIS is the most useful tool that is helpful in the indication of appropriate locations for landfill sitting. In addition, the possibility of GIS usage in WLC and having cell information allows that the traits of the TaqTaq sub-district land structure are examined accurately at small area.

A multi criterion decision analysis also highly contributes in the process of observing various criteria regarding site determination assessment problem as a result it helps decision makers in their selections. GIS gathering with decision analysis as decision assisting system can help responsible figures in each site selection problem as an effective tool. Therefore, for making the ultimate decision, field investigation of the recommended landfill sites should be proposed with the consideration of the costs, social, administrative, and political aspects. The recommended procedures can help responsible figures in the disposal

and solid waste management activities.one site that is very adequate for landfilling, but it is about 3 km away from TaqTaq Town.

References

Al-Hanbali, A., Alsaaideh, B. and Kondoh, A. 2011. Using GIS-based weighted linear combination analysis and remote sensing techniques to select optimum solid waste disposal sites within Mafraq City, Jordan. *Journal of Geographic Information System*, 3(4), p.267.

Aziz, R.S., Hamadamin, R.R. and Omer, M.K. 2019. August. Land Cover Classification in Rural Area of Koya District Using Remote Sensing and GIS. In *Koya on the Road of Civilization, KRC 2019 International Conference*. Koya University, pp.355-362.

Bingemer, H.G. and Crutzen, P.J. 1987. The production of methane from solid wastes. *Journal of Geophysical Research Atmospheres*, 92(D2), pp.2181-2187.

Gorsevski, P.V., Donevska, K.R., Mitrovski, C.D. and Frizado, J.P. 2012. Integrating multi-criteria evaluation techniques with geographic information systems for landfill site selection: A case study using ordered weighted average. *Waste Management*, 32(2), pp.287-296.

Kouloughli, S. and Kanfoud, S. 2017. Municipal solid waste management in constantine, algeria. *Journal of Geoscience and Environment Protection*, 5(1), pp.85-93.

Limoodehi, F.A., Tayefeh, S.M., Heydari, R. and Abdoli, M.A. 2017. Life cycle assessment of municipal solid waste management in Tehran. *Environmental Energy and Economic Research*,1(2), pp.207-218.

Moeinaddini, M., Khorasani, N., Danehkar, A., Zienalyan, M. and Darvishsefat, A.A. 2010. Siting MSW landfill using weighted linear combination and analytical hierarchy process (AHP) methodology in GIS environment (case study: Karaj). *Waste Management*, 30(5), pp.912-920.

Omang, D.I., John, G.E., Inah, S.A. and Bisong, J.O. 2021. Public health implication of solid waste generated by households in Bekwarra Local Government area. *African Health Sciences*, 21(3), pp.1467-14763.

Rezaeisabzevar, Y., Bazargan, A. and Zohourian, B. 2020. Landfill site selection using multi criteria decision making: Influential factors for comparing locations. *Journal of Environmental Sciences*, 93, pp.170-184.

Schübeler, P., Christen, J. and Wehrle, K. 1996. *Conceptual Framework for Municipal Solid Waste Management in Low-income Countries* Vol. 9. St. Gallen: SKAT Swiss Center for Development Cooperation.

Siddiqui, M.Z., Everett, J.W. and Vieux, B.E. 1996. Landfill siting using geographic information systems: A demonstration. *Journal of Environmental Engineering*, 122(6), pp.515-523.

Tomlinson, R.F. 1990. *Current and Potential Uses of Geographic Information Systems: The North American Experience*. In: Peuquet, D. and Marble, D. (eds.), Introductory Readings in Geographic Information Systems, New York, Taylor and Francis.

Zadeh, M.A., Ngah, I., Shahabi, H. and Zadeh, E.A. 2013. Evaluating AHP and WLC methods in site selection of waste landfill (Case Study: Amol, North of Iran). *Journal of Basic and Applied Scientific Research*, 3(5), pp.83-88.

Evaluation and Assessment of Existing Design Codes and Standards for Building Construction: A State of the Art

Diman N. Abdulqader and Dawood S. Atrushi

Department of Civil Engineering, College of Engineering, University of Duhok,
Duhok, Kurdistan Region – F.R. Iraq

Abstract—**Building design codes (BDC) are used to control the construction industry in general and building design in particular. The BDC offers the construction sector with a standard language and set of requirements. There are several BDCs developed and utilized for construction purposes throughout the world. Certain design codes are employed in structural design to assure the structure's health and safety, as well as its cost-effectiveness. It also assures that the structure is sufficiently sturdy to endure all potential climatic conditions, bear its intended load, and is integrated to ensure effective use of building materials and resources. This research aims to compare various building construction design codes to identify and explore the most appropriate standard in terms of safe design, economics, and availability of details. In Kurdistan and different parts of Iraq, many international companies have designed building structures with various codes during the past 20 years. This is a bad condition since the government has no control over the construction of the buildings, which includes both the code and the building materials. There is currently no overview of the design codes in use in Kurdistan, nor is it clear whether they are congruent with what students' study in institutions.**

Index Terms—**Building design codes, Codal comparison, Comparative analysis, Specification and standards.**

I. Introduction

There is a scarcity of land globally due to the continued usage of landmass and the expanding population. The majority of the world's population lives in major cities. When building development grows, several criteria must be met, including the capacity to safely endure earthquake loads, counteract wind loads, and preserve structural stability (O'Bannon, 1973), in other words, the structure must be in balance throughout the forces operating on it (Koti, 2017). Several international and national standards are created and implemented in general building design to ensure life safety

and damage reduction. Many nations have developed seismic criteria and compared them to international counterparts to analyze and quantify the discrepancies (Nahhas, 2011; Marino, Nakashima and Mosalam, 2005; Fenwick, Lau and Davidson, 2002; Noor, Ansari and Seraj, 1997; McIntosh and Pezeshk, 1997).

Natural catastrophes such as earthquakes, landslides, tsunamis, and fires create significant damage by demolishing structures, interrupting transportation networks, and killing or trapping people and animals, among other things. Recent earthquakes in densely populated areas have revealed that existing structures built without seismic resistance are a major source of danger and the cause of the vast majority of casualties (Varum, 2003). Natural disasters like this pose a hazard to development (Ben-Joseph, 2012). Civil engineers, on the other hand, as designers, may be able to help mitigate damage by appropriately constructing structures or making other value judgments. This includes an understanding of the behavior of the construction and structural materials, of earthquakes, and the extent to which structural engineers utilize their knowledge to make appropriate decisions when constructing reinforced concrete structures (TQ and Given, 2017).

Reinforced concrete has earned a unique spot in the current construction of a variety of structures because of its numerous advantages as a composite material. It has mainly replaced prior materials such as stone, wood, and other natural elements due to its form flexibility and superior performance. It also plays an important role in structural structures such as multistorey frames, bridges, and foundations, among others. With the rapid growth of urban populations in both developed countries and emerging, reinforced concrete has become a popular option for residential construction (Bhavsar, et al., 2014).

Design strategies must merge in the era of globalization to produce structures with a constant risk of sustaining a specified level of damage or collapsing (Singh, Khose and Lang, 2012; Canisius, Baker and Diamantidis, 2011). It has been revealed that the design of high structures in seismically active areas differs significantly by region, with some countries requiring detailed performance-based studies and others requiring only a basic design based on force reduction factors. Moving on, each country has developed its own set of standards for creating safe structures according to

its own experiences with construction methods, nature, and materials since the dawn of time (Keeler and Vaidya, 2016; Australia, 2012). Buildings are constructed in accordance with the national design and seismic standards. All seismic code structural engineering designs must be based on ideal mechanics assumptions. These necessary codes must be experimentally validated, as well as rational, reasonable, and effective. These codes must be updated on a regular basis (Wagh, Narkhede and Salunke, 2018).

Structural design indicates to the precise calculation of type, size, materials, and the appropriate configuration until a comprehensive drawing is generated (Mosley, Hulse and Bungey, 2012; OBrien and Dixon, 1995) that can bear loads in a safe and serviceable manner (Oyenuga, 2011). All aspects of the structure, for example, the beam, slab, column, foundation, and roof, are designed (Nwofor, Sule and Eme, 2015). The introduction of new structural codes, design philosophies, and materials stimulates study into comparative analyses of structural design codes. Such studies give insight into the varied methods of codified structural design in various nations, indicating how much one code varies or agrees with another in terms of locally adjusted, safety, complexity, and details (Baltimore, 2009; Hitchin, 2008; Clemmensen, 2003). They are also valuable in nations where more than one code for structural design is permitted since they aid in deciding which code has a greater factor of safety than another (Tabsh, 2013; Bano, Izhar and Mumtaz,). However, most buildings across the globe have structural designs that are based on international and national norms of practice (Van der Heijden and De Jong, 2009; Liebing, 1987; Regulation, 2004). These direct the engineer's assessment of the general structural scheme, detailed analysis, and design (Allen and Iano, 2019; Billington, 1985). The codes of practice are essential tools created by knowledgeable engineers and related professionals that provide a framework for resolving concerns of structural safety and serviceability (Nwofor, Sule and Eme, 2015; Cheng, 2013; Franklina and Mensahb, 2011). Prescriptive and performance requirements are the two types of requirements that are commonly found in codes (Al-Fahad, 2012; Meijer and Visscher, 2008; Bartlett, Halverson and Shankle, 2003; Melkers and Willoughby, 1998). Differences, at times significant, might be found between the codes in the data provided for actions, in the requirements for assessing section resistance, as well as other code requirements for details, durability, and so on (Bakhoum, Mourad and Hassan, 2016).

A code is a legal document that governs a set of laws (Vaughan and Turner, 2013). A building code is a document that contains standardized rules that determine the minimum acceptable level of safety for both buildings as well as nonbuildings. Safety standards, as well as product standards, are two categories of codes (Dollet and Guéguen, 2022; Cote and Grant, 2008; Listokin and Hattis, 2005). These algorithms are based on engineers' experience, unique situations, behaviors, and experimental work. They safeguard the structures from numerous threats such as fire and structural collapse, as well as amenity concerns such as lighting, sanitation, moisture, sound insulation, and ventilation. Furthermore, codes are important

instruments for accomplishing societal aims, for example, sustainability and energy efficiency, as they cover all areas of building, for instance, the use of construction materials, seismic design, electrical, structural integrity, plumbing, and safe exits. Building codes categorize structures by applying and utilizing numerous criteria; for instance, schools and business buildings are classified as separate occupation groups with varying performance requirements (Standard, 1986).

Disputes on recent construction projects frequently include extremely technical problems, complex factual scenarios, and legal problems. The ancients were interested with how the environment can be affected back thousands of years. The modern civil justice system is based on biblical principles and reflects a four-thousand-year evolution of opinions and construction knowledge that began with the development of western civilization. The Code of Hammurabi covers the earliest known principles of construction law. Hammurabi was Babylon's sixth king, reigning from 1792 BC to 1750 BC. The Hammurabi Code contained 282 laws inscribed on twelve stone tablets that were displayed in public. One of the initial written legal codes in recorded history solves construction-related problems (Heady, Currie and Llp, 2012).

The history of every organization reflects the beliefs and actions of prearranged groups inside it. These operations were a modification to suit changing conditions, and they resulted in the formation of the American Concrete Institute (ACI) (ACI, 2002). ACI's history is intertwined with the evolution of concrete technology (Committee, 2008). The joint committee on reinforced concrete was formed in 1904 and generates many drafts of the concrete code (Institute, 2011). ACI 318 criteria became the only document in the United States dealing with concrete design in 1930 (Nowak and Szerszen, 2003). The American Concrete Institute (ACI) Committee 318 Building Code Criteria for Structural Concrete (ACI-318) (Nowak and Rakoczy, 2012) specifies the design requirements for special moment frames. ACI 318 is the major document for the design of concrete buildings in the United States, and it contains the unique criteria for inspection, structural construction concrete, and design materials (Poston and Dolan, 2008; ACI Committee, 2005). It also provides the resistance parameters, load factors, and design resistance. The code applies to both prestressed and precast concrete, as well as prestressing and reinforcing steel (Standard, 2011). Furthermore, regulations apply to diaphragms and frame components that are not specified as part of the seismic force-resisting system. The multiple interconnected criteria are covered in some parts of ACI 318 (Standard, 2011), which are not always presented in a logical arrangement, making application difficult for all except the most experienced designers. The code was accepted and made legal by the IBC. The ACI 318 code is revised every 3–4 years to reflect changes in the engineering sector (Moehle, Hooper and Lubke, 2016).

According to our present understanding and information, the great majority of structures in earthquake-prone locations in Europe built before the 1980s are seismically weak. Furthermore, before the implementation of contemporary seismic-oriented design philosophies, a substantial percentage

of extant reinforced concrete building structures were constructed before the 1970s using plain reinforcing bars (Rodrigues, et al., 2013; Rodrigues, Varum, and Costa, 2010). From 1971 through 1990, the code committee worked to acquire a draft of technical documents designated as an international investigation (Athanasopoulou, et al., 2018). For a limited number of years, this comprehensive set of coordinated Eurocodes (EC) for the geotechnical and structural design of buildings and civil engineering works was at first presented as Euronorme Voluntaire (ENV) standards, anticipated for use in conjunction with national application documents (NADs) as an alternative to national codes, for example, British code (BS8110 1997). Subsequently, Euronorme (EN) versions have approximately replaced these, with each member state of the European Union adopting a National Annex (NA) with domestically decided parameters to apply the ECs as a national standard (Reynolds, Steedman and Threlfall, 2007). Before the establishment of the Eurocodes, the British standard codes of practice were used to perform the same role as the Eurocodes, raising several problems about the disparities in building infrastructure (Nwofor, Sule and Eme, 2015). The Eurocodes are a brand-new collection of European structural design codes for buildings and all civil engineering projects. The Eurocodes were created as part of a broader European coordination process, rather than to simply replace any national codes (Shodolapo and Kenneth, 2011; Marpal, 2010; Liew, 2009; Jawad, 2006; Bond, et al., 2006; Moss and Webster, 2004; Institution, 2004).

The Eurocode framework, which includes 58 standard documents, is depicted in Fig. 1. The European code is based on the concept of limit states and contains standards for both serviceability and strength. Resistance and load factors design are the terms used to represent them. Also specified in the code are material strength reduction factors (Al-Taie, Al-Ansari and Knutsson, 2014). The relevant standards in the design of concrete structures are EC0: Basis of structural design (Gulvanessian, Calgaro and Holicky, 2012; Standard, 2002; Gulvanessian, 2001), EC1: Actions on structures (Gulvanessian, Formichi and Calgaro, 2009; Standard, 2006; Eurocode, 2006; Gulvanessian and Holický, 1996), and EC2: Design of concrete structures (Walraven, 2008; Code, 2005; Beeby and Narayanan, 2005; Standard, 2004). The goals of these ECs are to establish universal design standards and methodologies for meeting the stipulated requirements for stability, fire resistance, and mechanical resistance, regulating the construction industry as well as features of durability and economy. Furthermore, they facilitate a shared knowledge of structural design among users, owners, and operators, as well as contractors, designers, and producers of construction materials. Furthermore, they allow the trade of construction services across European Union member states and act as a single platform for construction-related research and development. Furthermore, they promote the operation of the single market by reducing impediments caused by nationally defined procedures. Furthermore, enhance the European building industry's competitiveness (Wimo, et al., 2011). The ACI and BS limit state concepts (Ultimate Design Method) are also used by EC2 (Jawad, 2006). EC2, or more specifically BS EN 1992, governs the design of structures and civil engineering works made of precast, prestressed, reinforced, and plain concrete was released in 2004 (Bisch, et al., 2012; Marpal, 2010; En, 2004). Since April 2010, EC2 has become the standard code for the design of reinforced concrete structures in the United Kingdom, and the previous BS8110-1997 (Higgins and Rogers, 1998) has been phased out (Mosley, Hulse and Bungey, 2012; Reichert, 2005).

Most Iraqi civil engineers, particularly those in the Kurdistan area, are aware of the ACI code; although, it is vital to educate them on the other recent British and European codes. Before Eurocode 2 and BS 8110 become heavily engaged in our design lives, most engineers will need to be confident that they can be used as a practical design tool. Knowledge must be expanded to include all facets of each component, as well
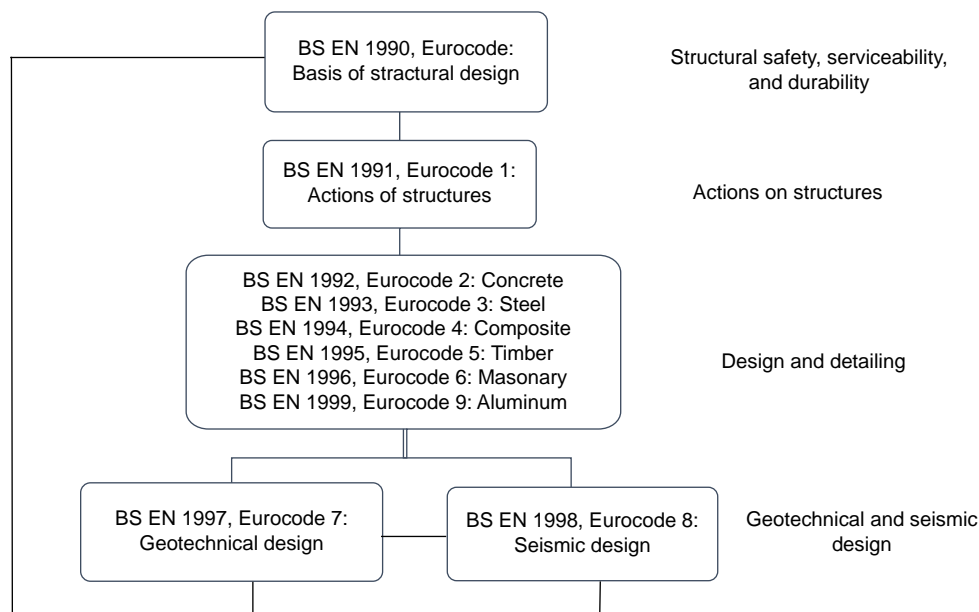

Fig. 1. Eurocodes structure, (Produced from Al-Taie, Al-Ansari and Knutsson, 2014).

as the economic and conservative outcomes. A conservative design is that design in which the designer focuses more on the loads and other predictable problems, and so makes more preventative decisions, basically to add a higher design factor of safety. Because the majority of the nations mentioned have approved and followed European Standards in the design and analysis of their structures. The objective of this paper is to explore and survey the feasibility of the existing different design codes which are utilized in the construction sector for the building purpose. By collecting data and making a review and comparison between some published works. Since the building codes affect the structure properties, materials selection, the area, natural, and economic aspects.

II. Statement of Problem

Most countries require the structural design to adhere to a single code or guideline. Reinforced concrete design, for example, is carried out by the ACI 318 code (ACI318M-11 2011) (Standard, 2011) in the United States, the CAN/CSA CSA-A23.3 (2004) (Association, 2004) standard in Canada, the Eurocode-EC2 (2002) (Beeby, Narayanan and Narayanan, 1995) in Europe, and AS3600 (2009) (Gilbert, Mickleborough, and Ranzi, 2016) in Australia. Some countries around the world, however, accept structural design based on one of several codes. The Kurdistan Region of Iraq (KRI) is one of these regions, and it allows the reinforced concrete design of building structures to adhere to either the ACI 318, the BS 8110 code (1997), or the Turkish code. Although the British Standard is being phased out in the UK owing to the introduction of Eurocode EC2, this is not the situation in the KRI, where BS 8110 is still widely utilized by a large number of consultants. The importance of the study is that for the past 20 years, many international companies have designed building structures in Kurdistan and different parts of Iraq with various codes. This is an unfortunate situation because the government has no control over the building structures, taking into account both the code and the building materials. At present, there is no overview concerning the design codes used in Kurdistan, and whether it is consistent with what students learn at the universities. Therefore, the uniqueness of this study is that this inadequate condition will be clarified by identifying the most appropriate construction design code for the region that has not been investigated yet. For reinforced concrete design, it is expected that both ACI 318 and EC2 are about equally followed. Because the design requirements of the ACI 318 and EC2 codes differ, it is necessary to analyze the structural demand in the two codes and decide which one has a greater factor of safety for a particular limit condition. Furthermore, the construction community has limited knowledge of and interest in the newly developed Eurocodes. It will take effort and time to learn how to use the new Eurocode 2. As a result, applying programming approaches to new design aspects will aid designers in the transition to new code adaptation. However, it is not obvious which design code fits the most; consequently, the research will focus on this problem, by studying this gap in design codes in the Kurdistan region.

III. Methodology

In this paper, a comprehensive literature review was done for the findings of the 43 studies of previous research relating to the comparison of building construction design codes. The researchers are taken from different websites and journals, for example, World Conference, Journal of Structural Engineering, Nigerian Journal of Technology, Journal of Engineering, Asian Journal of Civil Engineering and Applied Sciences, Structures and Materials Journal, IJCIET, Journal of Engineering and Development, IJERT, International Journal of Engineering Technology and Sciences and more. Taken from databases such as Google Scholar, Elsevier, MDPI, IEEE explore, Science Direct, Semantic scholars, Academia, and more. That their duration varied from 1992 until 2021.

There are various building construction design codes to explore and survey. Since some countries build and design their standards whereas others use international codes such as ACI, UBC, IBC, and EC. But this study aims to find out the most appropriate design code for the Kurdistan Region by reviewing and evaluating the various existing design codes as shown in Table I, and probably the need for a unified design code in Kurdistan. Furthermore, in the next section which is the result and discussion, there will be a comparison between these standards and their effect on the analysis and design process.

As shown in Table I, the review covers various design codes such as North American design codes for instance ACI (Committee, 2008), UBC (Conrad and Winkel, 1998), IBC (Code, 1997), ASCE (Vesilind, 1995), AISC (Muir and Duncan, 2011), NBC (Canada, 2015; Code, 1990), and CSA (Association, 2005), South American codes, for example, NBR (NBR, 2006), NSR (Galíndez and Thomson, 2007), CEC (Canchig Cola, 2016), and NCh (Guendelman, Saragoni and Verdugo, 2012), Asian codes such as IS (Haldar and Singh, 2009), BSLJ (Itabashi and Fukuda, 1999), NZS (Authority, 1992), BNBC (Islam, et al., 2011), NBC India (de León, 2010), NSCP (Deepshikha and Basu, 2011), ICS (Habibi and Asadi, 2013), TEC (Şengöz, 2007), OSC (Al-Sayed and Waris, 2017), ISC (Amer, Sobaih and Adel, 2016), and SBC (Shuraim, et al., 2007), European codes like EC (Gulvanessian, Calgaro and Holicky, 2012), BS (Arya, 2018), African codes such as ECP (Committee, 2007), RPA (Belazougui, 2017), and Australian code AS (Menegon, et al., 2018). The proportion of the applying building design codes of the researchers that have been studied in their researches is illustrated in Fig. 2.

Moreover, a comparison based on some parameters will be prepared, such as the details, economic aspect, base shear which is an estimate of the maximum predicted lateral stress on the structure's base as a result of seismic activity, as it is computed by using the soil material, building code lateral force, and seismic zone formulas (Davidson, 2008; Balendra, et al., 2002; Mehrabian, 1996). The displacement that is known as the lateral displacement of the story relative to the base is indicate to as story displacement. The lateral force-resisting system can reduce the building's excessive lateral displacement. Typically, the story drift ratio around the

TABLE I
A Summary of Some Findings of Previous Research Relating to the Comparison of Building Construction Design Code

S. No.	Reference	Source	Design codes	Origin	Typical model	Software used	Parameters studied	Findings of the study
1	Bose, Dubey and Yazdi, 1992	Earthquake Engineering Tenth World Conference	BSLJ, IS, NBC, NZS, UBC	Japan, India, Canada, New Zealand, United States	10-storey framed buildings	-	Details	The ratio for structural behavior coefficient of ductile performance is greatest in IS and lowest in NZS. The NZS does not offer the necessary data about the seismic hazard, hence the distribution of base shear for UBC, NBC, and NZS is more accurate.
2	Shkoukani, 1993	The First Palestinian Engineering Conference	ACI, EC	Palestine	RC rectangular beam	-	Economic	The quantity of shear reinforcement estimated by the ACI Code is less than the EC calculates. The elastic analysis moment and partial safety factors used in the EC code for loads and materials were less than those used in the ACI code.
3	Atique and Wadud, 2001	The Eighth East Asia-Pacific Conference on Structural Engineering and Construction	BNBC, UBC, NBC India	Bangladesh	A series of multistorey concrete buildings 10, 15, 20, and 25 stories	-	Base shear, Conservative	The BNBC provides the lowest value for base shears for both types of structures. BNBC and NBC-India are the least conservative, with nearly identical base shear. UBC is more conservative than both these codes. UBC is 2.61 times more conservative than the BNBC in general.
4	Fenwick, et al., 2002	Bulletin of the New Zealand Society for Earthquake Engineering	NZS, Draft NZ/Australian Loadings Standard, UBC, IBC, EC	New Zealand	A series of multistorey concrete frame buildings of 6, 12, 18, and 24 stories	SAP2000	Base shear	The required strength levels determined by the New Zealand Loadings Standard and the draft NZ/Australian Standard were significantly lower than those determined by the UBC and IBC codes of practice, and significantly lower than the comparable values in EC, but EC has the least amount of deflection. EC requires over 4.0 times and the UBC and IBC over 1.9 times, the base shear required by the NZS.
5	Moss and Webster, 2004	Journal of Structural Engineering	BS, EC	United Kingdom	-	-	Details	EC was less prescriptive and had a wider scope than BS. As a result, the use of EC would allow designs that would not ordinarily be permitted in the United Kingdom and would allow designers to profit from the significant improvements in concrete technology that had occurred.
6	Faizian and Ishiyama, 2004	World Conference on Earthquake Engineering	BSLJ, IBC, ICS	Japan, United States, Iran	-	-	Details	Although the ICS is quite close to the IBC, the BSLJ differs significantly from the other two codes. The significance of a building is addressed in the ICS and IBC but not in the BSLJ.
7	Jawad, 2006	Journal of Engineering and Development	ACI, BS, EC	Iraq	Rectangular beam section	-	Details, Economic	EC is more liberal in terms of strength design and partial safety factors than the ACI Code. In terms of design methodology, EC and BS are not very different from ACI code, but the EC and ACI Codes are more comprehensive than the BS. The ACI Code outcomes diverge on the less economical side.
8	Doğangün and Livaoğlu, 2006	Journal of Seismology	TEC, IBC, UBC, EC	Turkey	6 and 12 storey RC buildings	SAP2000	Details, Displacement, Base shear	The EC's recommendations are equally critical for the Turkish community. Even though all domains of the response spectrum are defined differently in the EC, they are not defined differently in the UBC or TEC. The maximum lateral displacement values for the buildings are given by EC, whereas the lowest values are given by UBC. For equivalent ground types listed in other codes, EC often delivers the larger base shear.

(Contd...)

TABLE I
(Continued)

S. No.	Reference	Source	Design codes	Origin	Typical Model	Software used	Parameters studied	Findings of the study
9	Ng, Loo and Bong, 2006	6th Asia-Pacific Structural Engineering and Construction Conference (APSEC)	BS, EC	Malaysia	RC rectangular beam	-	Conservative	Both standards provide nearly identical flexural strength for reinforced beams. The requirement for ductility in beams is stricter in EC, where the maximum allowable neutral axis depth is lower than in BS. In general, BS demonstrated greater flexural strength.
10	Parisi, 2008	Practice Periodical on Structural Design and Construction	ASCE, EC	Italy	-	-	Details	By demonstrating the design philosophy, the EC aims at understanding ideas and holding up resolutions, whereas the ASCE code appears more user affable and exercise-oriented, executing the design process with comprehensive information.
11	Hawileh, Malhas and Rahman, 2009	Structural engineering and mechanics journal	ACI, EC	UAE	-	-	Conservative, Economic	The EC flexural requirements are a little more conservative, with only a slight practical variation from the ACI standards. The EC provisions have a higher safety factor than the ACI.
12	Mourad and Hassan, 2009	13th International Conference on Structural and Geotechnical Engineering (ICSGE)	ECP, EC, UBC	Egypt	Rectangular beam sections	-	Details	ECP has a lower maximum ground acceleration than UBC and EC. This might be related to the fact that EC and UBC cover an extensive variety of seismic zones that vary along a large region, but ECP lacks seismic detail. Furthermore, it may be concluded that EC limitations are stricter than UBC limits.
13	Marpal, 2010	International Journal of Engineering Technology and Sciences	BS, EC	Malaysia	RC slabs	Microsoft Excel Spreadsheet	Details	BS is usually used to influence the design of constructions in Malaysia. It has long been assumed that the design process will not change as a result of the use of EC. Also, demonstrate that this program serves the research aims of developing software to assist designers in the use of BS and EC.
14	Franklina and Mensahb, 2011	Journal of Basic and Applied Scientific Research	BS, EC	South Africa	4-storey RC building	PROKON 32	Economic	There was a virtually slight difference between BS and EC in terms of the difficulty of the calculations required or the conclusions achieved. In a continuous beam, the maximum span moments, the EC moments, are less than the BS values.
15	Imashi and Massumi, 2011	Asian Journal of Civil Engineering	ICS, IBC	Iran	12-storey building	-	Details	The findings demonstrate the importance of reviewing the Iranian seismic code and developing more suitable relations to achieve economic and functional goals. In the Iranian seismic code, shear force values are higher than in IBC.
16	Adewuyi and Franklin, 2011	Journal of Engineering and Applied Sciences	BS, EC	Nigeria	Simply supported beam	PROKON 32	Economic	The results demonstrate that the EC moments are lower than the BS values. BS estimations are usually greater in the shear force envelopes.
17	Singh, Khose and Lang, 2012	15th World Conference on Earthquake Engineering	ASCE, EC, NZS, IS	India	8-storey RC frame building	SAP2000	Base shear	The codes also differ greatly in terms of minimum design base shear. The design base shear required by EC is similar to that required by NZS but IS outcomes in the lowest design base shear for a given hazard.

(Contd...)

TABLE I
(Continued)

S. No.	Reference	Source	Design codes	Origin	Typical Model	Software used	Parameters studied	Findings of the study
18	Santos, Lima and Arai, 2012	Structures and Materials Journal IBRACON	EC, ASCE, South American codes (NSR, CEC, NCh, NBR)	United States	10-storey RC building	SAP2000	Displacement	In the EC consideration, the displacements achieved with this code are much lower than those produced with the other codes.
19	Xiaoguang, et al., 2012	12th World Conference on Earthquake Engineering, Lisbon.	Korea, Japan, China, Nepal, India, Indonesia, Iran, Turkey code	Korea, Japan, China, Nepal, India, Indonesia, Iran, Turkey	4-storey RC building	-	Details	The degree of seismic fortification in China and Japan is high, whereas it is lower in Turkey and Korea. Except for Korea, the seismic design code of construction has reflected the influence of site conditions. Because there is no design seismic response spectrum material in Korea's building seismic design code.
20	Landingin, et al., 2013	The Open Construction and Building Technology Journal	NSCP, EC, IBC	Portugal	4-storey RC building	SAP2000	Safe, Conservative, Details	The EC provisions were seen to be safer. According to the findings, EC was shown to be more conservative than NSCP and IBC. The EC examined the consequences of seismic activities in both directions, whereas the NSCP and IBC standards do not.
21	Tabsh, 2013	Structural Engineering and Mechanics journal	ACI, BS	UAE	Beams, slender, and columns cross-sections	-	Economic	The ACI code outcomes in larger cross-sections and higher reinforcement ratios. ACI code has a lower shear strength than BS code.
22	Alnuaimi, Patel and Al-Mohsin, 2013	Practice periodical on structural design and construction	ACI, BS	Oman	Rectangular beam sections	-	Economic	For the same design load, the BS code needs less reinforcement than the ACI code. The ACI code requires a higher minimum area of flexural reinforcement than the BS code. The minimum area of shear reinforcement needed by the ACI code is less than that required by the BS code.
23	Bhavsar, et al., 2014	International Journal of Scientific and Research Publications	IS, EC	India	8-storey RC building	ETABS	Economic	The infrastructures of Gulf nations are typically impressive since they generally follow EURO standards for building development. In EC, the area of reinforcement required in the column is less than in IS. This is due to the greater modulus of elasticity in EC. As a result, in EC, column ductility is regulated by the modulus of elasticity but IS is controlled by the area of reinforcement.
24	Itti, Pathade and Karadi, 2014	Structural Engineering Forum of India (SEFI)	IS, IBC	India	10-storey RC building	ETABS	Base shear, Displacement	IS Code buildings have a greater base shear and higher displacements than the IBC design code. The base shear of IS buildings is higher than that of IBC buildings by 26.10%. Displacements of IS are higher than that of IBC by 11.00%.
25	Chebihi and Laouami, 2014	2nd European Conference on Earthquake Engineering and Seismology	RPA, UBC, EC	Algeria	10-storey RC building	SAP2000	Base shear, Displacement	RPA results are close to those of UBC, but EC provides smaller base shears and displacements than RPA and UBC.
26	Al-Taie, Al-Ansari and Knutsson, 2014	Engineering Journal	ACI, EC	Iraq	Foundation element	STAAD Pro V8i, SAFE	Details	Eurocode allows the user additional freedom to adopt their standards. EC is preferable since it contains all design and construction standards for all types of buildings, as well as modern and traditional materials. It also includes national annex national defined parameters NDPs. As a result

(Contd...)

TABLE I
(Continued)

S. No.	Reference	Source	Design codes	Origin	Typical Model	Software used	Parameters studied	Findings of the study
								of these factors, the Eurocode is the best international code. In the design, EC employs three ways. Materials characteristics, action loads, and soil resistance are examples of these. In addition, the code incorporates the British standards and code that are well-known in several nations throughout the world.
27	Nandi and Guha, 2014	International Journal of Engineering Research & Technology (IJERT)	IS, BS, EC	India	Slab, beam, column, and foundation element	-	Economic	The IS code allows for a larger slab area of steel than the BS and EC codes. The maximum steel beam area permitted by the EC code is greater than that permitted by the IS and BS codes. The maximum area of steel for a column is specified by the BS code rather than the IS and EC codes. Minimum steel is required for foundations following the EC regulation.
28	Nwofor, Sule and Eme, 2015	International Journal of Civil Engineering and Technology (IJCIET)	BS, EC	Nigeria	RC beam cross-section	Microsoft Excel Spreadsheet	Economic, Conservative	The EC requires less reinforcement at the spans and supports than the BS, showing that the BS requires more shear reinforcement. In comparison to EC, the BS code applies larger partial safety factors to loads at the ultimate limit state, whereas the EC is more conservative regarding partial safety factors for loadings. EC provides a more economical design with the necessary margin of safety.
29	Karthiga, et al., 2015	International Journal of Research in Engineering and Technology (IJRET)	IS, EC, ASCE, BS	India	11-storey building	STAAD Pro V8i, SAFE	Economic, Displacement	The EC standards were the most economical, whereas the Indian norms were the least economical. IS had the highest shear value during the pushover analysis. The minimum displacement as it is derived from the displacement data. In comparison to the IS, the percentage rise for EC is 22%, the ASCE percentage increase is 20%, and the BS percentage increase is 19%. As a result, buildings constructed following the IS are more rigid, attracting larger seismic forces.
30	Bakhoum, Mourad and Hassan, 2016	Journal of Advanced Research	ECP, ASCE, ACI, AISC, EC	Egypt	Beams and columns considering steel, concrete, and composite materials	-	Economic	In comparison to the EC, ACI defines lower values for the ultimate dead load factor. ECP and AISC require larger sections than EC, which is more economic by 2–10% depending upon the resistance of steel and the reinforcement ratio. In comparison to ACI and ECP, the findings reveal that for the same section dimensions, EC produces the highest axial strength.
31	Koti, 2017	International Journal for Innovative Research in Science & Technology (IJIRST)	IS, ASCE, EC	India	25-storey RC building	SAP2000	Economic	EC has more reinforcement area than other codes, therefore it is clear that the structure analyzed for IS is more rigid, whereas the structure analyzed for EC is more ductile. The drift ratio exceeds the intended limitations in the case of an IS design building due to gross section stiffness in design.
32	Nahhas, 2017	Open Journal of Earthquake Research	UBC, SBC	Saudi Arabia, USA	10-storey building	ETABS	Base shear	It is discovered that SBC base shear is higher in most sites, and the same is true for overturning moments. SBC seismic maps may be inaccurate.

(Contd...)

TABLE I
(Continued)

S. No.	Reference	Source	Design codes	Origin	Typical Model	Software used	Parameters studied	Findings of the study
33	Waris, Al-Jabri and EL-Hussain, 2017	World Conference on Earthquake (16WCEE)	OSC, UBC, IBC	Oman	Three buildings of 4, 10, and 14, storey	ETABS	Base shear, Conservative, Economic	UBC base shear values were 7.6, 4.1, and 3.2 times higher and IBC offered base shear values that were 6.3, 3.3, and 2.6 times higher than those from OSC. As compared to OSC, both UBC and IBC give relatively conservative seismic loads on structures. OSC enhances the economics of earthquake design.
34	Nwoji and Ugwu, 2017	Nigerian Journal of Technology (NIJOTECH)	BS, EC	Nigeria	2-storey building	CSC Tedds	Economic, Safe	The EC values are lower than the BS values in terms of column load and moments. The EC is more flexible, safer, and easier to use than the BS, and it will give more economical sections and is technologically more advanced.
35	Taha and Hasan, 2018	Eurasian Journal of Science and Engineering (EAJSE)	ISC 2014, ISC 1997	Kurdistan Region/Iraq	RC building consists of 5, and 15 storey	ETABS	Base shear	The ISC 2014 requirements result in a significant rise in base shear forces. As observed, ISC 1997 simply analyzes the important class of the building, but ISC 2014 bases the selection on the building irregularity and height, which is more sensible, because the dynamic behavior of the structure is connected to these two elements.
36	Wagh, Narkhede and Salunke, 2018	International Journal of Science Technology and Engineering (IJSTE)	IS, EC, NZS	India	25-storey RC frame building	ETABS	Base shear, Displacement	When compared to the EC and NZS Codes, the IS has the lowest base shear, as calculated according to EC is higher than IS by 79% whereas according to the NZS is higher than IS by 44%. The story displacements and drifts for EC are the lowest as compared to IS and NZS. The IS provides no modeling rules, leaving it up to the capabilities of individual designers.
37	Donduren and Omeed, 2018	Journal of International Environmental Application and Science	TEC, ACI, EC	Turkey	Masonry building properties	-	Details	The Turkish Code's material selection is similar to those of other standards. As observed in the Turkish code, the requirements are based on the seismic zone of the building according to its seismic location. The ACI and EC, on the other hand, base their masonry construction guidelines on the materials available for usage in the structure.
38	Gadade, et al., 2018	Resincap Journal of Science and Engineering	BNBC, ICS, EC, IS, NSCP, IBC	India	-	-	Conservative, Base shear	In comparison to other standards and procedures, BNBC is the least conservative. The Iranian seismic code is quite similar to the American code, whereas the Japanese code differs significantly from the other two codes. It was discovered that EC is more conservative than NSCP and IBC. IS Code buildings have a greater base shear than IBC Code.
39	Izhar and Dagar, 2018	International Journal of Civil Engineering and Technology (IJCIET)	IS, BS, EC, ACI, CSA	India	11-storey building	STAAD Pro V8i	Economic	Shear and flexural reinforcement are the least from the IS code and the most from the CSA. The longitudinal reinforcement for columns is the minimum from EC and the most from CSA. The longitudinal and transverse reinforcing of the slab is the least for EC and the most for ACI.

(Contd...)

TABLE I
(Continued)

S. No.	Reference	Source	Design codes	Origin	Typical Model	Software used	Parameters studied	Findings of the study
40	Al-Obaidi, Jokhio and Abu-Tair, 2019	IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)	American (ASCE), Australian codes (AS)	UAE	Two buildings 70-storey RC building	ETABS	Economic, Conservative, Safe, Base shear, Details	The Australian code is considerably more economical than the American code, but because the American code is more conservative, it is more commonly used internationally among engineers. ASCE in the base shear results, which demonstrates more than AS by 15%. The ASCE is more detailed than the AS, whereas the AS is more generic, with just one time-history equation, making the Australian code safer because the structure is subject to more earthquake time.
41	Pacheco, et al., 2019	Journal of Structural Engineering	EC, ACI	Portugal	RC beams	-	Conservative, Safe	There were no major variations in the bias of the ultimate moment between the two comparison vectors of ACI versus EC requirements or recycled versus natural coarse aggregate. ACI and EC produce similar bias.
42	Rajeev, Meena and Pa(llav, 2019	Journal of Infrastructures	IS, BS, EC	India	4-storey RC framed building	SAP2000	Displacement, Economic	IS is a more secure design principle with more reserve strength. When compared to the BS and EC codes, the Indian code has a higher displacement capacity, the IS provides a 19% and 26% increase in displacement capacity compared to the BS, and EC, correspondingly. The IS appears to give more steel than the BS and EC. The quantity of steel required to comply with the IS is 40.6% and 35.1% more than the BS, and EC, correspondingly.
43	Arezoumandi, et al., 2021	Journal of Structural and Construction Engineering (JSCE)	ACI, EC, AS, BSLJ	Japan	-	-	Conservative	According to the findings of this study, ACI, EC, AS, and JSCE regulations are conservative for shear strength values of 88, 78, 95, and 100%, respectively. This demonstrates that the Japanese code relations are conservative for all data.

ACI: American Concrete Institute, EC: Euro Code, BS: British Standards, IS: Indian Standards, UBC: Uniform Building Code, IBC: International Building Code, ASCE: American Society of Civil Engineers, BSLJ: Building Standard Law of Japan, NBC: National Building Code of Canada, CSA: Canadian Standards Association, AS: Australian Standards, NZS: New Zealand Standard, BNBC: Bangladesh National Building Code, NBC India: National Building Code of India, NSCP: National Structural Code of the Philippines, ICS: Iranian Seismic code, TEC: Turkish Earthquake Code, OSC: Oman Seismic Code, ISC: Iraq Seismic Code, SBC: Saudi Building Code, ECP: Egyptian Code of Practice, RPA: Algerian Seismic Code, NBR: Brazilian National Standards, NSR: Colombian Standard, CEC: Ecuadorian Standard, NCh: Chilean Standard, AISC: American Institute of Steel Construction

building's middle level is more essential than that at the top (Lian and Wang, 2006; Pekau, Zielinski and Lin, 1995). The final parameter to be compared is the conservative parameter. Accordingly, Fig. 3 depicts the percentage of utilizing of these parameters in different research.

The studies are taken from several destinations and countries throughout the world for taking different aspects and circumstances as much as possible; their origin is illuminated in Fig. 4.

IV. Results and Discussion

For the construction, design, and operations of buildings, common language and requirements are prepared in the form of codes and standards. These documents provide the necessary guidelines for designing and constructing buildings that are safe, secure, healthy, energy-efficient, and accessible.

There are several building design codes that have been produced and are in use around the world. Some conclusions can be drawn by comparing different design codes.

A. In Terms of Details

There are several outputs as a result of the obtained data. According to some research, certain existing design regulations and previously utilized standards lacked information. Bose, Dubey and Yazdi, 1992 concluded that the NZS did not give the required information regarding seismic hazards, since there were insufficient specifics in terms of seismic design and seismic zones. Mourad and Hassan, 2009 findings reveal that EC and UBC cover a wide range of seismic zones across a massive territory, whereas ECP lacks seismic detail. Furthermore, according to the conclusion of (Xiaoguang, et al., 2012), there is no design seismic response spectrum material in Korea's building seismic design code.
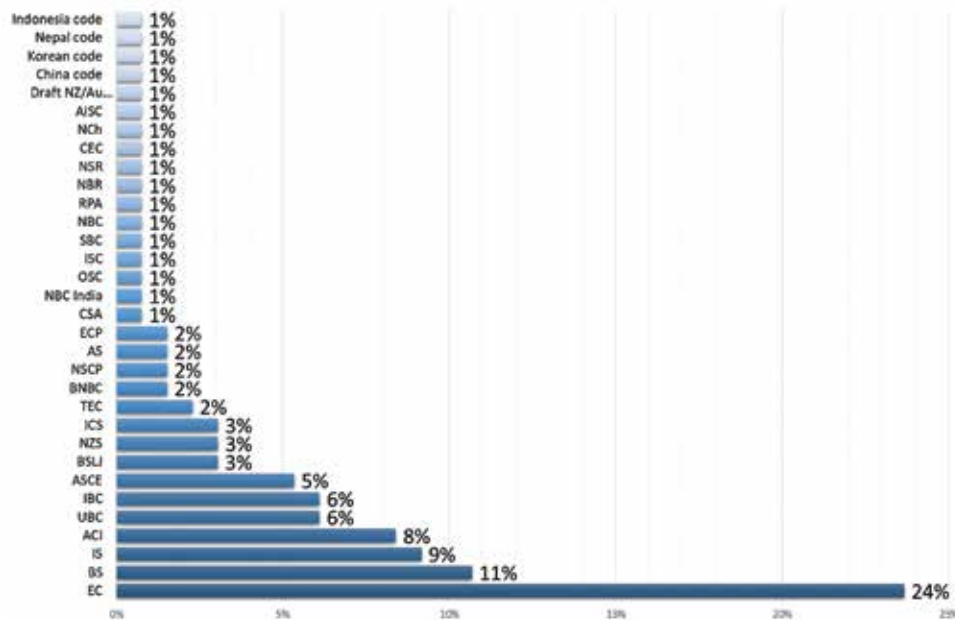
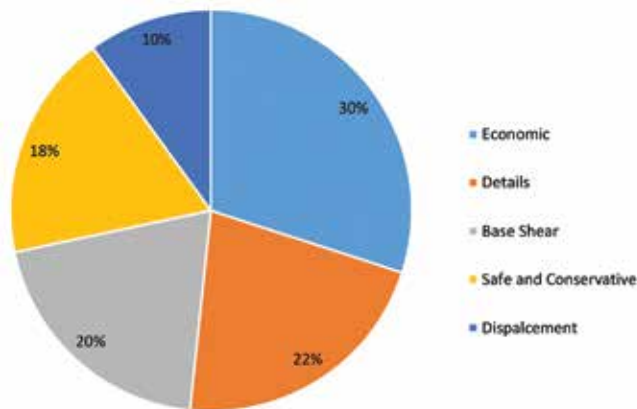Fig. 2. The ratio of applying building design codes.

Fig. 3. The ratio of studied parameters.

Unlike (Landingin, et al., 2013) said, the EC looked at the effects of seismic events in both directions, but the NSCP and IBC regulations do not. Even though all domains of the response spectrum are defined differently in the EC, as outlined by (Doğangün and Livaoğlu, 2006) they are not defined differently in the UBC or TEC. Al-Taie, Al-Ansari and Knutsson, 2014 also argue that EC is preferable to ACI since it contains all design and construction standards for all types of buildings, as well as modern and traditional materials, and in design, EC employs three methods, including materials characteristics, action loads, and soil resistance. Donduren and Omeed, 2018 show that the EC and ACI codes' guidelines on the specifics of the materials accessible are more than TEC.

Furthermore, several studies have found that design codes, such as BS, which (Moss and Webster, 2004) presented, have a restricted reach, because EC was less restrictive and had a broader reach than BS, it allowed for designs that would not normally be authorized in the United Kingdom and allowed designers to benefit from considerable advancements in concrete technology. According to the findings of (Marpal,

2010) is frequently utilized to influence building design. In Malaysia, it is thought that BS is frequently utilized to influence construction design. It has long been considered that using EC will have little effect on the design process. Furthermore, show how this application contributes to the research goals of building tools to help designers use BS and EC. The EC and ACI codes, according to (Jawad, 2006) are more thorough than the BS. The significance of a building is addressed in the IBC and ICS, but not in the BSLJ, according to (Faizian and Ishiyama, 2004). Imashi and Massumi, 2011 findings highlight the significance of revising the Iranian seismic code and building more appropriate relationships in order to meet economic and functional objectives. And, as (Parisi, 2008) clarifies that the EC aims at understanding ideas and holding up resolutions, whereas the ASCE code appears more user affable and exercise-oriented, executing the design process with detailed information. Last of all, according to (Al-Obaidi, Jokhio and Abu-Tair, 2019) the ASCE is more detailed than the AS, whereas the AS is more general, with just one time-history equation, making the Australian code safer because the structure is exposed to greater earthquake time. Fig. 5 depicts the fraction of research lacking in details for all of these studies.

B. Economic Aspect

In terms of economics, the data demonstrate that building construction design is becoming more cost-effective. According to (Shkoukani, 1993), ACI is more economic than EC since the ACI Code estimates less shear reinforcement than the EC does. The EC codes elastic analysis moment and partial safety factors for materials and loads were lower than the ACI code. According to (Al-Obaidi, Jokhio and Abu-Tair, 2019), AS is more cost-effective than ASCE. Furthermore, according to (Koti) IS code is more economic than EC and has a larger reinforcement area than other codes. As well
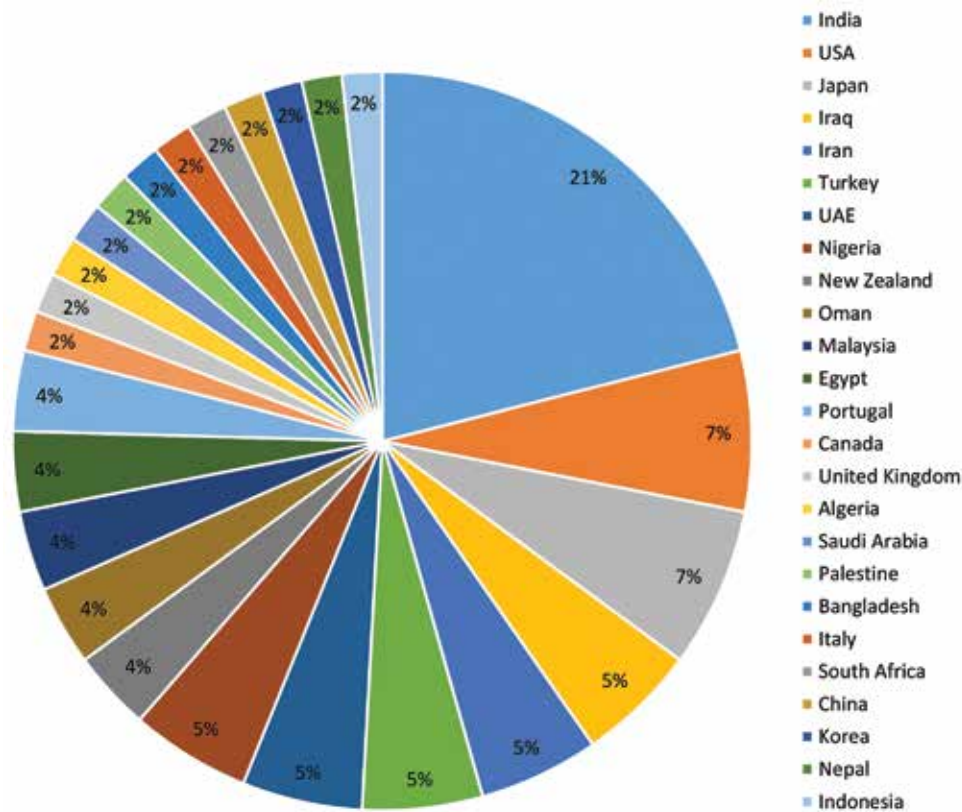
Fig. 4. The ratio of various origin.
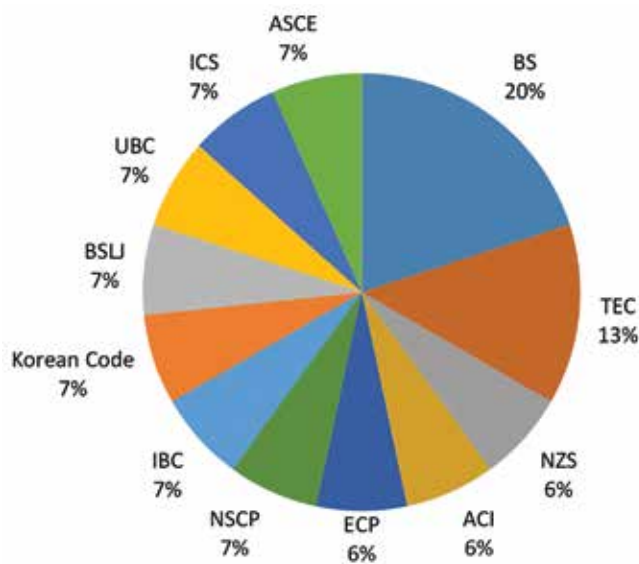
Fig. 5. The ratio of lack in details parameter.

the BS code outcomes in smaller cross-sections and lower reinforcement ratios than the ACI code as represented by (Tabsh, 2013; Alnuaimi, Patel and Al-Mohsin, 2013).

Whereas the majority of research found that EC is the most cost-effective design code. As stated by (Rajeev, Meena and Pallav, 2019; Karthiga, et al., 2015; Nandi and Guha, 2014; Bhavsar, et al., 2014). According to (Bhavsar, et al., 2014), the area of reinforcement required in the column in EC is smaller than in IS. This is owing to the fact that EC has a higher modulus of elasticity in EC. The IS code permits for a

bigger slab area of steel than the BS and EC codes, whereas the EC rule requires minimum steel for foundations, as found by (Karthiga, et al., 2015; Nandi and Guha, 2014). In addition, (Rajeev, Meena and Pallav, 2019), show that the IS appears to provide more steel than the BS and EC, with the IS requiring 40.6% and 35.1% more steel than the BS and EC, respectively.

Furthermore, the results show that Eurocode is better and preferable than ACI by (Waris, Al-Jabri and EL-Hussain, 2017; Bakhoum, Mourad and Hassan, 2016; Hawileh, Malhas and Rahman, 2009; Jawad, 2006; Izhar and Dagar). Jawad, 2006 presented that the ACI Code outcomes diverge on the less economical side. According to, the EC regulations have a greater safety factor than the ACI. Bakhoum, Mourad and Hassan, 2016 illuminate that, as compared to the EC, the ECP, ACI, and AISC codes demand bigger sections than EC, which is more economic by 2–10% depending on the steel resistance and reinforcement ratio. The OSC code is based on EC, according to (Waris, Al-Jabri and EL-Hussain, 2017). As a result of both UBC and IBC, OSC improves the economics of earthquake design. Furthermore, (Izhar and Dagar, 2018) demonstrated that EC provides the least longitudinal reinforcement for columns, whereas CSA provides the highest. EC has the least longitudinal and transverse slab reinforcement, whereas ACI has the greatest. Furthermore, EC is more economic than BS as concluded by (Nwoji and Ugwu, 2017; Nwofor, Sule and Eme, 2015; Franklina and Mensahb, 2011; Adewuyi and Franklin, 2011). The EC moments are lower than the BS values as the results demonstrated by (Franklina and Mensahb, 2011; Adewuyi and Franklin, 2011). Nwofor, Sule and Eme, 2015

explain that the EC requires less reinforcement at the spans and supports than the BS, showing that the BS requires more shear reinforcement. EC provides a more economical design with the appropriate safety margin. Nwoji and Ugwu, 2017 show that in terms of column load and moments, the EC values are lower than the BS values, resulting in more economical sections. Fig. 6 depicts a summary of the most cost-effective design code.

C. In Terms of Base Shear

In terms of base shear, however, the design codes comparison yielded different findings. As a result, (Wagh, Narkhede and Salunke, 2018) indicates that the codes differ substantially in terms of minimum design base shear when compared to the EC and NZS Codes, with the IS having the lowest base shear, as determined by the EC is greater than IS by 79%, and the NZS is higher than IS by 44%. Similarly, (Singh, Khose and Lang, 2012) findings show that the codes differ significantly in terms of the minimal design base shear. EC's design base shear is identical to NZS's but IS yields the lowest design base shear for a given danger.

However, it was found by (Gadade, et al., 2018) that the IS Code buildings have a greater base shear than IBC Code. Furthermore, (Itti, Pathade and Karadi, 2014) represent that IS Code buildings have a greater base shear than IBC design code, the base Shear of IS buildings is higher than that of IBC buildings by 26.10% (Chebihi and Laouami, 2014) show that RPA outcomes are close to those of UBC, but EC provides smaller base shears than RPA and UBC. Waris, Al-Jabri and EL-Hussain, 2017 represent that UBC base shear values were 7.6, 4.1, and 3.2 times higher and IBC offered base shear values that were 6.3, 3.3, and 2.6 times higher than those from OSC which is based on EC.

The outcomes by (Doğangün and Livaoğlu, 2006) for equivalent ground types listed in other codes, EC often delivers the larger base shear. Correspondingly (Nahhas, 2017) discovered that SBC base shear is higher in most sites than UBC, and the same is true for overturning moments. SBC seismic maps may be inaccurate. Furthermore, (Taha and Hasan, 2018) observed that the ISC 2014 requirements result in a significant rise in base

shear forces as compared to ISC 1997. However, (Al-Obaidi, Jokhio and Abu-Tair, 2019) stated that ASCE in the base shear results, which demonstrates more than AS by 15%. Atique and Wadud, 2001 illustrate that BNBC provides the lowest value for base shears compared to both codes UBC and NBC-India. Finally, (Fenwick, Lau and Davidson, 2002) demonstrate that EC requires up to 4.0 times, but the UBC and IBC up 1.9 times, the base shear required by the NZS.

The proportion of all these studies for the design code which provided a minimum base shear is represented in Fig. 7.

D. In Terms of Displacement

Some studies show the disparity in displacement values. As found by (Santos, Lima and Arai, 2012) in the EC consideration, the displacements achieved with this code are substantially smaller than those created with the ASCE and other South American codes such as (NSR, CEC, NCh, and NBR). As well (Chebihi and Laouami, 2014) present that EC delivers smaller displacements than RPA and UBC. Results of (Wagh, Narkhede and Salunke, 2018) display that story displacements for EC are the lowest as compared to IS and NZS. The IS provides no modeling restrictions, leaving it up to the capabilities of individual designers. However, (Itti, Pathade and Karadi, 2014) summarize that IS Code buildings have a larger displacement than IBC Code structures, with IS displacements being 11.00% higher than IBC. Apart from (Rajeev, Meena and Pallav, 2019) the Indian code IS has a larger displacement capacity than the BS and EC codes; the IS provides a 19% and 26% increase in displacement capacity, respectively, over the BS, and EC. Otherwise, (Karthiga, et al., 2015) demonstrate how the minimal displacement is calculated using displacement data. In contrast to the IS, the EC has increased by 22%, the ASCE has increased by 20%, and the BS has increased by 19%. As a result, structures built in compliance with the IS are more rigid and draw more seismic pressures. Doğangün and Livaoğlu, 2006 explain that the highest lateral displacement values for the buildings are supplied by EC, whereas the lowest values are given by UBC. Fig. 8 shows the proportion of design codes that provided minimal displacement based on all of these studies.

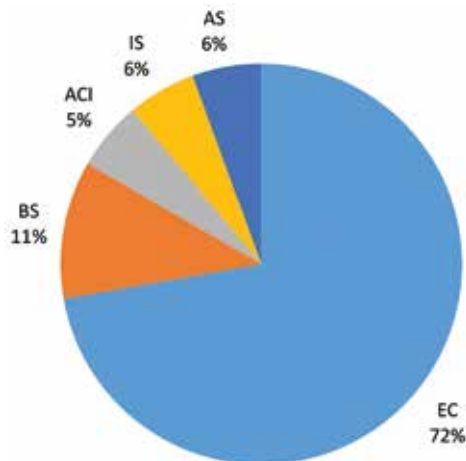### E. Safe and Conservative Aspects

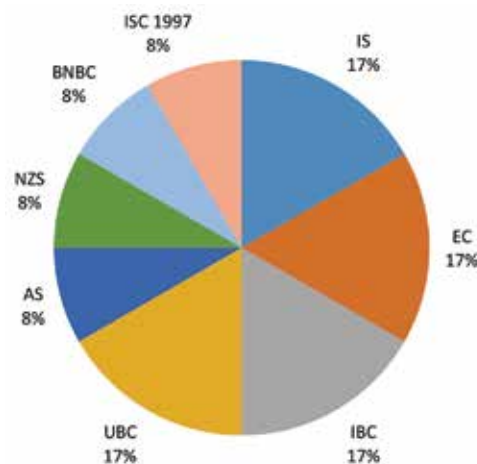Fig. 6. The ratio of economic parameter showing the most cost-effective design code.

Fig. 7. The ratio of the base shear parameter.

Regarding the safe design of construction, the majority of the research concludes that Eurocode is the safest and most conservative. According to the findings of (Landingin, et al., 2013) the EC provisions were seen to be safer, and EC was shown to be more conservative than NSCP and IBC. In comparison to EC, the BS code applies larger partial safety factors to loads at the ultimate limit state, whereas the EC is more conservative in terms of partial safety factors for loadings, the EC is more flexible, the requirement for EC is stricter, and safer than the BS as discussed by (Nwoji and Ugwu, 2017; Nwofor, Sule and Eme, 2015; Ng, Loo and Bong, 2006). Likewise, (Hawileh, Malhas and Rahman, 2009) the EC code flexural requirements were revealed to be slightly more conservative, with only a little practical deviation from the ACI norms. In comparison to other standards and procedures (Gadade, et al., 2018) illuminated that BNBC is the least conservative, and discovered that EC is more conservative than NSCP, Iranian code, and IBC. In addition, (Atique and Wadud, 2001) demonstrate that BNBC and NBC-India are the least conservative, UBC is more conservative than both these codes, on average, UBC is 2.61 times more conservative than the BNBC. According to the findings of (Arezoumandi, et al., 2021) ACI, EC, AS, and BSLJ regulations are conservative for shear strength values of 88, 78, 95, and 100%, respectively. This demonstrates that the Japanese code relations are conservative for all data. But (Pacheco, et al., 2019) present that there are no major variations in the bias of the ultimate moment between the two comparison vectors of EC versus ACI requirements.

However, the results of a few of them were deduced differently. According to (Al-Obaidi, Jokhio and Abu-Tair, 2019) the American code ASCE is more conservative, and it is the most widely utilized among engineers abroad. The ASCE is more detailed and complex than the AS, whereas the AS is more generic, with just one time-history equation, making the Australian code safer because the structure is subject to more earthquake time. Finally, (Waris, Al-Jabri and EL-Hussain, 2017) add that, compared to OSC, both UBC and IBC provide structural seismic loads that are rather conservative. Fig. 9 depicts the fraction of all design codes that are the safest and most conservative.

V. Evaluation and Assessment

According to the outcomes of this study and a review of past research, this study recommends the following for future research:

1. Quality control skills should be plentiful among the stakeholders involved in any initiatives that are undertaken.
2. The site engineers must be recruited based on their understanding of the applicable design codes. Weather and environmental conditions must be taken into account, and materials must be maintained, controlled, and manufactured.
3. Defects and violations should be addressed, and appropriate social, technological, and administrative frameworks should be created and implemented.
4. The establishment of a regulatory framework is accompanied by other regulations and features. Financial, economic, and social regulations, as well as a well-functioning market and well-informed customers, are all examples. Additional
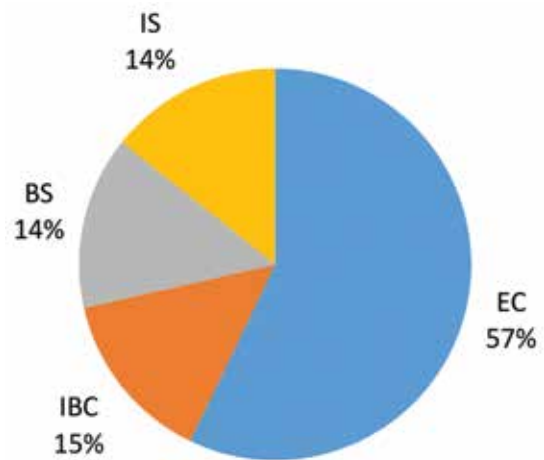
Fig. 8. The ratio of displacement parameter.
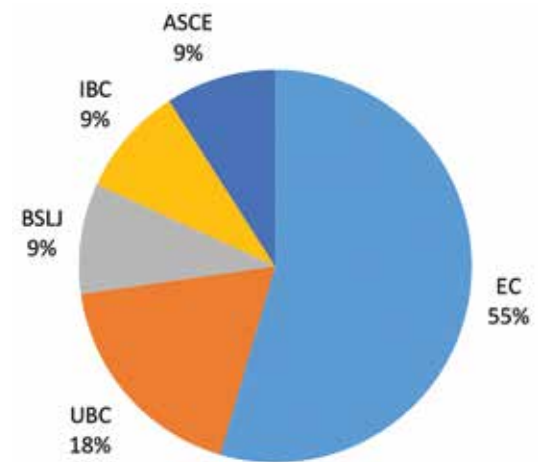
Fig. 9. The ratio of the conservative parameter.

research into these laws and elements may be possible.

The findings of this study would benefit the construction industry right away, allowing them to establish general properties and cost-effective solutions by utilizing the most appropriate unified construction building design code that is justified nationally and chosen as the country's most common construction regulatory. This will enable a responsive market with universal norms that protect and prioritize health and safety.

VI. Conclusion

The following conclusions were reached after extensive research on the comparison of several countries' building construction design codes:

1. There are substantial variances in construction requirements, as well as a considerable shift in structural analysis and design processes being used on structures in the KRI region. The research reported in this paper contributes to a better understanding of the significance of dealing with design codes and choosing the best one.
2. Almost every study that compared design codes in terms of detail available discovered that Eurocode is the most thorough code. Furthermore, provide all relevant information

for the analysis and design process to develop a structure in the safest and most secure manner feasible. Because it encompasses all design and construction rules for all types of buildings, as well as new and historic materials, the EC is the most sophisticated international standard and is preferred.

3. The majority of studies explored that the most economic design code is EC.

4. The findings differed in terms of the structure's base shear and displacement; however, the design codes for a minimal base shear are EC, IS, and American codes, with EC providing the smallest displacement.

5. Most of the studies show that EC is safest and most conservative compared to other codes.

6. All developments in the Kurdistan Region of Iraq (KRI) should adhere to the same national criteria. The various judgments of fault by different code enforcement authorities and organizations, or in different geographical zones, might be another field for further investigation. The Kurdistan region of Iraq is organized into four governorates; each one is using a set of building design codes randomly.

REFERENCES

ACI Committee, 2005. Building Code Requirements for Structural Concrete (ACI 318-05) and Commentary (ACI 318R-05). American Concrete Institute, United States.

ACI, C. 2002. Building Code Requirements for Reinforced Concrete (ACI318-02). American Concrete Institute, United States.

Adewuyi, A.P. and Franklin, S.O. 2011. Analytical investigation of prestressed concrete structures incorporating combined post-tensioned and post-compressed reinforcements. *ARPN Journal of Engineering and Applied Science*, 6, p.12.

Al-Fahad, J.Y. 2012. *Reform of Building Codes, Regulations, Administration and Enforcement in Kuwait: Within the Legal, Administrative, Technical and Social Framework*. Loughborough University, United Kingdom.

Allen, E. and Iano, J. 2019. *Fundamentals of Building Construction: Materials and Methods*. John Wiley and Sons, United States.

Alnuaimi, A.S., Patel, I.I. and Al-Mohsin, M.C. 2013. Design results of RC members subjected to bending, shear, and torsion using ACI 318: 08 and BS 8110: 97 building codes. *Practice Periodical on Structural Design and Construction,* 18, pp.213-224.

Al-Obaidi, A., Jokhio, G. and Abu-Tair, A. 2019. Comparison between American and Australian code through seismic effects. *In: IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, United States. pp.1-8.

Al-Sayed, M. and Waris, S.M.B. 2017. *Seismic Analysis of Residential Units in Oman*. PhD Thesis, Sultan Qaboos University, Oman.

Al-Taie, E., Al-Ansari, N. and Knutsson, S. 2014. The need to develop a building code for Iraq. *Engineering*, 6, pp.610-632.

Amer, A.N., Sobaih, M. and Adel, A. 2016. Investigation of effects of capacity spectrum method on performance evaluation of multi-story buildings according to the IRAQI seismic code requirements. *Open Journal of Civil Engineering,* 6, pp.420-441.

Arezoumandi, M., Mohaghegh Zadeh Davani ASL, M. S., Mobargha, S.O. and Afshar, S. 2021. Verification of shear strengths' provisions of existing codes for self-consolidating concrete. *Journal of Structural and Construction Engineering*, 7, 172-183.

Arya, C. 2018. *Design of Structural Elements: Concrete, Steelwork, Masonry and Timber Designs to British Standards and Eurocodes*. CRC Press, United States.

Association, C.C.S. 2005. CSA A23. 3-04: Design of Concrete Structures. *Technical Committee on Reinforced Concrete Design*. Canadian Standards Association, Rexdale.

Association, C.S. 2004. CSAA23. 3-04 Design of Concrete Structures: Structures Design. Canadian Standards Association**,** Mississauga. pp.62-64.

Athanasopoulou, A., Dimova, S., Fuchs, M., Sousa, M.L., Pinto, A., Nikolova, B. and Iannacone, S. 2018. *State of Eurocode 8 Implementation in the European Union*. 16th European Conference on Earthquake Engineering, Thessaloniki. pp.18-21.

Atique, F. and Wadud, Z. 2001. A comparison of BNBC-93 with other building codes with respect to earthquake and wind analysis. *The Eighth East Asia-Pacific Conference on Structural Engineering and Construction*. Nanyang Technological University, Singapore.

Australia, S.W. 2012. *Safe Design of Structures: Code of Practice*. Safe Work Australia, Australia.

Authority, B.I. 1992. *The New Zealand Building Code Handbook*. Standards New Zealand.(NZBC), New Zealand.

Bakhoum, M.M., Mourad, S.A. and Hassan, M.M. 2016. Comparison of actions and resistances in different building design codes. *Journal of Advanced Research,* 7, pp.757-767.

Balendra, T., Lam, N., Wilson, J.L. and Kong, K. 2002. Analysis of long-distance earthquake tremors and base shear demand for buildings in Singapore. *Engineering Structures,* 24, pp.99-108.

Baltimore, H. 2009. Proposed Changes to the 2009 Editions of the International Building Code. *International Code Council 2009/2010 Code Development Cycle.*

Bano, S., Izhar, T. and Mumtaz, N. 2019. Design of RC Member Using Different Building Code: A Review. *International UKIERI Concrete Congress Congress.*

Bartlett, R., Halverson, M.A. and Shankle, D.L. 2003. *Understanding Building Energy Codes and Standards*. Pacific Northwest National Lab (PNNL), Richland, WA (United States).

Beeby, A.W. and Narayanan, R. 2005. *Designers' Guide to EN 1992-1-1 and EN 1992-1-2. Eurocode 2: Design of Concrete Structures: General Rules and Rules for Buildings and Structural Fire Design*. Thomas Telford Ltd., United Kingdom.

Beeby, A.W., Narayanan, R.S. and Narayanan, R. 1995. *Designers' Handbook to Eurocode 2: Design of Concrete Structures*. Thomas Telford Ltd., United Kingdom.

Belazougui, M. 2017. *Algerian Seismic Building Code: Main Features Of The New Draft RPA 2015*. 16th World Conference on Earthquake, Santiago Chile.

Ben-Joseph, E. 2012. Codes and standards. *In: The Oxford Handbook of Urban Planning*. Oxford University Press, United Kingdom.

Bhavsar, M.J., Choksi, K.N., Bhatt, S.K. and Shah, S.K. 2014. Comparative study of typical RCC building using Indian standards and Euro standards under seismic forces. *International Journal of Scientific and Research Publications,* 4, pp.1-4.

Billington, D.P. 1985. *The Tower and the Bridge: The New Art of Structural Engineering*. Princeton University Press, United States.

Bisch, P., Carvalho, E., Degee, H., Fajfar, P., Fardis, M., Franchin, P., Kreslin, M., Pecker, A., Pinto, P. and Plumier, A. 2012. *Eurocode 8: Seismic Design of Buildings Worked Examples*. Publications Office of the European Union, Luxembourg.

Bond, A., Brooker, O., Harris, A., Harrison, T., Moss, R., Narayanan, R. and Webster, R. 2006. *How to Design Concrete Structures Using Eurocode 2*. Concrete Centre, United Kingdom.

Bose, P., Dubey, R. and Yazdi, M. 1992. *Comparison of Codal Provisions Suggested by Various Countries*. Earthquake Engineering, Tenth World Conference, Rotterdam, Netherlands. pp.5747-5750.

Canada, N.R.C.O. 2015. *National Building Code of Canada, 2015*. National Research Council Canada, Canada.

Canchig Cola, M.V., 2016. *Comparative analysis of the structural design of a reinforced concrete housing project applying the Standards of the Ecuadorian Construction Code (CEC 2002) and the Ecuadorian Construction Standard (NEC 2015)* (Bachelor's thesis, Quito / UIDE / 2016).

Canisius, T.G., Baker, J. and Diamantidis, D. 2011. *Structural Robustness Design for Practising Engineers*. COST Action TU0601 Robustness of Structures, Switzerland. p.139.

Chebihi, A. and Laouami, N. 2014. *Comparative Seismic Study between Algerian Code (RPA99), European Code (EC8) and American Code (UBC97)*. 2nd European Conference on Earthquake Engineering and Seismology, Istanbul, Turkey.

Cheng, F. 2013. Using building codes and standards to improve construction quality and safety. In: *ICCREM 2013:* International Conference on Construction and Real Estate Management 2013.

Clemmensen, B., 2003, November. Building Codes–a good tool in the right context. In: *Paper presented by the Chairman, Canadian Commission on Building and Fire Codes to the Global Summit on Performance-Based Building Codes*.

National Research Council Canada. Associate Committee on the National Building Code, 1990. National building code of Canada. Associate Committee on the National Building Code, National Research Council.

Code, P. 2005. *Eurocode 2: Design of Concrete Structures-Part 1-1: General Rules and Rules for Buildings*. British Standard Institution, London.

Code, U.B. 1997. *International Building Code*. International Code Council, USA.

Committee, A. 2008. *Building Code Requirements for Structural Concrete (ACI 318-08) and Commentary*. American Concrete Institute, United States.

Committee, E.P. 2007. *ECP-203: 2007-Egyptian Code for Design and Construction of Concrete Structures*. HBRC, Giza.

Conrad, R.T. and Winkel, S.R. 1998. *Design Guide to the 1997 Uniform Building Code*. John Wiley and Sons, United States.

Cote, A.E. and Grant, C.C. 2008. Codes and standards for the built environment. In: *Fire Protection Handbook*. vol. 20. National Fire Protection Association, Massachusetts. p.1-61.

Davidson, B. 2008. *Base Shear Scaling*. Proceedings NZSEE Technical Conference, New Zealand.

De León, A.E.D. 2010. National building code of India and the international building code: An introduction. In: *Indo-US Forensic Practices: Investigation Techniques and Technology*. American Society of Civil Engineers, United States. pp.111-120.

Deepshikha and Basu, T. 2011. A review on synthesis and characterization of nanostructured conducting polymers (NSCP) and application in biosensors. *Analytical Letters,* 44, pp.1126-1171.

Doğangün, A. and Livaoğlu, R. 2006. A comparative study of the design spectra defined by Eurocode 8, UBC, IBC and Turkish Earthquake Code on R/C sample buildings. *Journal of Seismology*, 10, pp.335-351.

Dollet, C. and Guéguen, P. 2022. Global occurrence models for human and economic losses due to earthquakes (1967-2018) considering exposed GDP and population. *Natural Hazards*, 110, pp.349-372.

Donduren, M.S. and Omeed, A. 2018. Comparison of the Masonry building codes Turkish, Euro and USA building codes. *Journal of International Environmental Application and Science,* 13, pp.89-96.

EN, B. 2004. *Eurocode 2: Design of Concrete Structures-Part 1-1: General Rules and Rules for Buildings*. European Committee for Standardization (CEN), Belgium.

Eurocode, B. 2006. *Actions on Structures-Part 1-7: General Actions-Accidental Actions*. British Standard, London.

Faizian, M. and Ishiyama, Y. 2004. Comparison of Seismic Codes of 1981 Japan (BSLJ), 2000 USA (IBC) and 1999 Iran (ICS). 10th World Conference on Earthquake Engineering, Canada.

Fenwick, R., Lau, D. and Davidson, B. 2002. A comparison of the seismic design requirements in the New Zealand loadings standard with other major design codes. *Bulletin of the New Zealand Society for Earthquake Engineering,* 35, pp.190-203.

Franklina, S.O. and Mensahb, K.K. 2011. A Comparative Study of EC2 and BS8110 Beam Analysis and Design in a Reinforced Concrete Four Storey Building. *Journal of Basic and Applied Scientific Research*.

Gadade, M.S.R., Student, P., Vaidkar, N. and Kalwane, U. 2017. Literature Review on Comparative Study on Seismic Analysis of RC Frame Structure By Various Code. *International Journal of Agriculture, Environment and Biotechnology*, 10, pp.703-707.

Galíndez, N. and Thomson, P. 2007. Performance of steel moment-frame buildings designed according to the Colombian code NSR-98. *Engineering structures,* 29, pp.2274-2281.

Gilbert, R.I., Mickleborough, N.C. and Ranzi, G. 2016. *Design of Prestressed Concrete to AS3600-2009*. CRC Press, United States.

Guendelman, T., Saragoni, G. and Verdugo, R. 2012. Chilean Emergency Seismic Design Code for Buildings after El Maule 2010 Earthquake. *Proc. 15 WCEE, paper*, p.4480.

Gulvanessian, H. 2001, EN1990 Eurocode-Basis of Structural Design. Proceedings of the Institution of Civil Engineers-Civil Engineering. Thomas Telford Ltd., United Kingdom. pp.8-13.

Gulvanessian, H. and Holický, M. 1996. *Designers' Handbook to Eurocode 1: Basis of Design*. Thomas Telford, United Kingdom.

Gulvanessian, H., Calgaro, J.A. and Holicky, M. 2012. *Designers' Guide to Eurocode: Basis of Structural Design*. ICE Publishing, Telford.

Gulvanessian, H., Formichi, P. and Calgaro, J.A. 2009. *Designers' Guide to Eurocode 1: Actions on Buildings: EN1991-1-1 and-1-3 TO-1-7*. Thomas Telford Ltd., United Kingdom.

Habibi, A. and Asadi, K. 2013. Seismic Performance of RC Frames Irregular in Elevation Designed Based on Iranian Seismic Code. *Journal of Rehabilitation in Civil Engineering*, 1, pp.40-55.

Haldar, P. and Singh, Y. 2009. Seismic performance and vulnerability of Indian code designed RC frame buildings. *ISET Journal of Earthquake Technology*, 46, pp.29-45.

Hawileh, R.A., Malhas, F.A. and Rahman, A. 2009. Comparison between ACI 318-05 and Eurocode 2 (EC2-94) in flexural concrete design. *Structural Engineering and Mechanics: An International Journal*, 32, pp.705-724.

Heady, E.J., Currie, S. and LLP, H. 2012. Construction law-the history is ancient!! *Common Sense Contract Law*.

Higgins, J. and Rogers, B. 1998. Designed and Detailed (BS 8110: 1997). *British Cement Association*.

Hitchin, R. 2008. Can Building Codes Deliver Energy Efficiency. Royal Institution of Chartered Surveyors, United Kingdom.

Imashi, N. and Massumi, A. 2011. A comparative study of the seismic provisions of Iranian seismic code (standard no. 2800) and international building code 2003. Asian Journal of Civil Engineering (Building and Housing). 12. 579-596.

Institute, A. 2011. Building Code Requirement for Structural Concrete (ACI 318-11) and Commentary. America Concrete Institute Farmington Hills, Michigan.

Institution, B.S. 2004. *Eurocode 2: Design of Concrete Structures: Part 1-1: General Rules and Rules for Buildings*. British Standards Institution, United Kingdom.

Islam, A.S., Jameel, M., Uddin, M. and Ahmad, S.I. 2011. Simplified design guidelines for seismic base isolation in multi-storey buildings for Bangladesh National Building Code (BNBC). *International Journal of Physical Sciences,*

6, pp.5467-5486.

Itabashi, M. and Fukuda, H. 1999. The Japanese Building Standard Law and a series of steels for earthquake-resistant building structures. *Technology, Law and Insurance,* 4, pp.37-44.

Itti, S., Pathade, A. and Karadi, R.B. 2014. A Comparative Study on Seismic Provisions Made in Indian and International Building Codes for RC Buildings. *International Journal of Civil, Architectural, Structural and Construction Engineering*, 8, pp.457-463.

Izhar, T. and Dagar, R. Comparison of reinforced concrete member design methods of various countries. *International Journal of Civil Engineering and Technology (IJCIET)*, 9, pp.637-646.

Jawad, A.A.H. 2006. Strength design requirements of ACI-318M-02 code, BS8110, and Eurocode 2 for structural concrete: A comparative study. *Journal of Engineering and Development*, 10, pp.22-28.

Karthiga, S., Titus, H.E., Hazarika, R.R. and Harrish, M. 2015. Design and comparison of a residential building (G+ 10) for seismic forces using the codes: IS1893, Euro code 8, ASCE 7-10 and British code. *International Journal of Research in Engineering and Technology*, 5, pp.205-209.

Keeler, M. and Vaidya, P. 2016. *Fundamentals of Integrated Design for Sustainable Building*. John Wiley and Sons, United States.

Koti, K.N.V. 2017. Comparative analysis of a high-rise structure using various international codes. *International Journal for Innovative Research in Science and Technology*, 4, 6.

Landingin, J., Rodrigues, H., Varum, H., Arêde, A. and Costa, A. 2013. Comparative analysis of RC irregular buildings designed according to different seismic design codes. *The Open Construction and Building Technology Journal,* 7, pp.221-229.

Lian, W. and Wang, S. 2006. Discussion on inter-storey displacement angle of high-rise building. *Building Structure,* 36, pp.49-55.

Liebing, R.W. 1987. *Construction Regulations Handbook*. Wiley-Interscience, New Jersey.

Liew, Y. 2009. British standard (BS 8110) and Eurocode 2 (EC2) for reinforced concrete column design. *Microsoft incorporation (2010).* Microsoft Excel. University of Technology Malaysia, Malaysia.

Listokin, D. and Hattis, D.B. 2005. Building codes and housing. *Cityscape*, 8, pp.21-67.

Marino, E.M., Nakashima, M. and Mosalam, K.M. 2005. Comparison of European and Japanese seismic design of steel building structures. *Engineering Structures,* 27, pp.827-840.

Marpal, K.A.F. 2010. *Comparison of Slab Design between Bs 8110 and Eurocode 2 by Using Microsoft Excel.* Universiti Malaysia Pahang, Malaysia.

Mcintosh, R. and Pezeshk, S. 1997. Comparison of recent US seismic codes. *Journal of Structural Engineering,* 123, pp.993-1000.

Mehrabian, A. 1996. *Uncertainty Evaluation of the 1994 UBC Base Shear Formula for Concrete Moment-Resisting Frames*. San Jose State University, San Jose, CA.

Meijer, F. and Visscher, H. 2008. Building regulations from an European perspective. *Proceedings of COBRA*, Delft University of Technology, Netherlands.

Melkers, J. and Willoughby, K. 1998. The state of the states: Performance-based budgeting requirements in 47 out of 50. *Public Administration Review,* 58, pp.66-73.

Menegon, S.J., Wilson, J.L., Lam, N.T. and Mcbean, P. 2018. RC walls in Australia: Seismic design and detailing to AS 1170.4 and AS 3600. *Australian Journal of Structural Engineering,* 19, pp.67-84.

Moehle, J.P., Hooper, J.D. and Lubke, C.D. 2016. *Seismic Design of Reinforced Concrete Special Moment Frames*. US Department of Commerce, National Institute of Standards and Technology, United States.

Mosley, W.H., Hulse, R. and Bungey, J.H. 2012. *Reinforced Concrete Design: To Eurocode 2*. Macmillan International Higher Education, London.

Moss, R. and Webster, R. 2004. EC 2 and BS 8110 compared. *Structural Engineer*, 82, pp.33-38.

Mourad, S.A. and Hassan, M. 2009. The response modification factor in seismic design codes for steel MRF. *In: Conference: 13th ICSGE-International Conference on Structural and Geotechnical Engineering*, 627-629.

Muir, L. and Duncan, C.J. 2011. *The AISC 2010 Specification and the 14th Edition Steel Construction Manual*. Vol. 2011. Structures Congress, United States. pp.661-675.

Nahhas, T.M. 2011. A comparison of IBC with 1997 UBC for modal response spectrum analysis in standard-occupancy buildings. *Earthquake Engineering and Engineering Vibration,* 10, pp.99-113.

Nahhas, T.M. 2017. A comparison of Saudi building code with 1997 UBC for provisions of modal response spectrum analysis using a real building. *Open Journal of Earthquake Research,* 6, p.98.

Nandi, L. and Guha, P. 2014. Design comparison of different structural elements by using different international codes. *International Journal of Engineering Research and Technology*, 3, 2161-2164.

NBR, A. 2006. 15421: Design of earthquake-resistant structures-Procedure. *Rio de Janeiro*.

Ng, C.K., Loo, S.W. and Bong, Y.P. 2006. *Reinforced concrete beams at ultimate flexural limit state: Comparison of BS 8110 and Eurocode 2*. Proceedings of the 6th Asia-Pacific Structural Engineering and Construction Conference (APSEC 2006), Kuala Lumpur, Malaysia, pp.5-6.

Noor, M., Ansari, M. and Seraj, S. 1997. Critical evaluation and comparison of different seismic code provisions. *Journal of Civil Engineering*, 25, pp.1-20.

Nowak, A.S. and Rakoczy, A.M. 2012. Reliability-based calibration of design code for concrete structures (ACI 318). Anais do 54 Congresso Brasileiro do Concreto, Brasil.

Nowak, A.S. and Szerszen, M.M. 2003. Calibration of design code for buildings (ACI 318): Part 1-Statistical models for resistance. *ACI Structural Journal,* 100, pp.377-382.

Nwofor, T., Sule, S. and Eme, D. 2015. A comparative study of Bs8110 and Eurocode 2 standards for design of a continuous reinforced concrete beam. *Journal Impact Factor,* 6, pp.76-84.

Nwoji, C. and Ugwu, A. 2017. Comparative Study of BS 8110 and Eurocode 2 in structural design and analysis. *Nigerian Journal of Technology,* 36, pp.758-766.

O'Bannon, R.E. 1973. *Building Department Administration: International Conference of Building Officials*. ICC, United Arab Emirates.

Obrien, E.J. and Dixon, A.S. 1995. *Reinforced and Prestressed Concrete Design: The Complete Process*. CRC Press, United States.

Oyenuga, V. 2011. *Reinforced Concrete Design-a Consultant/Computer Based Approach*. 2nd ed. Astro Limited, Lagos Nigeria.

Pacheco, J., De Brito, J., Chastre, C. and Evangelista, L. 2019. Uncertainty models of reinforced concrete beams in bending: Code comparison and recycled aggregate incorporation. *Journal of Structural Engineering,* 145, 04019013.

Parisi, M.A. 2008. The Eurocode for earthquake-resistant design: An outline. *Practice Periodical on Structural Design and Construction*. 13, pp.161-166.

Pekau, O., Zielinski, Z. and Lin, L. 1995. Displacement and natural frequencies of tall building structures by finite story method. *Computers and Structures,* 54, pp.1-13.

Poston, R.W. and Dolan, C.W. 2008. Reorganizing ACI 318. *Concrete International*, 30, pp.43-47.

Rajeev, A., Meena, N.K. and Pallav, K. 2019. Comparative study of seismic design and performance of omrf building using Indian, British, and European codes. *Infrastructures,* 4, p.71.

Regulation, B. 2004. Reform of Building Regulation: Productivity Commission Research Report. *Productivity Commission, November.*

Reichert, J. 2005. *Will Change Management Master the Challenge of Introducing the Eurocodes to the UK?.* ProQuest Dissertations Publishing, University College of London, United Kingdom.

Reynolds, C.E., Steedman, J.C. and Threlfall, A.J. 2007. *Reynolds's Reinforced Concrete Designer's Handbook.* CRC Press, United States.

Rodrigues, H., Arêde, A., Varum, H. and Costa, A.G. 2013. Experimental evaluation of rectangular reinforced concrete column behaviour under biaxial cyclic loading. *Earthquake Engineering and Structural Dynamics,* 42, pp.239-259.

Rodrigues, H., Varum, H. and Costa, A. 2010. Simplified macro-model for infill masonry panels. *Journal of Earthquake Engineering,* 14, pp.390-416.

Santos, S.H.D., Lima, S. and Arai, A. 2012. Comparative study of codes for the seismic design of structures. *Revista IBRACON de Estruturas e Materiais,* 5, pp.812-819.

Şengöz, A. 2007. *Quantitative Evaluation of Assessment Methods in the 2007 Turkish Earthquake Code.* Middle East Technical University, Turkey.

Shkoukani, H.T. 1993. *General Rules for Design of Reinforced Concrete Structures According to Eurocode No. 2: Comparative with the ACI Code.* Proceedings of the First Palestinian Convention in Civil Engineering, Palestine.

Shodolapo, O. and Kenneth, K. 2011. A comparative study of EC2 and BS8110 beam analysis and design in a reinforced concrete four storey building. *Journal of Basic and Applied Scientific Research,* 1, pp.3172-3181.

Shuraim, A., Al-Haddad, M., Al-Zaid, R., Mirza, R. and Al-Sheref, K. *Seismic Provisions in the Saudi Building Code.* Proceedings of the 7th Saudi Engineering Conference. Citeseer, Riyadh.

Singh, Y., Khose, V.N. and Lang, D.H. 2012. *A Comparative Study of Code Provisions for Ductile RC Frame Buildings.* Proceedings of the 15th World Conference on Earthquake Engineering, Lisbon, Portugal, p.24-28.

Standard, A.A. 2011. *Building Code Requirements for Structural Concrete (ACI 318-11).* American Concrete Institute, United States.

Standard, B. 1986. *Structural use of concrete. BS8110.* British Standard, United Kingdom.

Standard, B. 2002. *Eurocode-Basis of structural design. Eurocode 0.* British Standard, United Kingdom.

Standard, B. 2004. *Eurocode 2: Design of Concrete Structures Part 1-1: General Rules and Rules for Buildings.* British Standard, United Kingdom. pp.230.

Standard, B. 2006. *Eurocode 1: Actions on structures.* British Standard, United Kingdom.

Tabsh, S.W. 2013. Comparison between reinforced concrete designs based on the ACI 318 and BS 8110 codes. *Structural Engineering and Mechanics,* 48, pp.467-477.

Taha, B.O. and Hasan, S.A. 2018. A comparative study of the seismic provisions between Iraqi Seismic codes 2014 and 1997 for Kurdistan Region/Iraq. *Eurasian Journal of Science and Engineering,* 4, pp.180-192.

Tq, A. and Given, D. 2017. 14.37. 5 Test 37 with Solutions. *Objectives and Methods of Analysis and Design, and Properties of Concrete and Steel.* Design of Concrete Structures, Kharagpur.

Van der Heijden, J. and De Jong, J. 2009. Towards a better understanding of building regulation. *Environment and Planning B: Planning and Design,* 36, pp.1038-1052.

Varum, H.S.A. 2003. *Seismic Assessment, Strengthening and Repair of Existing Buildings.* Universidade de Aveiro, Portugal.

Vaughan, E. and Turner, J. 2013. The value and impact of building codes. *Environmental and Energy Study Institute White Paper,* 20, pp.501-517.

Vesilind, P.A. 1995. Evolution of the American Society of Civil Engineers code of ethics. *Journal of Professional Issues in Engineering Education and Practice,* 121, p.4-10.

Wagh, A., Narkhede, T. and Salunke, P. 2018. Codal comparison of seismic analysis of a high-rise structure. *International Journal of Science Technology and Engineering,* 5, pp.39-45.

Walraven, J. 2008. *Eurocode 2: Design of concrete structures EN1992-1-1.* Symposium Eurocodes: Background and Applications, Brussels.

Waris, M., Al-Jabri, K. and El-Hussain, I. 2017. *Comparison of Oman Seismic Code for Buildings with International Counterparts.* World Conference on Earthquake (16WCEE), Santiago, Chile.

Wimo, A., Jönsson, L., Gustavsson, A., Mcdaid, D., Ersek, K., Georges, J., Gulacsi, L., Karpati, K., Kenigsberg, P. and Valtonen, H. 2011. The economic impact of dementia in Europe in 2008-cost estimates from the Eurocode project. *International journal of geriatric psychiatry,* 26, pp.825-832.

Xiaoguang, C., Jingshan, B., Youwei, S., Jianyi, Z. and Yudong, Z. 2012. Comparison of *Seismic Fortification Criterion of Eight Asian Countries.* 15th World Conference on Earthquake Engineering, Lisbon, Portugal.

Some Enzymatic and Non-enzymatic Antioxidants Response under Nickel and Lead Stress for Some Fabaceae Trees

Sargul A. Khudhur and Ikbal M. Albarzinji

Department of Biology, Faculty of Health and Science, Koya University,
Kurdistan Region - F.R. Iraq

Abstract—**This study investigates the effects of soil contamination by nickel and lead on some enzymatic and non-enzymatic antioxidants in addition to the nitrate reductase (NR) enzyme activity for Gleditsia triacanthos*, Leucaena leucocephala, and Robinia pseudoacacia* plant species. The results of this study show a significant increase in peroxidase enzyme activity and a significant decrease in catalase enzyme activity, proline, total carotenoids, and total carbohydrate content of leaves of the three species with increasing the concentration of Ni and Pb except for the total carbohydrate, which increased only for *L. leucocephala* species. Each NR enzyme activity and ascorbic acid content are increased significantly with increasing the concentration of Ni and Pb for *G. triacanthos*, *L. leucocephala,* and on the contrary, they decreased significantly for *R. pseudoacacia* species. From the result, we can conclude a general increase or decrease in leaves content of some antioxidants content for all the species, whereas there is some peculiarity according to the plant species regarding other contents, which in turn reflects different mechanisms of these species to tolerant heavy metal stress.**

Index Terms—**Ascorbic acid, Catalase, Heavy metals, Nickel and Lead, Nitrate reductase, Peroxidase, Proline.**

I. Introduction

Soils can be contaminated with heavy metals through industrial waste, gasoline, paint, mining waste, fertilizers, pesticides, animal fertilizers, irrigation of wastewater, and other sources. Lead (Pb) and nickel (Ni) can be considered the most commonly found heavy metals in contaminated sites. Several authors have reported that the toxicity of Pb and Ni could produce oxidative cellular damage when generating reactive oxygen species (ROS), which in turn leads to several detrimental effects on plant cells, so antioxidant enzymes can reduce or prevent the toxic effects of ROS induced by metal stress. (Nyiramigisha, et al., 2021; Kumari and Mishra, 2021 and Álvarez, et al., 2019).

ROS are usually produced due to heavy metals, high light intensity, high temperature, biotic stress, air pollutants, soil salinity, and drought, the imbalance between the ROS and antioxidants defense system in plants creates oxidative stress in the plants. ROS are produced in plants as by-products during many metabolic reactions, such as photosynthesis and respiration. The accumulation of ROS leads to disturbances in normal physiological processes and leads to damage to biomolecules, cells, and tissues. Disturbance in the ratio of pro-oxidant and anti-oxidant leads to potential damage to the cells and tissues. Disturbances in the normal redox state of cells can cause toxic effects. However, ROS has a role in signal transduction pathways by cellular antioxidant machinery, which involves detoxifying enzymes and non-enzymatic antioxidant compounds which can control toxic ROS. ROS are produced from molecular oxygen as a result of normal cellular metabolism. (Sachdev, et al., 2021; Ali, et al., 2020; Das and Aryadeep, 2014).

Antioxidants are the substances that are present in plants at the lower concentrations compared to that of oxidizable substrates, significantly delaying, or preventing the oxidation of substrates. Antioxidants such as carotenoids and ascorbic acid (AA) are non-enzymatic in nature whereas catalase (CAT) and peroxidase (POD) are the major enzymatic antioxidants. Enzymatic antioxidants function by breaking down and removing the free radicals from the cells while non-enzymatic antioxidants work by interrupting free radical chain reactions thus protecting cells from damage. The presence of antioxidant defense is universal in nature, and plants produce it to protect themselves from the ultraviolet light of the sun and ROS generated during photosynthesis that would cause irreparable damage to the plant tissues. Although the mechanisms of antioxidant defenses are changed from plant species to another, the existence of antioxidant defense is general (Ren, et al., 2021; Santos-Sánchez, et al., 2019; Kartoori, et al., 2018).

Gleditsia triacanthos, Leucaena leucocephala, and Robinia pseudoacacia plant species are planted as landscapes in the Iraqi Kurdistan region, and they are important for carbon sequestration, soil stabilization and re-vegetation of

landfills, mining areas, and wastelands, in biotherapy and landscaping. These plants produce high levels of biomass, grow rapidly, are easily cultivated, and most importantly, tolerate and accumulate high concentrations of heavy metals in the aboveground parts. These plants have numerous inherent characteristics that can be exploited to enhance phytoremediation and lower the cost of regeneration. These species can survive in severe environmental conditions with the exception of heavily frosted conditions and occur in a wide range of ecological settings. They are fast-growing species, capable of reaching maturity in 6–7 months to produce a vast amount of seeds that can germinate into numerous seedlings to carry on further remediation of the polluted site. They are endowed with high proficiency for nitrogen fixation through nodule formation and can substantially revitalize microbial mass and micro-bioactivities to pave the way for the re-establishment of self-sustaining plant communities over the polluted sites (Ssenku, et al., 2017).

One of the 14 essential mineral elements for plants is nickel which is required in small amounts for healthy growth and development. Higher concentrations of Ni in plant cells result in alterations at the physiological, biochemical, and cellular levels leading to severe damage to plants. The most common symptoms of Ni^{2+} toxicity in plants are inhibition of photosynthesis and mitotic activities, inhibition of sugar transport, and reductions in plant growth. Extremely high soil Ni concentrations have left some farmland unsuitable for growing crops, fruits, and vegetables.

In general, the value between 0.5 and 5 mg·kg^{-1} of plant dry weight is the accepted Ni content in plant tissues whereas, for soil is vary between 5 and 150 mg·kg^{-1} dry weight. Total uptake of Ni by plants are depend on many factors including soil solution concentration of Ni^{2+}, the metabolism of plant, soil solution acidity, exist of other metals, and soil contain of organic matter. Regarding lead it is not considered as essential element, also, for living organisms it is relatively considered as unavailable because they are immobe in the soil solution and their transport through the short transport system from roots to plants is very limited. In uncontaminated soil the concentration of Pb is varied within the range of 10–50 mg·kg^{-1}; however, its concentration can be expected to range from 30 to 100 mg·kg^{-1} in soil with low-level contamination (Kacálková, et al., 2014; Vatansever, et al., 2010). Ni and Pb uptake by plant roots is not mediated by the same mechanisms. Ni uptake is mainly carried out by roots through a passive diffusion and/or active transport. Pb is generally taken up through the root system through passive diffusion, mainly in the apoplast through the intercellular spaces, following the water movement within the plant (Amari, et al., 2017).

This study aimed to investigate the effects of Ni and Pb elements as soil polluted heavy metals on changes in some enzymatic and non-enzymatic antioxidants in addition to nitrate reductase (NR) in leaves of *G. triacanthos, L. leucocephala, and R. pseudoacacia* as three forest tree species belong to fabaceae family.

II. Materials and Methods

A. Plant Materials and Growth Conditions

The ripe seed pods of *G. triacanthos, L. leucocephala,* and *R. pseudoacacia* were collected on March 2, 2021, in Koya University- Campus (Koya city, Erbil, Iraq, located at 44°38 E, 36°4N and 570 m of altitude above sea level). The pods were collected directly from 10 years aged, with straight bales and well-formed trees by climbing the trees manually. The pods were soaked in ordinary (tap) water at room temperature for 24 h. A factorial, completely randomized design experiment with three replicates was conducted under the ambient environmental conditions with the temperature that varied from (14°C to 39°C) and relative humidity from 19% to 64%.

The healthy plant seeds were cultivated on 25 March 2021. Seeds were sown (2.5 cm depth) in black polyethylene bags (12 cm diameter and 25 cm height) filled with 3 kg of loamy soil mixed with Ni (0, 15, 30 or 45 mg NiCl$_2$ kg^{-1} soil) that was denoted as Ni_0, Ni_{15}, Ni_{30}, and Ni_{45} or Pb (0, 15, 30, or 45 mg PbCl$_2$ kg^{-1} soil) that denoted as Pb_0, Pb_{15}, Pb_{30}, and Pb_{45} and their interactions. (Abdullah, et al., 2018; Olorunmaiye, 2019 and Dirr and Heuser, 1987). The bags were suitably watered with ordinary (tap) water to provide possible uniform soil moisture conditions. 70 days after seed germination 3rd–4th fully matured leaves (Fig. 1) were taken to estimate some enzymatic and non-enzymatic antioxidants in addition to NR enzyme, as follows:

B. Extraction Some Enzymatic and Non-enzymatic Antioxidants and NR Enzyme Activity in the Leaves of the Studied Plant

In fresh leaf samples, the activity of the following enzymes was measured; POD enzyme activity was estimated according to the method described by (Nezih, 1985), where the absorbance was read at 420 nm and the activity of POD [µg.g^{-1} fresh weight] was calculated. The activity of the CAT enzyme was estimated according to the method of Aebi (1974), where the absorbance was read at 240 nm using a UV-spectrophotometer (GENESYS 10 UV Spectrometer), Thermo Electron Corporation, USA), and the activity of CAT [microgram. gram^{-1}] was calculated. Although NR enzyme is not an antioxidant, we measure it because has an important role in the Fabaceae family species used in the study, the activity of this enzyme was measured by preparing enzyme extract using the method (Ahmad, et al., 2010), for this enzyme the absorbance was read at 540 nm using a spectrophotometer (PD-303) and the activity of NR [µM.l^{-1}] was calculated. AA [g. l^{-1}] was determined by the method described by Elbsheer (2018). The Pr [µg.ml^{-1}] content in fresh leaf samples was determined by the method of Bates, Waldron, and Tears, (1973). The absorbance of the toluene layer was read at 520 nm, using a spectrophotometer (PD-303). The amount of total carotenoids (TC) [mg. g^{-1} fresh weight] was estimated according to the method of Lichtenthaler and Wellburn (1983) and total carbohydrates (TCHOs) [%] were determined by the method of Joslyn (1970).

C. Statistical Analysis

The experiment was conducted according to a factorial completely randomized design. Each treatment was replicated 3 times and ten bags were considered as an experimental unit. Treatment means were compared by the analysis of variance using the SAS Statistical program. Duncan's multiple range test ($\alpha \leq 0.05$) was used for comparing treatment means (Al-Mohammadi and Al-Mohammadi, 2002).

III. Results and Discussion

The results shown in Tables I-III indicated that the effect of Ni and Pb was significant on some leaves
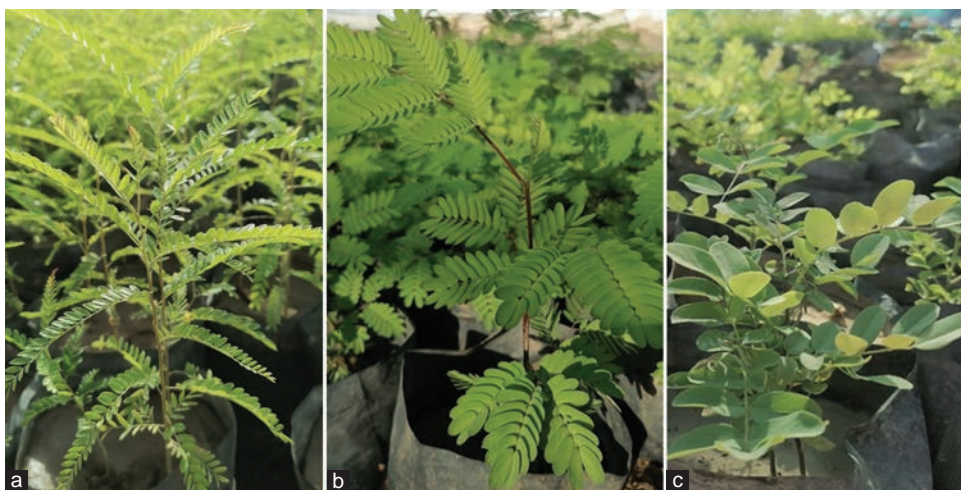
Fig. 1. Three month old of (a) *Gleditsia triacanthos*, (b) *Leucaena leucocephala,* and (c) *Robinia pseudoacacia* species.

TABLE I

EFFECTS OF NICKEL, LEAD, AND THEIR INTERACTIONS ON SOME ENZYMATIC, NON-ENZYMATIC ANTIOXIDANTS AND NITRATE REDUCTASE ACTIVITY OF *GLEDITSIA TRIACANTHOS* PLANT LEAVES

Treatments	Enzymatic Antioxidants and Nitrate Reductase			Non-Enzymatic Antioxidants			
	Catalase (microgram. Gram^{-1})	Peroxidase (microgram. gram^{-1})	Nitrate Reductase (micromole.L^{-1})	Ascorbic Acid (gram.L^{-1})	Proline (microgram.ml^{-1})	Total Carotenoids (mg. g^{-1} fresh weight)	Total Carbohydrate (%)
Ni concentration (mg.kg^{-1} Soil)							
Ni 0	327.5[a]	1692.5[b]	0.17[c]	1.50[d]	55.37[a]	0.367[a]	1.04[a]
Ni 15	277.5[b]	1873.3[ab]	0.14[d]	2.19[c]	44.50[b]	0.363[b]	0.66[c]
Ni 30	194.2[c]	1699.2[b]	0.18[b]	2.77[a]	37.34[c]	0.362[b]	0.76[b]
Ni 45	139.2[d]	2030.8[a]	0.20[a]	2.30[b]	35.74[c]	0.361[b]	0.30[d]
Pb concentration (mg.kg^{-1} Soil)							
Pb 0	269.2[a]	1837.5[a]	0.18[ab]	3.22[a]	48.70[a]	0.377[a]	0.60[b]
Pb 15	194.2[c]	1738.3[a]	0.14[c]	1.49[c]	43.42[b]	0.370[b]	0.80[a]
Pb 30	236.33[b]	1804.2[a]	0.18[b]	2.63[b]	39.30[c]	0.354[c]	0.78[a]
Pb 45	238.3[b]	1915.8[a]	0.19[a]	1.49[c]	41.52[bc]	0.351[c]	0.58[b]
Interactions between Ni and Pb							
Ni0 Pb0	496.7[a]	1710.0[b]	0.20[d]	3.94[c]	63.35[a]	0.387[a]	1.25[b]
Ni0 Pb15	240.0[d]	1753.3[b]	0.17[e]	1.41[h]	56.96[b]	0.370[b]	0.54[d]
Ni0 Pb30	216.7[d]	1593.3[b]	0.20[d]	0.53[j]	51.23[bc]	0.357[c]	1.48[a]
Ni0 Pb45	356.7[b]	1713.3[b]	0.11[g]	0.13[k]	49.92[c]	0.353[c]	0.87[c]
Ni15 Pb0	303.3[c]	1826.7[ab]	0.14[f]	1.63[g]	49.46[c]	0.370[b]	0.41[de]
Ni15 Pb15	200.0[de]	1960.0[ab]	0.17[e]	2.66[e]	45.88[c]	0.370[b]	0.99[c]
Ni15 Pb30	363.3[b]	1736.7[b]	0.08[h]	1.97[f]	37.69[fg]	0.357[c]	1.18[b]
Ni15 Pb45	243.3[d]	1970.0[ab]	0.17[e]	2.50[e]	44.92[cde]	0.350[c]	0.06[g]
Ni30 Pb0	136.7[f]	1873.3[ab]	0.22[c]	4.41[b]	42.46[def]	0.370[b]	0.31[e]
Ni30 Pb15	210.0[d]	1493.3[b]	0.14[f]	0.94[i]	34.04[g]	0.370[b]	1.17[b]
Ni30 Pb30	216.7[d]	1926.7[ab]	0.21[c]	3.09[d]	34.96[g]	0.353[c]	0.29[ef]
Ni30 Pb45	213.3[d]	1503.3[b]	0.14[f]	2.66[e]	37.89[fg]	0.350[c]	1.28[b]
Ni45 Pb0	140.0[f]	1940.0[ab]	0.17[e]	2.91[d]	39.54[efg]	0.380[a]	0.41[de]
Ni45 Pb15	126.7[f]	1746.7[b]	0.07[h]	0.72[j]	36.81[fg]	0.370[b]	0.51[d]
Ni45 Pb30	150.0[fe]	1960.0[ab]	0.23[b]	4.91[a]	33.27[g]	0.350[c]	0.16[fg]
Ni45 Pb45	140.0[f]	2476.7[a]	0.33[a]	0.66[j]	33.35[g]	0.350[c]	0.10[g]

*Means followed by the same letters with in columns are not significantly different at $P\leq0.05$ according to the Duncan test. Ni: Nickel, Pb: Lead

TABLE II
EFFECTS OF NICKEL, LEAD AND THEIR INTERACTIONS ON SOME ENZYMATIC, NON-ENZYMATIC ANTIOXIDANTS, AND NITRATE REDUCTASE ACTIVITY OF *LEUCAENA LEUCOCEPHALA* PLANT LEAVES

Treatments	Enzymatic Antioxidants and Nitrate Reductase			Non-Enzymatic Antioxidants			
	Catalase (microgram.gram⁻¹)	Peroxidase (microgram.gram⁻¹)	Nitrate Reductase (micromole.L⁻¹)	Ascorbic Acid (gram.L⁻¹)	Proline (microgram.ml⁻¹)	Total Carotenoids (mg.g⁻¹ fresh weight)	Total Carbohydrate (%)
Ni concentration (mg.kg⁻¹ Soil)							
Ni 0	200.83[a]	1366.7[b]	0.17[c]	0.74[d]	75.05[a]	0.377[a]	0.22[c]
Ni 15	133.33[b]	1535.8[b]	0.14[d]	2.93[a]	58.0[d]	0.366[b]	0.23[c]
Ni 30	132.50[b]	1801.7[a]	0.18[b]	1.66[c]	66.55[b]	0.363[b]	0.45[a]
Ni 45	131.67[b]	1989.2[a]	0.20[a]	2.03[b]	60.05[c]	0.363[b]	0.37[b]
Pb concentration (mg.kg⁻¹ Soil)							
Pb 0	195.8[a]	1659.2[a]	0.18[ab]	1.08[c]	63.80[c]	0.368[b]	0.31[b]
Pb 15	150.0[b]	1644.2[a]	0.14[c]	1.05[c]	70.42[a]	0.370[ab]	0.30[b]
Pb 30	131.7[bc]	1644.2[a]	0.18[b]	2.07[b]	60.50[d]	0.372[a]	0.26[b]
Pb 45	120.8[c]	1745.8[a]	0.19[a]	3.16[a]	63.92[b]	0.358[c]	0.40[a]
Interactions between Ni and Pb							
Ni0 Pb0	310.0[a]	1143.3[e]	0.20[d]	0.16[n]	87.04[b]	0.390[a]	0.38[bcd]
Ni0 Pb15	220.0[b]	1520.0[cde]	0.17[e]	0.22[n]	71.15[f]	0.370[bc]	0.06[h]
Ni0 Pb30	133.3[c]	1320.0[de]	0.20[d]	1.84[f]	82.54[d]	0.377[b]	0.19[fgh]
Ni0 Pb45	140.0[c]	1483.3[cde]	0.11[g]	0.72[l]	59.46[h]	0.370[bc]	0.23[efg]
Ni15 Pb0	200.0[b]	1510.0[cde]	0.14[f]	1.03[jk]	31.15[k]	0.360[d]	0.10[gh]
Ni15 Pb15	113.3[cd]	1560.0[cde]	0.17[e]	1.47[h]	84.73[c]	0.370[bc]	0.36[cde]
Ni15 Pb30	133.3[c]	1560.0[cde]	0.08[h]	1.09[ij]	24.23[l]	0.370[bc]	0.06[h]
Ni15 Pb45	86.7[d]	1513.3[cde]	0.17[e]	8.13[a]	91.85[a]	0.363[cd]	0.39[bcd]
Ni30 Pb0	136.7[c]	1836.7[bcd]	0.22[c]	0.34[m]	70.50[f]	0.360[d]	0.23[efg]
Ni30 Pb15	150.0[c]	1960.0[abc]	0.14[f]	1.59[g]	69.73[f]	0.370[bc]	0.50[ab]
Ni30 Pb30	113.3[cd]	1853.3[bcd]	0.21[c]	2.09[e]	48.27[j]	0.370[bc]	0.56[a]
Ni30 Pb45	130.0[cd]	1556.7[cde]	0.14[f]	2.63[d]	77.69[e]	0.350[e]	0.52[ab]
Ni45 Pb0	136.7[c]	2146.7[ab]	0.17[e]	2.78[c]	66.50[g]	0.360[d]	0.51[ab]
Ni45 Pb15	116.7[cd]	1536.7[cde]	0.07[h]	0.91[k]	56.08[i]	0.370[bc]	0.27[def]
Ni45 Pb30	146.7[c]	1843.3[bcd]	0.23[b]	3.25[b]	86.92[b]	0.370[bc]	0.21[fg]
Ni45 Pb45	126.7[cd]	2430.0[a]	0.33[a]	1.16[i]	30.69[k]	0.350[e]	0.48[abc]

*Means followed by the same letters with in columns are not significantly different at P≤0.05 according to the Duncan test. Ni: Nickel, Pb: Lead

antioxidants for *G. triacanthos, L. leucocephala,* and *R. pseudoacacia* species, where the activity of POD and NR were increased with increasing Ni and Pb concentrations for the three species, except NR enzyme activity for *R. pseudoacacia* which decreased with increasing the elements concentration. The highest activity of POD (2030.8, 1989.2, and 2418.3 microgram. gram⁻¹) for the three species, respectively, and (0.20 micromole.l⁻¹) NR activity for each of G. triacanthos and *L. leucocephala* was recorded from 45 mg.kg⁻¹ NiCl₂ or PbCl₂ concentrations, whereas the lowest activities (1692.5, 1366.7, and 1767.5 microgram. gram⁻¹ and 0.17 micromole.l⁻¹) were recorded in the control treatment, same results were obtained from each (Bartkowiak, Lemanowicz, and Lamparski, 2000; Matraszek, 2008). Whereas the activity of CAT enzyme decreased in the three species by increasing the concentrations to 45 mg.kg⁻¹ of Ni and Pb as compared to the control treatment, these results were confirmed by Yan, et al. (2008) and Dey, et al. (2007) who found a decrease in CAT enzyme activity with increasing the concentration of Ni and Pb. The interactions between concentrations of Ni and Pb also had significant effects on all the study parameters for different plants Hussain, et al. (2020) and Andresen, Edgar, and Hendrik (2018). The differences between the species response to Ni and Pb may be due to the uptake of heavy metals from the soil solution depending on species, form and concentration of metal, the soil or nutrient solution acidity and organic matter composition. (Amari, et al., 2017). Pb and Ni are considered as a dangerous pollutant of plants, because they affect many biological processes such as photosynthesis, carbohydrate synthesis, and cell wall by decrease of exchange between inner and outer. (Taha, et al., 2008). Plant exposure to higher concentrations of heavy metals can increase the production of (ROS) and change antioxidant response (Shahid, et al., 2015; Gratão et al., 2005). Antioxidants are the first line of defense against the damages caused by free radicals and are critical for the optimum health of plant cells. Plant antioxidants play a significant role in assisting plant development through a wide variety of mechanisms and functions (Rajput, et al., 2021). In fact, these enzymes are key components in preventing the oxidative stress in plants as the activity of one or more of these enzymes are generally increased in plants when exposed to stressful conditions (Das, et al., 2014). In recent years, a new role for ROS has been identified: the control and regulation of biological processes, such as growth and development. The use of ROS as signaling molecules by plant cells suggests that, during the course of evolution, plants were able to achieve a high degree of control over ROS toxicity and are now using ROS as signaling molecules. (Mittler, et al., 2004).

TABLE III
EFFECTS OF NICKEL, LEAD, AND THEIR INTERACTIONS ON SOME ENZYMATIC, NON-ENZYMATIC ANTIOXIDANTS, AND NITRATE REDUCTASE ACTIVITY OF *ROBINIA PSEUDOACACIA* PLANT LEAVES

Treatments	Enzymatic Antioxidants and Nitrate Reductase			Non-Enzymatic Antioxidants			
	Catalase (microgram. gram^{-1})	Peroxidase (microgram. gram^{-1})	Nitrate Reductase (micromole.L^{-1})	Ascorbic Acid (gram. L^{-1})	Proline (microgram.ml^{-1})	Total Carotenoids (mg. g^{-1} fresh weight)	Total Carbohydrate (%)
Ni concentration (mg.kg^{-1} Soil)							
Ni 0	285.83[a]	1767.5[b]	0.127[a]	0.77[c]	51.52[a]	0.380[a]	1.22[b]
Ni 15	255.00[b]	2291.7[a]	0.112[b]	0.91[b]	46.82[b]	0.370[b]	1.33[a]
Ni 30	174.17[c]	2390.0[a]	0.124[a]	1.31[a]	45.14[c]	0.370[b]	0.77[d]
Ni 45	150.00[d]	2418.3[a]	0.093[c]	0.35[d]	40.75[d]	0.370[b]	0.93[c]
Pb concentration (mg.kg^{-1} Soil)							
Pb 0	226.7[a]	1914.2[b]	0.102[c]	1.38[a]	58.94[a]	0.373[a]	0.94[b]
Pb 15	209.2[a]	2084.2[b]	0.143[a]	0.89[b]	43.58[b]	0.373[a]	1.15[a]
Pb 30	208.3[a]	2688.3[a]	0.099[c]	0.52[c]	38.25[c]	0.373[a]	1.23[a]
Pb 45	220.8[a]	2180.8[b]	0.112[b]	0.54[c]	43.46[b]	0.373[a]	0.93[b]
Interactions between Ni and Pb							
Ni0 Pb0	360.0[a]	1543.3[d]	0.083[i]	2.16[b]	49.38[ef]	0.380[a]	1.70[a]
Ni0 Pb15	250.0[b]	1626.7[d]	0.170[b]	0.28[hi]	49.31[ef]	0.380[a]	0.97[cde]
Ni0 Pb30	276.7[b]	2090.0[bcd]	0.126[de]	0.53[f]	48.30[f]	0.380[a]	1.18[b]
Ni0 Pb45	256.7[b]	1810.0[cd]	0.130[cd]	0.09[i]	58.85[c]	0.380[a]	1.03[bcde]
Ni15 Pb0	270.0[b]	1583.3[d]	0.120[e]	0.19[hi]	63.31[b]	0.370[b]	0.94[de]
Ni15 Pb15	246.7[b]	2150.0[bcd]	0.130[cd]	0.81[e]	39.84[h]	0.370[b]	1.15[bc]
Ni15 Pb30	250.0[b]	3243.3[a]	0.060[j]	1.00[d]	41.46[h]	0.370[b]	1.58[a]
Ni15 Pb45	253.3[b]	2190.0[bcd]	0.136[c]	1.63[c]	35.96[i]	0.370[b]	1.64[a]
Ni30 Pb0	156.7[cd]	2146.7[bcd]	0.106[fg]	2.91[a]	51.24[de]	0.370[b]	0.45[g]
Ni30 Pb15	186.7[c]	2736.7[ab]	0.180[a]	1.53[c]	43.54[g]	0.370[b]	0.89[e]
Ni30 Pb30	153.3[cd]	2620.0[abc]	0.110[f]	0.47[fg]	40.88[h]	0.370[b]	1.09[bcd]
Ni30 Pb45	200.0[c]	2056.7[bcd]	0.100[gh]	0.34[gh]	51.62[d]	0.370[b]	0.64[f]
Ni45 Pb0	120.0[d]	2383.3[bcd]	0.100[fgh]	0.25[hi]	71.69[a]	0.370[b]	0.68[f]
Ni45 Pb15	153.3[cd]	1823.3[cd]	0.090[h]	0.97[de]	41.54[h]	0.370[b]	1.58[a]
Ni45 Pb30	153.3[cd]	2800. 0[ab]	0.100[gh]	0.093[i]	22.35[k]	0.370[b]	1.06[bcde]
Ni45 Pb45	173.3[c]	2666.7[abc]	0.080[i]	0.093[i]	27.42[j]	0.370[b]	0.41[g]

*Means followed by the same letters within columns are not significantly different at *P*≤0.05 according to the Duncan test. Ni: Nickel, Pb: Lead

Results of the study showed in Tables I-III the effects of Ni and Pb and their interactions on some non-enzymatic antioxidants for G. triacanthos, *L. leucocephala,* and *R. pseudoacacia* leaves. It shows that proline (Pr) and TC were decreased with increasing the concentration of Ni and Pb for the three species. TCHO decreased with increasing concentration for only G. triacanthos *and R. pseudoacacia,* but not for *L. leucocephala where* the TCHO increased with increased the elements concentration, same results were confirmed by Mame and Al-Rashed (2021); Singh, et al. (2012) and Tzvetkova and Kolarov (1996). The AA content increased with increasing the concentration of Ni and Pb for G. triacanthos *and L. leucocephala* but not for *R. pseudoacacia where it* decreased with increasing the elements concentrations (Bielen, et al., 2013; Gad, El-Sherif, and El-Gereedly, 2007). However, at high concentrations it is actually harmful to plants, and leads to programmed cell death (Gill and Tuteja, 2010), Ni and Pb may accumulate at such low concentrations in the leaves, so it was not detected. In this work, we assumed that each of *G. triacanthos*, *L. leucocephala,* and *R. pseudoacacia* plants may be better adapted and are tolerant to heavy metal stress conditions. However, highest concentration of Ni and pb reduced photosynthetic attributes, decreased pigment contents, the activity of non-enzymatic antioxidants AA, Pr, TC, and TCHO, in association with relatively high concentrations

of Ni and Pb, where stresses increase the amount of ROS, thereby changes the formation and imbalance the contents of photosynthetic pigments, and non-enzymatic antioxidants. Plants increase antioxidant defense mechanisms under abiotic stresses such as drought, excessive watering, extreme temperatures (cold, frost, and heat), salinity, UV-B radiation and mineral toxicity, to alleviate oxidative damage. Different responses of antioxidants are stimulated to avoid ROS damages to cellular constituents, as well as to sustain normal growth and development. Excess levels of ROS are damaging to the plant; thus, to remain the balance of cellular redox, both non-enzymatic and enzymatic systems are motivated to repair the toxic levels of ROS. The balance between the producing and eliminating the ROS is persistent by enzymatic and non-enzymatic antioxidants. In plants, the appearance of many antioxidant enzymes is correlated positively with high levels of tolerance against the abiotic stresses, so, the increase in the activities of enzymes are closely related to the decrease in oxidative damage. The activation of some enzymes leads to plant protection against oxidative damage. Thus, in plants a complex enzymatic system has developed to scavenge extra ROS also, to protect them from the oxidative stress (Gull, Lone, and Wani, 2019; Caverzan, Casassola, and Brammer, 2016)

IV. Conclusions and Recommendations

The contamination of soil with heavy metals is a serious problem that affects the physiological process in plants including antioxidants defense systems. From this study, it is concluded that each of *G. triacanthos*, *L. leucocephala,* and *R. pseudoacacia* plant species were a response to soil increase in Ni and Pb heavy metals concentration like each other regarding the decrease in the CAT enzyme activity and each of Pr, and TC content, also they sharing the increase in POD activity, whereas *R. pseudoacacia* species unlike the two other species in decreasing the activity of NR enzyme and AA content with the increase in Ni and Pb concentrations. However, *L. leucocephala* is unlike the two other species in increasing CHO content with increasing the concentration of Ni and Pb elements. On the other hand, this study indicated that *L. leucocephala* was revealed to be more tolerant to Ni and Pb stress than other plant species. It can be recommended to conduct genetic studies regarding the responses to heavy metal stress to understand the differences between the same family species at the gene level.

References

Abdullah, K.H., Ahmed, G.B., Selah-Alden, M.T., Hassan, H.N., Mahmood, M.J., Hameed, N.A. and Amin, S.M., 2018. Overcoming seed dormancy of *Robinia pseudoacacia* L. and *Ceratonia siliqua* L. species using different pretreatments in Malta forest nursery-Duhok. *Journal of University of Duhok*, 21(1), pp.1-7.

Aebi, H.E., 1974. Catalase. In: Methods of Enzymatic Analysis. Vol. 2. Academic Press Inc., United States, pp.673-684.

Ahmad, S., Fazili1, I.S., Haque, R., Khan, S.N. and Abdin, M.Z., 2010. Standardization and estimation of nitrate reductase activity in the leaves of *Ammi majus L.* (Bishops weed) in relation to sulphur deficiency and seed yield. *Australian Journal of Crop Science*, 4(7), pp.515-522.

Ali, S., Hayat, K., Iqbal, A. and Xie, L., (2020). Implications of abscisic acid in the drought stress tolerance of plants. *Agronomy*, 10, p.1323.

Al-Mohammadi, S.H. and Al-Mohammadi, F., (2002). "*Statistics and Experimental Design*". Dar Osama for Publishing and Distribution, Amman, Jordan. pp.375.

Álvarez, S.P., Cabezas-Montero, D., Debora-Duarte, B.N., Tapia, M.A., Sida-Arreola, J.P., Sánchez, E. and Héctor-Ardisana, E.F., 2019. Used response to antioxidative enzymes in rice under stress due to lead and nickel. *Revista Mexicana de Ciencias Agrícolas*, 10(1), pp.51-62.

Amari, T., Ghnaya, T. and Abdelly, C., 2017. Nickel, cadmium and lead phytotoxicity and potential of halophytic plants in heavy metal extraction. *South African Journal of Botany*, 111, pp.99-110.

Andresen, E., Edgar, P. and Hendrik, K., 2018. Trace metal metabolism in plants. *Journal of Experimental Botany*, 69(5), pp.909-954.

Bartkowiak, A., Lemanowicz, J. and Lamparski, R., 2020. Assessment of selected heavy metals and enzyme activity in soils within the zone of influence of various tree species. *Scientific Reports*, 10, p.14077.

Bates, L.S., Waldron, R.P. and Tears, I.D., 1973. Rapid determination of free proline for water stress studies. *Plant and Soil*, 39, pp.205-207.

Bielen, A., Remans, T., Vangronsveld, J. and Cuypers, A., 2013. The influence of metal stress on the availability and redox state of ascorbate, and possible interference with its cellular functions. *International Journal of Molecular Sciences*, 14, pp.6382-6413.

Caverzan, A., Casassola, A. and Brammer, S.P., 2016. Antioxidant responses of

wheat plants under stress. *Genetics and Molecular Biology*, 39(1):1-6.

Das, K. and Roychoudhury, A., 2014. Reactive oxygen species (ROS) and response of antioxidants as ROS-scavengers during environmental stress in plants. *Frontiers in Environmental Science*, 2, p.53.

Dey,S.K., Dey, J., Patra, S. and Pothal, D., 2007. Changes in the antioxidative enzymes activities and lipid peroxidation in wheat seedlings exposed to cadmium and lead stress. *Brazilian Journal of Plant Physiology*, 19(1), pp.53-60.

Dirr, M.A. and Heuser, C.W., 1987. *The Reference Manual of Woody Plant Propagation*. Varsity Press Inc., Athens, GA. p.239.

Elbsheer, S.E., 2018. *Development of Methods for the Determination of Vitamin C Content in Some Dry Fruits and Leaves, Laloub and Mesquite* (*Balanites aegyptiaca and Prosopis juliflora*). Sudan University of Science and Technology College of Graduate Studies, Sudan. (PhD Thesis). Available from: https://repository.sustech.edu/handle/123456789/22634 [Last accessed on 2022 Jul 01].

Gad, N., El-Sherif, M.H. and El-Gereedly, N.H., 2007. Influence of nickel on some physiological aspects of tomato plants. *Australian Journal of Basic and Applied Sciences*, 1(3), pp.286-293.

Gill, S.S. and Tuteja, N., 2010. Reactive oxygen species and antioxidant machinery in abiotic stress tolerance in crop plants. *Plant Physiology and Biochemistry*, 48, pp.909-930.

Gull, A., Lone, A.A. and Wani, N.U., 2019. Biotic and abiotic stresses in plants. In: *Abiotic and Biotic Stress in Plants*, Intechopen, London, UK.

Hussain, S., Khaliq, A., Noor, M.A., Tanveer, M., Hussain, H.A., Hussain, S., Shah, T. and Mehmood, T., 2020. *Metal Toxicity and Nitrogen Metabolism in Plants: An Overview*. Springer Nature, Singapore. Available from: https://doi.org/10.1007/978-981-13-7264-3_7 [Last accessed on 2022 Jul 01].

Joslyn, M.A., 1970. *Analítico: Methods in Food Analysis. Physical, Chemical, and Instrumental Methods of Analysis*, 2nd ed. Academic Press, New York. p.844.

Kacálková, L., Tlustoš, P. and Száková, J., 2014. Chromium, nickel, cadmium, and lead accumulation in maize, sunflower, willow, and poplar. *Polish Journal of Environmental Studies*, 23(3), pp.753-761.

Kartoori, S., Sumalatha, G.M., Shuba, A.C., Komala, N.T. and Biradar Patil, N.K., 2018. Role of enzymatic antioxidants defense system in seeds. *International Journal of Current Microbiology and Applied Sciences*, 7, pp.584-594.

Kumari, S. and Mishra, A., 2021. *Heavy Metal Contamination. Life sciences*. Intechopen, London, UK. Available from: https://www.intechopen.com/chapters/72968 [Last accessed on 2022 Jul 01]. doi: 10.5772/intechopen.93412

Lichtenthaler, H.K. and Wellburn, A.R., 1983. Determinations of total carotenoids and chlorophylls a and b of leaf extracts in different solvents. *Biochemical Society Trtansaction*, 11(5), pp.591-592.

Mame, S.O. and Ali Al-Rashed, H.S., 2021. Impact of ascorbic acid on the concentration of Carotene and Prolin in the legume (*Vicia faba*) Plants grown with Soil treated with Pb and Ni. *Ecology, Environment and Conservation*, 27, pp.S169-S172.

Matraszek, R., 2008. Nitrate reductase activity of two leafy vegetables as affected by nickel and different nitrogen forms. *Acta Physiologiae Plantarum*, 30(3), p.361-370.

Mittler, R., Vanderauwera, S., Gollery, M. and van Breusegem, F., 2004. Reactive oxygen gene network of plants. *Trends in Plant Science*, 9(10), pp.490-498.

Nezih, M., 1985. The peroxidase enzyme activity of some vegetables and its resistance to heat. *Journal of the Science of Food and Agriculture*, 36, pp.877-880.

Nyiramigisha, P., Komariah. and Sajidan., (2021). Harmful impacts of heavy metal contamination in the soil and crops grown around dumpsites. *Agricultural Science*, 9, pp.271-282.

Rajput, V.D., Harish, Singh, R.K., Verma, K.K., Sharma, L., Quiroz-Figueroa, F.R., Meena, M., Gour, V.S., Minkina, T., Sushkova, S. and Mandzhieva, S., 2021. Recent developments in enzymatic antioxidant defence mechanism in

plants with special reference to abiotic stress. *Biology*, 10, p.267.

Ren, R., Li, Z., Zhang, L., Zhou, H., Jiang, X. and Liu, Y., 2021. Enzymatic and non-enzymatic antioxidant systems impact the viability of cryopreserved *Paeonia suffruticosa* pollen. *Plant Cell, Tissue and Organ Culture (PCTOC)*, 144, pp.233-246.

Sachdev, S., Ansari, S.K., Ansari, M.I., Fujita, M. and Hasanuzzaman, M., 2021. Abiotic stress and reactive oxygen species: Generation, signaling, and defense mechanisms. *Journal Antioxidants*, 10, p.277.

Santos-Sánchez, N.F., Salas-Coronado, R., Villanueva-Cañongo, C. and Hernández-Carlos, B., 2019. Antioxidant compounds and their antioxidant mechanism. In: Shalaby E. (editor), Antioxidants, IntechOpen, London. DOI: 10.5772/intechopen.85270

Shahid, M., Khalid, S., Abbas, G., Shahid, N., Nadeem, M., Sabir, M., Aslam, M. and Dumat, C., 2015. Heavy metal stress and crop productivity. In: *Crop Production and Global Environmental Issues.* Springer International Publisher, Switzerland. Available from: https://doi.org/10.1007/978-3-319-23162-4 [Last accessed on 2022 Jul 01].

Singh, G., Agnihotri, R.K., Reshma, R.S. and Ahmad, M., 2012. Effect of lead and nickel toxicity on chlorophyll and proline content of Urd (*Vigna mungo* L.)

seedlings. *International Journal of Plant Physiology and Biochemistry,* 4(6), pp.136-141.

Ssenku, J.E., Ntale, M., Backeus, I. and Oryem-Origa, H., 2017. Hytoremediation potential of *Leucaena leucocephala* (Lam.) de Wit. for heavy metal-polluted and heavy metal-degraded environments. In: *Phytoremediation Potential of Bioenergy Plants.* Springer, Singapore. pp.189-209.

Taha, D.N., AL-Kuramy, A.A. and Al-Khalaf, A.K., 2008. Determination of lead, nickel and zinc in plants of North AL-Diwaniya City. *Journal of Kerbala University*, 6(4), p.58.

Tzvetkova, N. and Kolarov, D., 1996. Effect of air pollution on carbohydrate and nutrients concentrations in some decieduous tree species. *Bulgarian Journal of Plant Physiology*, 22(1-2), pp.53-63.

Vatansever, R., Ozyigit, I.I. and Filiz, E., 2017. Essential and beneficial trace elementsin plants, and their transport in roots. *Biotechnology and Applied Biochemistry*, 181(1), pp.464-482. Doi: 10.1007/s12010-016-2224-3

Yan, R., Gao, S., Yang, W., Cao, M., Wang, S. and Chen, F., 2008. Nickel toxicity induced antioxidant enzyme and phenylalanine ammonia-lyase activities in *Jatropha curcas* L. cotyledons. *Plant, Soil and Environment*, 54(7), pp.294-300.

Effect of Substrate Temperature on the Electrical Properties of Al-doped Zinc Oxide Films Deposited on Polyethylene Terephthalate

Mohammad G. Faraj

Department of Physics, Faculty of Science and Health, Koya University, Koya KOY45,
Kurdistan Region - Iraq

Abstract—To prepare homogeneous thin films of zinc oxide (ZnO) doped with aluminum (Al) on a polyethylene terephthalate (PET) substrate at different temperatures (200–250°C), the process is carried out by utilizing the chemical spraying pyrolysis approach. A study of the effects of substrate temperature on the Al-doped Zinc Oxide (AZO) films' electrical characteristics and roughness is performed. The measurements of atomic force microscopy show that the root mean square roughness of the AZO films is increased with the increase of PET substrate temperature. Hall measurements show that the electrical resistivity decreases as the substrate temperature increases. On the increment of substrate temperature, there is an increase in the carrier concentration value from 9.98×10^{19} to 5.4×10^{20} cm^{-3} and an increase in the carrier mobility value from 5.5 to 9.76 cm^2.(V. S)$^{-1}$.

Index Terms—Al-doped zinc oxide, Chemical spray pyrolysis, Electrical properties, Hall measurements, Polyethylene terephthalate.

I. Introduction

Zinc oxide (ZnO) is an important semiconductor material due to its wide applications in many different fields. Recently, ZnO has received an increasing attention by the researchers. At room temperature, ZnO is an n-type semiconductor with a direct band gap and high energy exciton bonds (Faraj and Ibrahim, 2011). ZnO has many applications because it is cheap, stable, readily available, and secure. These properties make it one of the best materials to be used in optoelectronics. Several research works have been previously reported to study ZnO (Faraj and Taboada, 2017; Faraj and Taboada, 2017; Faraj and Esia, 2019).

Post deposition treatments can easily be used to tune the physical properties of ZnO, such as the incorporation of gallium (Ga), tin (Sn), and aluminum (Al) (Park, Ikegami and Ebihara, 2006; Tynell, et al., 2013).

Researchers have employed gallium, aluminum, and indium as dopants to create n-type ZnO films with excellent quality and conductivity (Myong, et al., 1997; Assuncao, et al., 2003). Al doped ZnO films is considered to be one of the best elements for optoelectronics due to its excellent electrical conductivity (Gong, Lu and Ye, 2010; Raghu, et al., 2017). Several methods were used in the literature to prepare Al-doped Zinc Oxide (AZO) thin films, such as magnetron sputtering, atomic layer deposition, and spray pyrolysis (Zhou, et al., 2007; Kaurn, Mitra, and Yadav, 2015; Ayouchi, et al., 2003).

Plastic substrates are being used more and more in optoelectronics applications, particularly for energy-producing devices like solar cells (Ryu, Cha and Jung, 2010), as they can offer benefits including light weight, high shock resistance, and scalable roll-to-roll preparation processes. Due to their flexibility, light weight, and low cost, AZO films on flexible polymeric substrates as polyethylene terephthalate (PET) for such intended applications are in particular attracting a lot of attention (Jeong, et al., 2010). New to the ongoing work is the preparation and characterization of AZO films fabricated on PET plastic substrates with a chemical spraying pyrolysis technique at different temperatures (200–250°C).

In this study, AZO thin films are deposited on PET plastic substrates by chemical spray pyrolysis at different substrate temperatures (200°C, 225°C, and 250°C) with steps of 25°C. Substrate temperature influences the roughness and electrical characteristics of the AZO films.

II. Experimental Procedure

A. Materials

In this experiment, PET plastic substrates were acquired from Penfibre Sdn. Bhd., company. Materials used in the preparation of AZO thin films include Zinc nitrate (Zn(NO$_3$)$_2$.6H$_2$O; 99.0%) and aluminum nitrate Al(NO$_3$)$_3$.9H$_2$O; 99.9%) were obtained from Sigma-Aldrich.

B. Deposition of AZO Thin Film

AZO thin films were preserved on PET plastic substrates using a chemical spray pyrolysis process. The PET plastic

substrates were immediately washed with ethanol for 10 min to remove contamination. After cleaning, purified water is used to rinse the substrate (DI water). After that, the sample was dried using nitrogen (N_2) gas. By dissolving 0.1 M watery arrangement of zinc nitrate in DI water, the antecedent arrangement (100 mL) was created. Then, 0.2 mL of nitric corrosive (HNO_3) was added to the arrangement to increase the zinc nitrate's solvency. The final mixture was vigorously mixed for 30 min at 25 °C with an appealing stirrer.

The precursor arrangement was split onto PET plastic substrates at a temperature of preserved ZnO thin films (200°C, 225°C, and 250°C). To begin the investigation into aluminum (Al) doping, aluminum nitrate in various groups of 3% was introduced. To obtain homogeneous thin films, the height of the splashing spout and the rate process were kept constant throughout the testimony operation, at 27 cm and 5 ml/min, respectively.

The arrangement was atomized using packed nitrogen at a weight of 1 bar. Using an optical reflectometer (Model: Filmetrics F20), the estimated thickness of the samples was 300 nm. Using the Hall measurement (HL5500PC) system, hall effect measurement was carried out to determine the electrical properties of AZO thin films, including resistivity, carrier type, mobility and concentration.

Hall mobility (μ_H) is calculated using the relation shown below (Bube; 2003):

$$\mu_H = \sigma. R_H \tag{1}$$

Where σ is the electrical conductivity at room temperature and R_H is the Hall coefficient, whereas the carrier concentration (n) can be determined using the relation (Bube; 2003):

$$n = \frac{1}{R_H.e} \tag{2}$$

Where e is the electron charge.

Atomic force microscopy (AFM) (model: ULTRA1Objective model) was used to examine the surface morphology of the AZO films). The temperate of the all samples were measured by electric hot Plate (model: Onilab). The melting point of PET plastic is 254°C (Faraj, Ibrahim and Ali, 2011), making the temperatures used in this study for AZO film deposition on PET plastic substrates appropriate.

III. Results and Discussion

Fig. 1 presents AFM images of the surface morphologies of AZO deposited on PET plastic substrates at different substrate temperatures. The surfaces of the AZO thin film products were smooth. For films with substrate temperatures of 200°C, 225°C, and 250°C, the estimated root mean square (RMS) surface roughness of the AZO samples was 12.3 nm, 19.64 nm, and 38.87 nm, respectively. As the substrate temperature rises, the RMS surface roughness also rises. Surface roughness was always brought on by the conical features that could be readily seen on the sample surface. It is crucial to remember that surface smoothness is a highly desired property for coatings used in optical applications since it lowers the reflection loss caused by surface scattering caused by roughness. This behavior is similar to earlier reports (Ghorannevis, et al., 2015; Faraj, et al., 2011).

Hall measurements were used to measure the electrical parameters. According to the electrical analysis, the films degenerate and exhibit n-type electrical conductivity. Fig. 2 shows the electrical resistivity of the AZO films as a function of substrate temperature. The film resistivity decreases dramatically as the substrate temperature is increased from 200°C to 250°C in 25°C steps. The evaluated film resistivity's were 18.54×10^{-2}, 3.2×10^{-2}, and 1.14×10^{-2} (Ω.cm) at different temperatures (200°C, 225°C, and 250°C). The Hall mobility and carrier concentration of the AZO films depend on the substrate temperature, as shown in Table 1. As the substrate temperature rises, the Hall measurement reveals an increase in the carrier concentration value from $9.98 \ 10^{19}$ to $5.4 \ 10^{20}$ cm^3 with a mobility value in the range of 5.5–9.76 cm^2 $(V. S)^{-1}$, as shown in Table 1.

The increased mobility of the charge carriers most likely causes the decrease in resistivity. This behavior agrees with that reported in the literature (Hu, et al., 2014; Wang, Li and Zhang, 2009; Zhao, et al., 2020). Since all samples exhibit low resistance and high carrier concentration, transparent front contacts for solar cells are possible.

TABLE I
THE DEPENDENCE OF THE HALL MOBILITY AND CARRIER CONCENTRATION OF THE AL-DOPED ZINC OXIDE FILMS ON THE SUBSTRATE TEMPERATURE

No.	T (C)	Hall mobility (cm²/V.S)	Carrier concentration (cm⁻³)
1	200	5.5	9.98×10^{19}
2	225	8.3	1.6×10^{20}
3	250	9.76	5.4×10^{20}
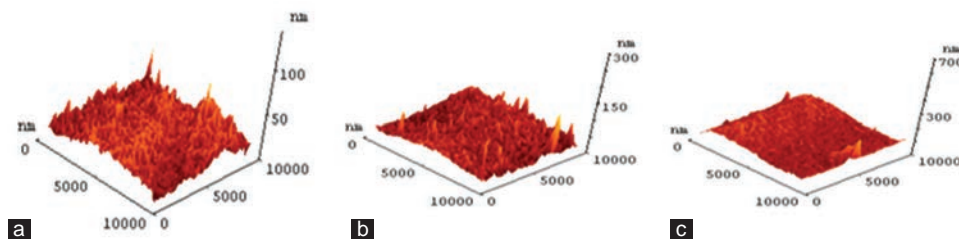
Fig. 1. Atomic force microscopy surface images of Al-doped zinc oxide films as a function of substrate temperature: (a) 200°C, (b) 225°C and (c) 250°C.
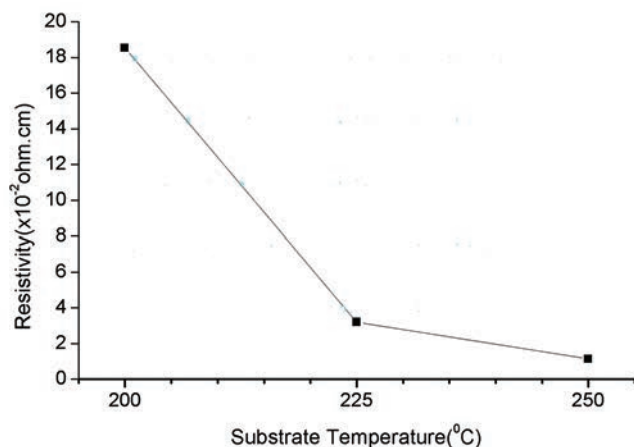
Fig. 2. Resistivity of Al-doped zinc oxide films as a function of substrate temperature.

IV. Conclusions

Chemical spray pyrolysis was used to deposit zinc oxide-doped aluminum (AZO) thin films at various substrate temperatures on PET plastic substrates. AFM and Hall measurements were used to examine the properties of AZO thin films. The effects of the substrate temperature on the roughness and electrical properties of the AZO films were studied. From AFM images, it was found that the RMS roughness of the AZO film surface increased as the substrate temperature increased. From the Hall measurements, it was found that AZO films with different substrate temperatures always exhibit n-type conductivity. From the investigations, it can be concluded that AZO thin film PET plastic substrates would be acceptable for use as transparent front contacts in thin film solar cells.

References

Assuncao, V., Fortunato, E., Marques, A., Aguas, H., Ferreira, I., Costa, M.E. and Martins, R., 2003. Influence of the deposition pressure on the properties of transparent and conductive ZnO: Ga thin-film produced by rf. sputtering at room temperature. *Thin Solid Films*, 427, pp.401-405.

Ayouchi, R., Leinen, D., Martín, F., Gabas, M., Dalchiele, E. and Ramos-Barrado, J.R., 2003. Preparation and characterization of transparent ZnO thin films obtained by spray pyrolysis. *Thin Solid Films*, 426, pp.68-77.

Bube, R.H., 2001. *Encyclopedia of Physical Science and Technology*. 3rd ed. Academic Press, USA.

Faraj, M.G. and Esia, M.H., 2019. Effect of polyimide substrate on the physical properties of Aluminum doped Zinc Oxide (AZO) thin films deposited by spray pyrolysis technique. *Digest Journal of Nanomaterials and Biostructures,* 14, pp.471-478.

Faraj, M.G. and Ibrahim, K., 2011. Optical and structural properties of thermally evaporated zinc oxide thin films on polyethylene terephthalate substrates. *International Journal of Polymer Science*, 2011, p4.

Faraj, M.G. and Taboada, P., 2017. Comparative studies of the properties of ZnO sprayed thin films on different polymer substrates for flexible solar cell applications. *Journal of Inorganic and Organometallic Polymers and Materials*, 27, pp.1405-1411.

Faraj, M.G. and Taboada, P., 2017. Structural and optical properties of ZnO thin films prepared by spray pyrolysis on PI plastic substrates at various temperatures for integration in solar cell. *Journal of Materials Science: Materials in Electronics*, 28, pp.16504-16508.

Faraj, M.G., Ibrahim, K. and Ali, M.K., 2011. PET as a plastic substrate for the flexible optoelectronic applications. *Optoelectronics and Advanced Materials Rapid Communications*, 5, pp.879-882.

Faraj, M.G., Ibrahim, K., Esia, M.H., Pakhuruddin, M.K. and Pakhuruddin, M.Z., 2011. Comparison of Zinc Oxide thin films deposited on the glass and polyethylene terephthalate substrates by thermal evaporation technique for applications in solar cells. *Optoelectronics and Advanced Materials Rapid Communications*, 4, pp.1587-1590.

Ghorannevis, Z., Hosseinnejad, M.T., Habibi, M. and Golmahdi, P., 2015. Effect of substrate temperature on structural, morphological and optical properties of deposited Al/ZnO films. *Journal of Theoretical and Applied Physics,* 9, pp.33-38.

Gong, L., Lu, L. and Ye, Z., 2010. Transparent and conductive Ga-doped ZnO films grown by RF magnetron sputtering on polycarbonate substrates. *Solar Energy Materials and Solar Cells*, 94, pp.937-941.

Hu, C.C., Lu, T.W., Chou, C.Y., Wang, J.T. and Hsu, C.Y., 2014. Optimization of AZO films prepared on flexible substrates. *Bulletin*, 37, pp.1275-1282.

Jeong, A.J., Shin, H.S., Choi, K.H. and Kim, H.K., 2010. Flexible Al-doped ZnO films grown on PET substrates using linear facing target sputtering for flexible OLEDs. *Journal of Physics D: Applied Physics*, 43, p.465403.

Kaurn, G., Mitra, A. and Yadav, K.L., 2015. Pulsed laser deposited Al-doped ZnO thin films for optical applications. *Progress in Natural Science: Materials International,* 25, pp.12-21.

Myong, S.Y., Baik, S.J., Lee, C.H., Cho, W.Y. and Lim, K.S., 1997. Extremely transparent and conductive ZnO: Al thin films prepared by photo-assisted metalorganic chemical vapor deposition (photo-MOCVD) using $AlCl_3(6H_2O)$ as new doping material. *Japanese Journal of Applied Physics*, 36, p. L1078.

Park, S.M., Ikegami, T. and Ebihara, K., 2006. Effects of substrate temperature on the properties of Ga-doped ZnO by pulsed laser deposition. *Thin Solid Films*, 513, pp.90-94.

Raghu, P., Srinatha, N., Naveen, C.S., Mahesh, M.H. and Angadi, B., 2017. Investigation on the effect of Al concentration on the structural, optical and electrical properties of spin coated Al: ZnO thin films. *Journal of Alloys and Compounds,* 694, pp.68-75.

Ryu, M.S., Cha, H.J. and Jang, J., 2010. Improvement of operation lifetime for conjugated polymer: Fullerene organic solar cells by introducing a UV absorbing film. *Solar Energy Materials and Solar Cells*, 94, pp.152-156.

Tynell, T., Yamauchi, H., Karppinen, M., Okazaki, R. and Terasaki, I., 2013. Atomic layer deposition of Al-doped ZnO thin films. *Journal of Vacuum Science and Technology*, A31, p.01A109.

Wang, S., Li, X. and Zhang, J., 2009. Effects of substrate temperature on the properties of heavy Ga-doped ZnO transparent conductive film by RF magnetron sputtering. *Journal of Physics: Conference Series*, 188, p.012017.

Zhao, M.J., Sun, Z.T., Hsu, C.H., Huang, P.H., Zhang, X.Y., Wu, W.Y., Gao, P., Qiu, Y., Lien, S.Y. and Zhu, W.Z., 2020. Zinc Oxide films with high transparency and crystallinity prepared by a low temperature spatial atomic layer deposition process. *Nanomaterials(Basel)*, 10, p.459.

Zhou, H.M., Yi, D.Q., Yu, Z.M., Xiao, L.R. and Li, J., 2007. Preparation of aluminum doped zinc oxide films and the study of their microstructure, electrical and optical properties. *Thin Solid Films*, 515, pp.6909-6914.

Employing Neural Style Transfer for Generating Deep Dream Images

Lafta R. Al-Khazraji[1,2], Ayad R. Abbas[1], and Abeer S. Jamil[3]

[1]Department of Computer Science, University of Technology-Iraq,
Baghdad, Iraq

[2]General Directorate of Education of Salahuddin Governorate,
Iraq

[3]Department of Computer Technology Engineering, Al-Mansour University College,
Baghdad, Iraq

Abstract—**In recent years, deep dream and neural style transfer emerged as hot topics in deep learning. Hence, mixing those two techniques support the art and enhance the images that simulate hallucinations among psychiatric patients and drug addicts. In this study, our model combines deep dream and neural style transfer (NST) to produce a new image that combines the two technologies. VGG-19 and Inception v3 pre-trained networks are used for NST and deep dream, respectively. Gram matrix is a vital process for style transfer. The loss is minimized in style transfer while maximized in a deep dream using gradient descent for the first case and gradient ascent for the second. We found that different image produces different loss values depending on the degree of clarity of that images. Distorted images have higher loss values in NST and lower loss values with deep dreams. The opposite happened for the clear images that did not contain mixed lines, circles, colors, or other shapes.**

Index Terms—**Deep dream, Gradient ascent, Gram matrix, Neural style transfer.**

I. Introduction

Deep learning (DL) has the ability to achieve complicated cognitive tasks exceeding the performance of humans. Because the performance of DL algorithms, such as deep CNN, gives great results compared to other machine learning machine learning (Abedi, et al., 2020; Alzubaidi, et al., 2021).

The deep dream had been developed as the newest DL technology that was produced by Google, which is a technique that works to improve the visual attributes of images (Mordvintsev, et al., 2015).

The deep dream is created through the repeating feeds of the image to the CNN model, where first, the low-level features (i.e., edges and lines) are detected by the first layer. After that, high-level features such as faces and trees appeared. In the end, all those features are collected for configuring combined effects such as trees or the whole structure (Khan, et al., 2020).

Neural style transfer (NST) is a technology that works under the umbrella of deep learning and is considered one of the attractive deep learning applications (Singh, et al., 2021).

In NST, the input consists of two images; a content image and a style image, whereas there is only one output image that combines the contents of the content images with the style of the style image (Li, 2018).

II. Motivations Behind the Study

Many people have a great passion for drawing, but unfortunately, they lack the skill to do it, and this is what is known today as NST. In recent times, deep dream emerged as a hot topic and are used in many important fields as a simulation tool; simulating hallucinations among psychiatric patients and drug addicts have become a very atheistic necessity to enable doctors and those responsible for psychiatric clinics to imagine a vision approaching the reality that these patients see.

Many cases combine hallucinations with seeing artistic images in a certain style, and therefore, this study came to combine the use of these two techniques.

III. Related Works

This section presents some of the most closely related works to NST and deep dream, starting with NST.

A. NST

Chen, et al. (2020) proposed a lightweight network to enable video style transfer by a knowledge distillation model. Two teacher networks were used; the former was used during the inference by taking the optical ow, whereas the latter did not. The output variation between those two

networks indicates that optical ow had improved, which is then embraced to distill the target student network. They employed the low degree distillation loss to constrain the student network output, which is stylized videos to simulate the low degree of the input video. They measured their work quantitatively and qualitatively; in the former, they compared their method with style Candy from other databases by calculating the temporal error and employing it for calculating the temporal consistency. They got the minimum error compared to the other methods. In the qualitative analysis, also, their method got higher temporal consistency.

Kotovenko, et al. (2019) presented a model based on a block that transforms the content; it works as a dedicated portion of the network to change in a content- and style-specific way. They focused on learning how it is necessary to transform the details of the content image by utilizing objects from the same category in both content and style images. In general, they improved that their model has the ability to stylize the content details from a single complex object class. Their model was evaluated by persons who had no art information and art experts, and both of them voted that the current model is better than and preferred compared to other previous studies. In addition, the objective and subjective evaluations indicated that their model is better than the listed methods in their related works section in terms of the quality of stylization.

Choi (2022) presented a model that used second-order statistics of the encoded features to construct an optimal arbitrary image style transfer technique. Their study had two contributions; in the former, they proposed a new technique for correlation-aware loss and feature alignment. This regular merging of loss and feature alignment techniques robustly match the statistics of the second order of the content features and the target style features, thus, increasing the decoder network style capacity. Whereas in the latter, they proposed a new component-wise style controlling technique. Their method could generate diverse styles depending on singular or multistyle images by utilizing style-specific components from second-order feature statistics. They proved that their model accomplishes improvements in the decoder network style capacity and style diversity without losing the capability to process in real time on GPU devices (under 200 ms). The pre-trained VGG16 network was used for the encoder, whereas for the decoder, trainable VGG16 was used. MS-COCO dataset was used as a content image dataset, whereas a small dataset that consisted of 22 images represents style images. Finally, they relied on increasing the training data to measure the losses of the networks.

B. Deep Dream

Yin, et al. (2020) presented a model called DeepInversion. Their model consisted of two sections; the first is teacher, whereas the second is called student. The teacher started with random noise, where the training dataset did not use any extra information. The DeepInversion model is built based on deep dream by improving the quality of images of the deep dream by adding a new feature to the image regularization. ImageNet and CIFAR-10 datasets had used to train the model.

Kiran (2021) presented a deep dream algorithm. The training is started as soon as the image is entered into the model. With this algorithm, the input image is modified by firing particular neurons. Thus, the layers had the ability to choose particular neurons to fire. This process is continued until all the required features by the input image become in the targeted layer. Hence, the more feeding images, the more able to extract more features. ReseNet, CNN, and ANN pre-trained model had used to extract the image features. At each layer, they executed gradient ascent to maximize the loss function.

El-Rahiem, et al. (2022) presented a multi-biometric cancellable scheme (MBCS) using deep dream. The purpose was to create secure and effective fingerprint cancellable patterns depending on the veins of the finger. The Inception v3 model had used to accomplish this study. They maximized the loss function using gradient ascent, as large the number of iterations as the loss maximized.

IV. BACKGROUND

This section browses the most important concepts relative to our study. As mentioned above, the model is divided into two parts; style transfer and deep dream. Some concepts related to style transfer, whereas others are used with the deep dream.

A. NST General Concepts

The concepts related to NST are VGG-19, Gram matrix, and gradient descent.

VGG-19 network

VGG-19 network is used as a pre-trained model, consisting of 19 layers to perform the convolution process; it has 16 convolutional layers and 5 max pooling layers. Each convolutional layer is followed by a rectified linear unit (ReLU) which works as an activation function. These layers (convolutional and ReLU) are fully connected with the max pooling layer. Unlike traditional convolution, the multiple layers of convolutional and non-linear activation functions enable to extraction of additional image features, whereas max-pooling layers are used to downsampling and select the most valued features. Finally, the fully connected layer is the last in this network. It takes the results of the previous layers, makes flattens, and delivers its output to the activation function (often SoftMax function). Fig. 1 shows the structure of the VGG-19 model (Rashid, et al., 2020; Sudha and Ganeshbabu, 2021; Xiao, et al., 2020).

Gram matrix

First, the features of content and style images have been extracted using the convolutional neural network (CNN). To make the style features that have been extracted more useful, one pre-processing additional step is required. Therefore, the Gram matrix is an important step to make the extracted style features more effective. All the extracted features by the CNN still retain the information related to the image's content, such as object structure and positioning. The gram matrix is then applied to those extracted features to remove
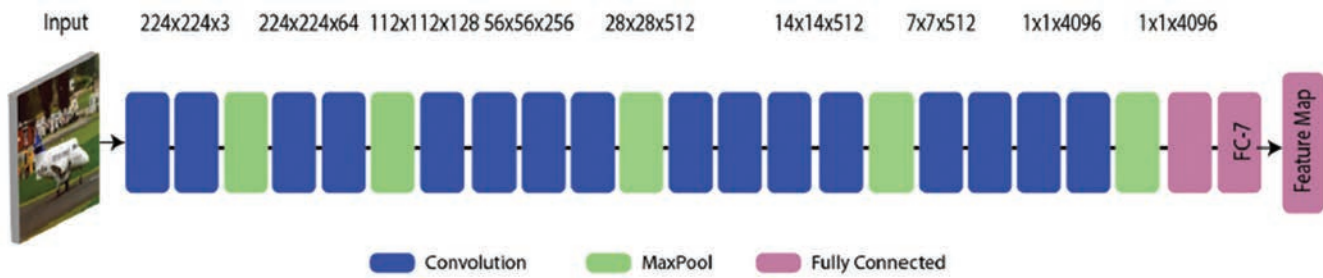
Fig. 1. VGG-19 Architecture.

the content-related information without affecting the style information. Hence, the Gram matrix assists for creating information on the texture related to the data when applied to the extracted features from CNN (Tuyen, et al., 2021).

3. Gradient decent

Gradient descent is an optimization mechanism that aims to minimize the loss function by computing the gradients required to update the parameters of the network. For deep learning, the most familiar algorithm to update the parameters that are used with the gradient is backpropagation, so, they work together as an effective learning algorithm for adjusting the errors by backward propagation through the network layers from the last one going back to the first layer (Wani, et al., 2020).

B. Deep Dream General Concepts

The concepts that relate to deep dream are Inception v3 and gradient ascent.

Inception v3

The Inception model works under the architecture of deep CNN, which aims to decrease the influence of computational effectiveness and low parameters in cases of application. The Inception-v3 model adopts a convolutional filter of various sizes; thus, different areas of receptive fields have been enabled (Cao, et al., 2021). Table I summarizes the Inception-v3 network architecture, where each module takes the output size of the previous one as its input (Cao, et al., 2021; Szegedy, et al., 2016).

Gradient ascent

The process of creating Deep Dream is based on using gradient ascent, which is used at each step of this process no matter the size of the image (it is used from the smallest to the largest image). The goal of gradient ascent is maximizing the loss function (El-Rahiem, et al., 2022).

V. Methodology

The model consists of two main techniques that work together to produce the desired system. These techniques are NST and deep dream. Hence, the model works as illustrated in the following steps:
1. Input the main image (content image) to the model.
2. Apply the NST model as shown below:
   - Choose one image among the style images to extract the desired style.

TABLE I
Network Structure of the Inception-v3 Model

Type	Patch size/Stride	Input size
Conv	3×3/2	299×299×3
Conv	3×3/1	149×149×32
Conv	3×3/1	147×147×32
Pool	3×3/2	147×147×64
Conv	3×3/1	7373×73×64
Conv	3×3/2	71×71×80
Conv	3×3/1	35×35×192
3×Inception	-------	35×35×288
3×Inception	-------	17×17×768
2×Inception	-------	8×8 × 1280
Pool	8×8	8×8 × 2048
Linear	logits	1×1 × 2048
SoftMax	classifier	1×1 × 1000

   - Extract content and style from the content and style image, respectively.
   - Compute the loss for each one of the images.
   - Compute the total loss.
   - Output the resulting image.
3. Input the resulting image from the previous model to the deep dream model, then apply the following steps:
   a. Use the Inception v3 pre-trained model as a deep CNN model to extract the image features (The extraction of the image features starts from low-level features and ends with high-level features).
   b. Compute the error loss.
   c. Compute the gradient using gradient ascent.
   d. Repeat until the required steps.
   e. Output the resulting image.
4. Present the final image.

In the below two subsections, we illustrate each one of the model components.

A. NST

First, the images must be resized to fit the input of VGG-19, which is 224 × 224. The NST aims to split the content and style of the image, and then, the content of one image is merged with the style of the other image creating a new image. The objects are similar to the content image, whereas colors and structures are similar to style images. Fig. 2 shows our NST proposed system.

VGG-19 pre-trained model is used to extract high- and low-level features from content and style images,

respectively, where it is trained on the ImageNet dataset that contains millions of images.

Gram matrix has been applied to retain the important features of content images like objects and remove the others with keeping the style of the style image. Equation 1 defines the Gram matrix (Tuyen, et al., 2021).

$$Gram = V^T.V \qquad (1)$$

Where, V is an arbitrary vector and multiplied by its transpose V^T.

For more explanation, when $V = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, then $V^T = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$

The stochastic gradient descent (SGD) is used to compute the gradient descent, in which we compute the gradient for one sample at a time then the values of the parameters are updated according to it. Equation 2 shows how SGD computes the parameters (Wani, et al., 2020).

$$w = w - \mu.\nabla E(w; x(i); y(i)) \qquad (2)$$

The $\nabla E(w; x(i); y(i))$ represents a gradient of the error loss, whereas the $\{x(i); y(i)\}$ is the training sample. μ is the learning rate.

After that, we compute the loss for both content and style images. Gatys, et al., 2015, proposed a way to compute the loss by computing the square of error loss between two features in the content image calculated from Equation 3.

$$\mathcal{L}_{content}(\vec{p}, \vec{x}, l) = \frac{1}{2} \sum_{ij} (F_{ij}^l - P_{ij}^l)^2 \qquad (3)$$

Here, $\vec{p}, \vec{x}$ are the original image, F_{ij}^l is the activation of the ith mask at the position j in the layer l. Whereas F^1, P^1 are the representative feature of both $\vec{x}$ and $\vec{p}$ respectively.

The correlation of features is computed by the Gram matrix G^l as in Equation 4.

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l \qquad (4)$$

Here, G_{ij}^l represents the inner product between vectorized feature map i and j in the layer l.

The texture of the style image that matches a given image style is generated using gradient descent to find another image matching the representation of the original image. To do this, it must reduce the difference between entries of the Gram matrix of the original and the generated image Gram matrix must be minimized by minimizing the mean-squared distance between them. The style loss calculated from Equation 5.

$$\mathcal{L}_{style}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_l E_l \qquad (5)$$

Here, w_l represent the weighting factors that each layer contributes to the total loss and $\vec{p}, \vec{x}$ are the original and generated image respectively.

Finally, we can get the output image by mixing the content from the content image and the style from the style image by minimizing the loss of content representation of one layer of the content image and the style representation of several layers of the style image. Equation 6 represents the total loss of the resulting image.

$$\mathcal{L}_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{content}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{style}(\vec{a}, \vec{x}) \qquad (6)$$

α and β represent the weighting factors for the content and style representation, respectively.

B. Deep Dream

The deep dream model was built using deep CNN, where the Inception v3 pre-trained model had used to build this model. Fig. 3 shows the steps of the deep dream proposed system.
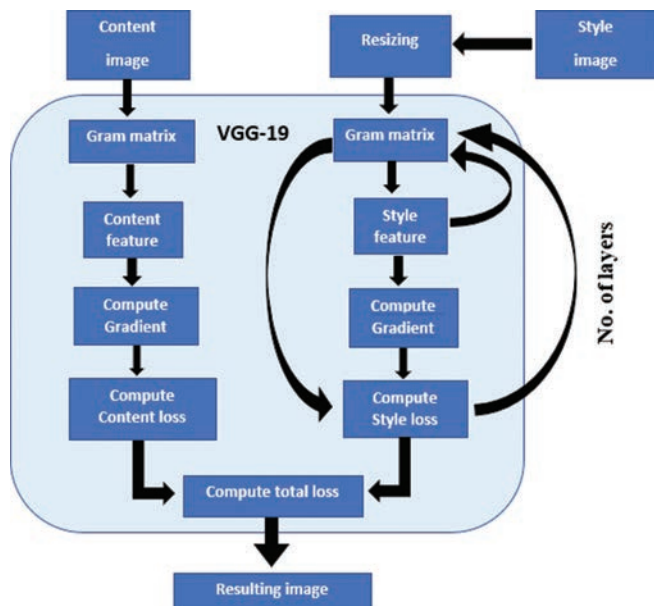
Fig. 2. Neural style transfer proposed system.
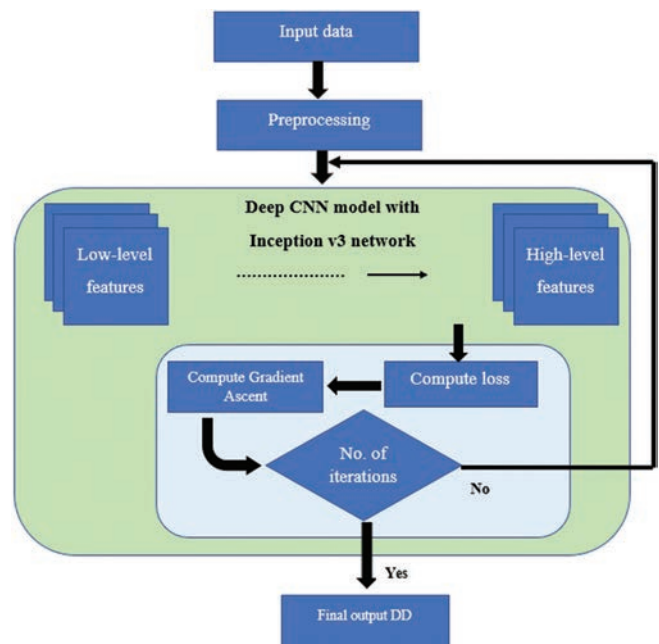
Fig. 3. Main steps of deep dream proposed system.
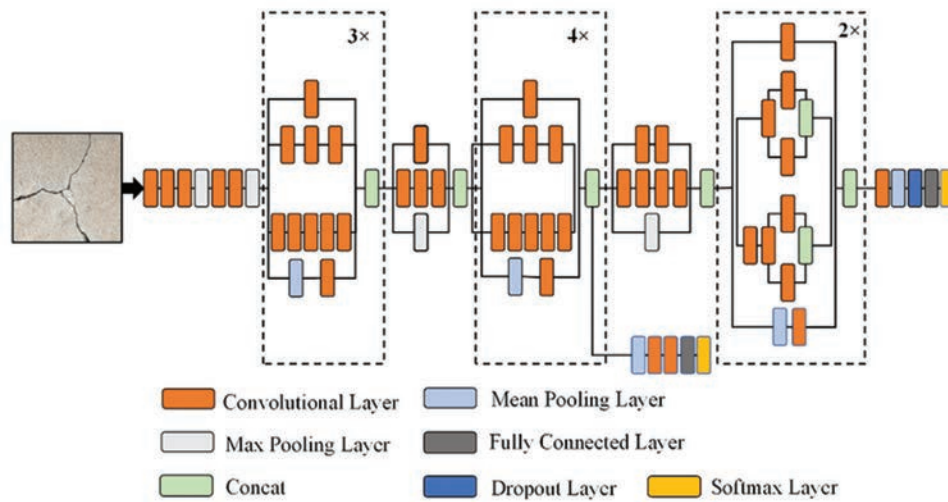
Fig. 4. Inception V3 layer architecture.
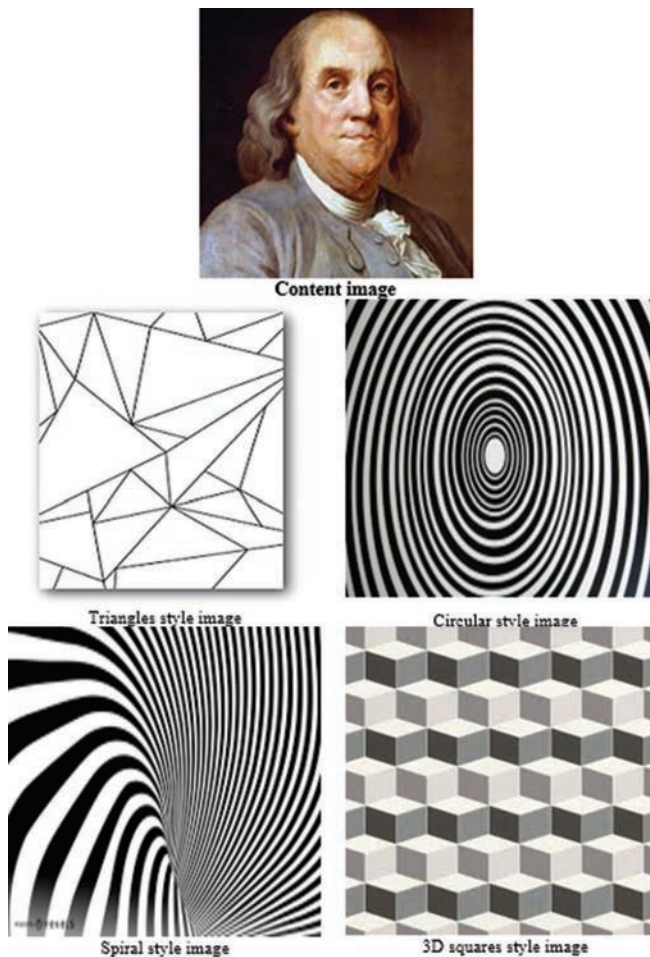
Fig. 5. Content and style images.

The input of the deep dream system is the image that resulted from the NST system. This image must be pre-processed by resizing it to be the size of that appropriate to the Inception v3 model. The size of the image was edited to 299 × 299. Then, the resized image had entered into the pre-trained CNN model, which is Inception v3.

The Inception v3 network consists of 42 layers of depth. It is a pre-trained network with an ImageNet dataset and works to reduce the dimensions of the network by reducing the number of parameters in each layer (Szegedy, et al., 2016).

Using Inception v3, we extract the low-level features at first; then, high-level features are extracted. The Inception v3 network is implemented as shown in Fig. 4 (Ali, et al., 2021), where the first block has been repeated 3 times, the second repeated 4 times, and the third block repeated 2 times.

Then, the loss is computed by adjusting the weights. To do that, at the chosen layers, we summed the activations. All layers have the same contribution, no matter whether it is large or small layers, because, at each layer, we normalized the loss. Our goal is to maximize the value of the loss; this had done using gradient ascent.

After calculating the loss for all layers that had been chosen, we should calculate the gradient w.r.t. the image, the result added to the original image. Each time we add the gradient to the original image, we contribute to creating an enhanced image by increasingly activating particular layers in the network. Adjusting weight to maximize the loss is the core process in a deep dream. The equation of computing the gradient ascent is the same equation of the gradient descent with only one difference represented by flipping the sign of the equation of the gradient descent. Hence, we adjusted the new weights based on the old weights and the gradient of the error function. The overall process was repeated many times; we got different results according to the number of repeated iterations as described in the next section.

VI. Results and Discussion

Our model used pre-trained networks; VGG-19 for NST and Inception v3 for a deep dream, both of them are trained on the ImageNet dataset.

We start with NST implementation. Here, there are two types of images, which are content images and style images, as shown in Fig. 5.

Fig. 6. The resulting images and their losses after 1000 iterations.

The resulting images after 1000 iterations with the loss of every image are shown in Fig. 6.

The loss is decreased as we increase the number of iterations because of using gradient ascent, which minimizes the loss function, thus, enhancing the accuracy of the resulting images. Furthermore, as going deeper through increasing the iterations, it is noticeable that the style of the style originated more and more in the content image, and the content of the content image still preserves the high-level features whereas some of the low-level features such as edges and lines are lost. Fig. 7 shows the resulting images after 10,000 iterations and the loss of every image.

All these output images had entered into the deep dream model after resizing it to 299×299 pixels to make the dreamed image.

In the deep dream algorithm, we used a 0.005 learning rate with 1000 steps to get the final results as shown in Fig. 8.

We measure the quality of the deep dream based on the loss value. The loss value must be balanced; a very high value of the loss may make the dreamed image unrecognizable, whereas a very low loss value does not achieve the purpose of the deep dream. The loss value of the NST works in opposite to the deep dream one, where we work to minimize the loss value. The value of the loss after the 10,000 iterations was 10,431.64 for the triangle style image, 23,225.48 for the circular style image, 40,709.7 for the spiral style image, and finally, 1788.9 for the 3D squares style image. It is clear that the lowest loss value, the better the image quality in NST.

Figs. 9 and 10 represent the loss values of NST and deep dream, respectively.

The deep dream algorithm maximizes the value of loss by increasing the number of steps. The loss value after 900

Fig. 7. The resulting images and their losses after 10,000 iterations.
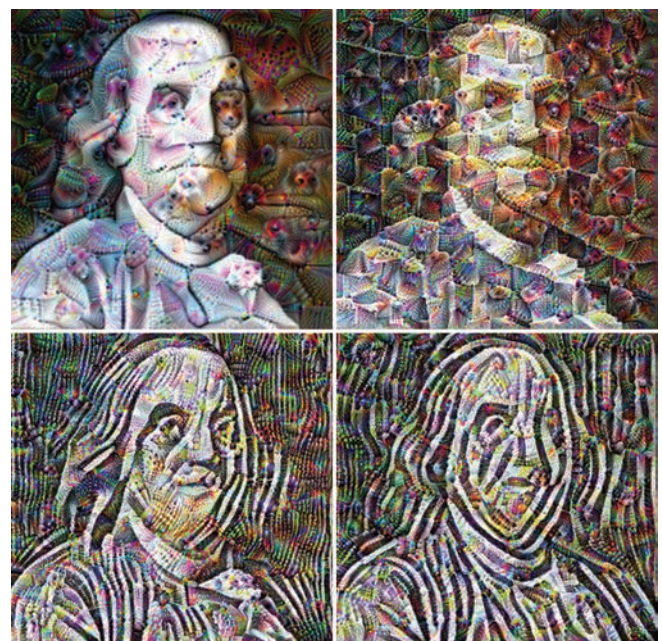
Fig. 8. Stylised images after applying the deep dream algorithm.

steps for the images that are stylized under the NST and then processed in the deep dream algorithm are 3.816556454 for the stylized triangle image, 3.1598 for the 3D squares stylized image, 3.0736 for the circular stylized image, and finally, 3.0087 for the stylized spiral image. The loss in Fig. 10 is for images that passed in the NST model and then entered into the deep dream model.
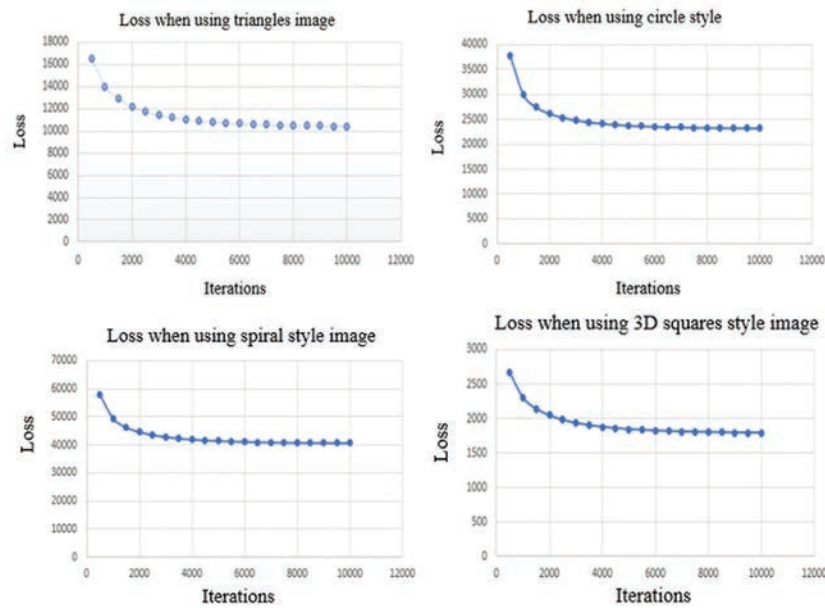
Fig. 9. The loss of the images in the style transfer model.
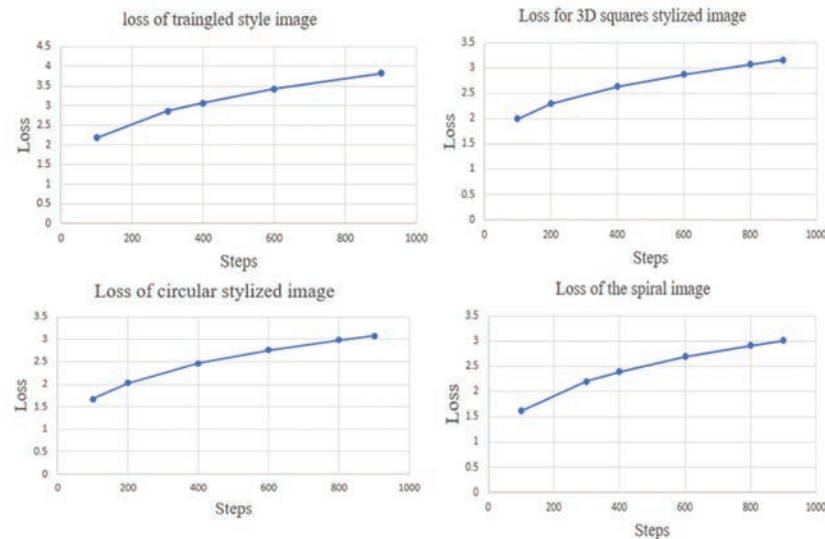
Fig. 10. The loss of the stylized images in the deep dream model.

Fig. 11. Applying deep dream on the original image.

Fig. 11 shows the output of the original image that passed in the deep dream model, with a loss value of 3.6324.

VII. Conclusion

Deep dream and NST represent the most modern deep learning techniques in recent years. This study presents a deep dream model that takes the output of NST images as input and then applies the deep dream algorithm. VGG-19 pre-trained network is used as a deep CNN network to implement the NST based on the Gram matrix that represents the core of this model and the gradient descent, which minimizes the loss function when the image is cleared and increases with the distorted images.

The results of the NST are the input to the deep dream; the loss varies among the different images in the deep dream.

The Inception v3 network is a pre-trained model used to build the deep dream algorithm. The loss here is maximized based on gradient ascent. The loss of the distorted images (i.e., circular and spiral stylized images) is less than the clear images (i.e., 3D squares and triangles images).

REFERENCES

Abedi, W.M., Nadher, I., Sadiq, A.T. and Al., E., 2020. Modified deep learning method for body postures recognition. *International Journal of Advanced Science and Technology*, 29, pp.3830-3841.

Ali, L., Alnajjar, F., Jassmi, H.A, Gochoo, M., Khan, W. and Serhani, M.A., 2021. Performance evaluation of deep CNN-based crack detection and localization techniques for concrete structures. *Sensors,* 21, p.1688.

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M. and Farhan, L., 2021. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8, pp.1-74.

Cao, J., Yan, M., Jia, Y., Tian, X. and Zhang, Z., 2021. Application of a modified Inception-v3 model in the dynasty-based classification of ancient murals. *EURASIP Journal on Advances in Signal Processing*, 2021, p.49.

Chen, X., Zhang, Y., Wang, Y., Shu, H., Xu, C. and Xu, C., 2020. Optical flow distillation: Towards efficient and stable video style transfer. In: *Lecture Notes in Computer Science (LNCS)*. Springer Science, Germany.

Choi, H.C., 2022. Toward exploiting second-order feature statistics for arbitrary image style transfer. *Sensors(Basel)*, 2022, p.2611.

El-Rahiem, B.A., Amin, M., Sedik, A., Samie, F.E. and Iliyasu, A.M., 2022. An efficient multi-biometric cancellable biometric scheme based on deep fusion and deep dream. *Journal of Ambient Intelligence and Humanized Computing*, 13, pp.2177-2189.

Gatys, L.A., Ecker, A.S. and Bethge, M., 2015. A Neural Algorithm of Artistic Style. arXiv Prepr. arXiv1508.06576.

Khan, A., Sohail, A., Zahoora, U. and Qureshi, A.S., 2020. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53, pp.5455-5516.

Kiran, T.T., 2021. Deep inceptionism learning performance analysis using TensorFlow with GPU-deep dream algorithm. *Journal of Emerging Technologies and Innovative Research*, 8, pp.322-328.

Kotovenko, D., Sanakoyeu, A., Ma, P., Lang, S. and Ommer, B., 2019. A content transformation block for image style transfer. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, United States, pp. 10024-10033.

Li, H., 2018. *A Literature Review of Neural Style Transfer*. Princeton University Technical Report, Princeton NJ, p.085442019.

Mordvintsev, A., Olah, C. and Tyka, M., 2015. Inceptionism: Going Deeper into Neural Networks. Available from: https://googleresearch.blogspot.co.uk/2015/06/inceptionism-going-deeper-into-neural.html [Last accessed on 2022 Aug 03].

Rashid, M., Khan, M.A., Alhaisoni, M., Wang, S.H., Naqvi, S.R., Rehman, A. and Saba, T., 2020. A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection. *Sustain*, 12, p.1-21.

Singh, A., Jaiswal, V., Joshi, G., Sanjeeve, A., Gite, S. and Kotecha, K., 2021. Neural style transfer: A critical review. *IEEE Access*, 9, pp.131583-131613.

Sudha, V. and Ganeshbabu, T.R., 2021. A convolutional neural network classifier VGG-19 architecture for lesion detection and grading in diabetic retinopathy based on deep learning. *Computers, Materials and Continua*, 66, pp.827-842.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, United States, pp.2818-2826.

Tuyen, N.Q., Nguyen, S.T., Choi, T.J. and Dinh, V.Q., 2021. Deep correlation multimodal neural style transfer. *IEEE Access*, 9, p.141329-141338.

Wani, M.A., Bhat, F.A., Afzal, S. and Khan, A.I., 2020. Advances in Deep Learning. Springer Nature, Singapore.

Xiao, J., Wang, J., Cao, S. and Li, B., 2020. Application of a novel and improved VGG-19 network in the detection of workers wearing masks. *Journal of Physics: Conference Series*, 1518, 012041.

Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K. and Kautz, J., 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, United States, pp.8712-8721.

Driver Drowsiness Detection Using Gray Wolf Optimizer Based on Voice Recognition

Sarah S. Jasim[1], Alia K. Abdul Hassan[2], and Scott Turner[3]

[1]Department of IT, Technical College of Management-Baghdad, Middle Technical University,
Baghdad, Iraq

[2]Department of Computer Science, University of Technology,
Baghdad, Iraq

[3]School of Engineering, Design, and Technology, Church Christ Church University,
Kent, UK

Abstract—**Globally, drowsiness detection prevents accidents. Blood biochemicals, brain impulses, etc., can measure tiredness. However, due to user discomfort, these approaches are challenging to implement. This article describes a voice-based drowsiness detection system and shows how to detect driver fatigue before it hampers driving. A neural network and Gray Wolf Optimizer are used to classify sleepiness automatically. The recommended approach is evaluated in alert and sleep-deprived states on the driver tiredness detection voice real dataset. The approach used in speech recognition is mel-frequency cepstral coefficients (MFCCs) and linear prediction coefficients (LPCs). The SVM algorithm has the lowest accuracy (71.8%) compared to the typical neural network. GWOANN employs 13-9-7-5 and 30-20-13-7 neurons in hidden layers, where the GWOANN technique had 86.96% and 90.05% accuracy, respectively, whereas the ANN model achieved 82.50% and 85.27% accuracy, respectively.**

Index Terms—**Drowsiness, Artificial neural network, Feature extraction, Gray Wolf Optimizer, Normalization, Mel-frequency cepstral coefficients, Linear prediction coefficients**

I. Introduction

Transportation safety officials have been concerned about distractions while driving as a significant contributor to traffic accidents (Dasgupta, et al., 2015; Zhang, et al., 2017). Such incidents necessitate research into on-board monitoring of the attentiveness level of the driver. Human alertness can be measured using various methods, including an electroencephalogram (EEG), ocular characteristics, blood samples, speech (Ooi, et al., 2016; Jasim and Hassan, 2022), and skin conductance. The EEG, electrooculography (EOG)

signals, and blood biochemical techniques have been deemed the most accurate for determining the state of alertness. However, because it relies on direct physical contact, they are challenging to put into widespread practice (Alzu'bi, Al-Nuaimy and Al-Zubi, 2013; Huo, Zheng and Lu, 2016). Driving can be viewed as a road, vehicle, and driver system. Roads and cars are already more stable thanks to research, but the driver is still the weakest link. Therefore, it has been argued that determining drivers' emotions should be a present study emphasis (Zhang, 2019). Emotions are a complex phenomenon that is difficult to define. It is a combination of mental and bodily sensations induced by thoughts. Some emotions cause avoidance or approach.

In contrast to avoidance, approach conduct is accompanied by good emotions (happiness) (Ooi, et al., 2016). Drivers' emotions strongly influence their driving behavior, which is closely linked to road safety. Therefore, understanding and monitoring driver emotions are believed to promote better driving ethics and the need to develop emotion-based accident prevention systems. Exhaustion, rage, and stress driving are all linked to traffic crashes (Dasgupta, et al., 2015).

Since the early 1960s, Automatic Speech Recognition (ASR) has been the most researched topic in speech processing, translating human spoken words into computer recognized words. It mainly entails extracting speech information and storing it in a data model. ASR comprises two phases (training and testing). In the training phase, known speech is recorded and stored in a speech database as a parametric representation (Badr and Abdul-Hassan, 2020; Abdul-Hassan and Hadi, 2020). In the testing phase, the ASR system compares the retrieved features to trained reference templates to identify the utterance (Gamit and Dhameliya, 2015).

The sympathetic nerve system includes the hippocampus, amygdala, and cingulate gyrus; these organs respond to attentiveness, showing a link with emotion. We can also employ sympathetic nervous system information from sweat glands to recognize emotions (Yoshida, et al., 2014). The present study aims to identify the emotion of drivers during speaking in two scenarios (drowsy and non-drowsy). These authors recognize that a lone driver would not necessarily be

speaking whilst driving, but this could be partially remedied by the system that requires the driver to speak every so often. However, in my study took this challenge, which is to rely on the driver's voice and analyze it based on the energy property of the human voice because the level of energy of his voice in a state of drowsiness decreases, therefore, this factor is essential in analyzing the characteristic of drowsiness.

The paper is organized as follows. Section II discusses the related work. Section III presents the proposed system and the voice-based algorithm. Section IV presents the results and discussion of the experiment. Finally, the paper is concluded in Section V.

II. Related Work

Recent advances in artificial intelligence (AI) have revolutionized image, signal, and voice detection (Yu, Wang and Zhang, 2021). Here are some recent works on emotion recognition:

An overarching paradigm for identifying drowsiness states using prosody, articulation, and other speech quality-related aspects were described by Krajewski, Batliner and Golz, 2009. Speech, speaker, and affect acoustic recognition were all used (frame-level-based speech features). In addition, features from perceptual and signal processing (such as fundamental frequency [FF], intensity [I], pause patterns [PP], formants [F], and cepstral coefficients [CC]) were utilized in the feature computation, used four different classifiers (SVM, KNN, MLP, and DT). The SVM classifier was 86.1% accurate.

Dasgupta, et al. (2015) introduced a voice and vision-based sleepiness detection system suitable for car drivers. It was done using an image-based algorithm that computes PERCLOS and a voice-based method that computes voiced speech ratio (VSR). In addition, face detection was performed with 94% accuracy using a support vector machine classifier (SVM).

Ooi, et al. (2016) used electrodermal activity (EDA) to study stress and anger as critical emotions. A simulated driving assignment containing neutral, stress, and anger scenarios was designed for emotional stimulation. First, the acquired EDA signals were filtered bandpass at 0.5–2 Hz and Fourier transformed. After that, the power spectral density mean, median, and standard deviation were calculated, and an f-test of two samples was used to examine the statistical significance of the data. A support vector machine classifier (SVM) was employed for classification and achieved 85% each for neutral stress, neutral anger, and 70% for stress anger, respectively.

Wankhade and Kharat (2017) presented a novel approach for emotion recognition using EEG signals and using a two-tier classifier based on both K-nearest neighbor (K-NN) and neural network (NN) classifiers. The proposed KNN-NN classifier achieved an accuracy of 97%.

Using an EEG-based emotion classification and a decision tree algorithm, Pane, et al., 2017, developed emotion categorization models for four states of mind: Happy, sad, angry, and relaxed. To divide the EEG signal into gamma, beta, alpha, and theta bands, researchers utilized an IIR bandpass filter with a Chebyshev type II window. Three algorithms were employed to classify emotions: Repeated Incremental Pruning to Produce Error Reduction (RIPPER), J4.8, and SVM. On average, RIPPER achieved 92.01% accuracy in the binary assessment of sad versus relaxed emotional state.

Greco, et al. (2019) investigated the possibility of combining electrodermal activity (EDA) and voice data to recognize human arousal levels during single effective word pronunciation. The support vector machine with a recursive feature elimination (SVM-RFE) was trained and tested on three datasets, using the two channels (speech and EDA) independently and combined. The results suggest that combining EDA and speech data improve the classifier marginally (+11.64%).

The three methods presented by Martin, et al., 2021, are based on the acoustic quality of the speech, reading faults, and a completely new method that relies on Automatic Speech Recognition systems' errors. Unweighted average recall (UAR) values of 74.2% were achieved with a classification system built on a principal components analysis (PCA) and logistic regression (LR) (Table I).

III. The Proposed System

At present, researchers in pattern recognition and machine learning are increasingly recognizing the importance of swarm optimization for reducing data dimensionality and improving classification accuracy. Features can be selected using various swarm optimization techniques, such as particle swarm optimization (PSO) or ant colony optimization (ACO). New swarm-inspired algorithms, such as Gray Wolf Optimization (GWO), have recently appeared (Hassan and Mohammed, 2020). GWO is a novel metaheuristic algorithm based on how Gray Wolves behave in the wild. The group used to have two species: Wolf males and wolf females to keep order (Xu, et al., 2019). Any pack has a social hierarchy that looks like this:

1. Alpha wolves (decide [α]) are dominating wolves. Wolves are at the forefront and represent the herd's leadership and the issuance of decisions.
2. Beta wolves (β) are second-level wolves. Beta's support the dominant decisions.

TABLE 1
Summary of Artworks Utilizing Hybrid Techniques

References	Hybrid features	Machine learning methods	Accuracy
Krajewski, Batliner and Golz, 2009	FF, I, PP, F, and CC	SVM	86.1%
Dasgupta, et al., 2015	VSR+PERCLOS	SVM	94%
Ooi, et al., 2016	EDA signals	SVM	85% and 70%, respectively
Wankhade and Kharat	EEG signals	KNN-NN	97%
Pane, et al., 2017	EEG signal	RIPPER	92.01%
Greco, et al., 2019	Speech+EDA	SVM-RFE	+11.64%
Martin, et al., 2021	AQS, RF, ASRse	PCA+LR	74.2%

3. Delta wolves (δ) are the third-level wolves who obey the alphas and betas.
4. The omegas wolves (ω) signify the pack's smallest alpha, beta, and delta wolves' scheme. The GWO algorithm ranks alpha wolves first, followed by beta and delta wolves. Therefore, omegas (ω) comprise this population cluster (Heidari and Pahlavani, 2017).

The general stages of applying GWO are as follows:

Start the Gray Wolf population Xi, in which I = 1, 2, 3, 4... n; Start a, A and C, where a is a vector which on linearly decreasing from 2 to 0, A and C are a coefficient matrix. N is the number of total iterations for optimization. The fitness of each candidate solution is computed through equations:

$$X(t+1) = X(t) - A.D \ldots\ldots\ldots \tag{1}$$

$$D = \left| C.X_p(t) - X(t) \right| \ldots\ldots\ldots \tag{2}$$

$$A = 2a.r_1 - a \ldots\ldots\ldots \tag{3}$$

Where, X and $X\alpha$ are *the first* fitness *search agent*, Xβ *is the second* fitness *search agent*, and $X\delta$ *is* the third fitness *search agent*. The optimal Gray Wolf position ($X\alpha$) has the best fitness value f($X\alpha$).

The operation starts by randomly generating the Gray Wolf population Xi and repeating operations until the stop condition is not satisfied. The primary step is to modify all search agents (each current search agent's position through equation) using Equation (2.45) and find the fitness values of all modified search agents.

$$X(t+1) = \frac{(X\alpha + X\beta + X\delta)}{3} \ldots\ldots\ldots \tag{4}$$

The stop condition is represented by the maximum iteration until there is no update of the current search agent (best result).

A. Proposed Gray Wolf Optimizer (GWO) with artificial neural network (GWOANN) based drowsiness detection system

The proposed method Gray Wolf Optimizer with Artificial Neural Network (GWO-ANN) was used for training the speech signal features for these reasons:
- Finding the best starting weights that are valid proved to be a challenging task with other approaches.
- There was a concern about potential overfitting in neural networks
- The trend toward minimizing the mean square error in the training stage.
- The initial range of the initial weights may put some constraints on the search space.

To overcome the above problems, the proposed classifier method is the GWO-ANN based on the relationship between input and output using GWO. The method used a specific weight for neurons obtained by GWO. These weights are efficient in the training process. First, the ANN is trained with GWO algorithms to find the optimum weights and biases. The neural network then employs one-step-forward ANN. Finally, the fitness function (Equation 1) is used to evaluate the results. The total MSE is used as a fitness function.

First stage: Using GWO for ANN training:
Selection: Size of individuals (Packs),
Improvement: Largest iterations are identified,
Formation: Using back-propagation algorithm in NN, implementation: GWO in the following equation uses to get the better value of weights and bias,

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (T_i - Y_i)^2} \tag{5}$$

Where, n shows the number of test samples, T is the target value, and Y is the predicted output. It means a better model.
Restoring: Optimum of weights and bias,
Second stage: GWO used to optimize back-propagation algorithms for training ANN:
Selection: GWO outputs as the beginning of weights and bias,
Restoring: Optimum training of ANN model.

B. Speech-Based Method

The GWO-ANN described above represents the training model of this method. A total of 18 features of the recorded speech are obtained from the previous steps and then used for training. The size of input features differs here depending on the specified energy threshold, which neglects the silence segments. Finally, each frame is input to the proposed method for training with its class label (drowsy or not). The general steps of the proposed method for detecting the driver's drowsiness are explained in Fig. 1.

Dataset

In this work, a speech dataset was recorded through (human interaction) as a large vowel phonation, single words, connected speech, reading the speech, and impromptu speaking as all forms of oral communication. The samples were selected because they were common phrases and had a mixture of required frequency components. A Samsung mobile device was used for 16 subjects (eight men and eight women), and each person had two samples for 12 phrases (in the sleepy and normal states). Examples of the speech data are shown in Fig. 2 for the normal and abnormal datasets. Signal energy has been used with edge detection algorithms to distinguish between voiced and silent clips because the level of energy of a person's voice in drowsiness decreases. Therefore, this factor is essential in analyzing the characteristic of drowsiness. Each frame has several samples (ω), where, ω < n where n is all sample size. Frame-by-frame, the intensity of a speaker's words is measured. Each sample has been squared, and the total of all squared samples has been computed. The following is the estimated energy equation:

$$Energy = \sum_{i=1}^{\omega} x_i^2 \tag{6}$$

Feature extraction

It is essential to execute a standard pre-processing step to reduce noise (small frames) and remove noise from voice signals to improve speech recognition results; the intensity of the speaker's words is measured frame by frame. Each
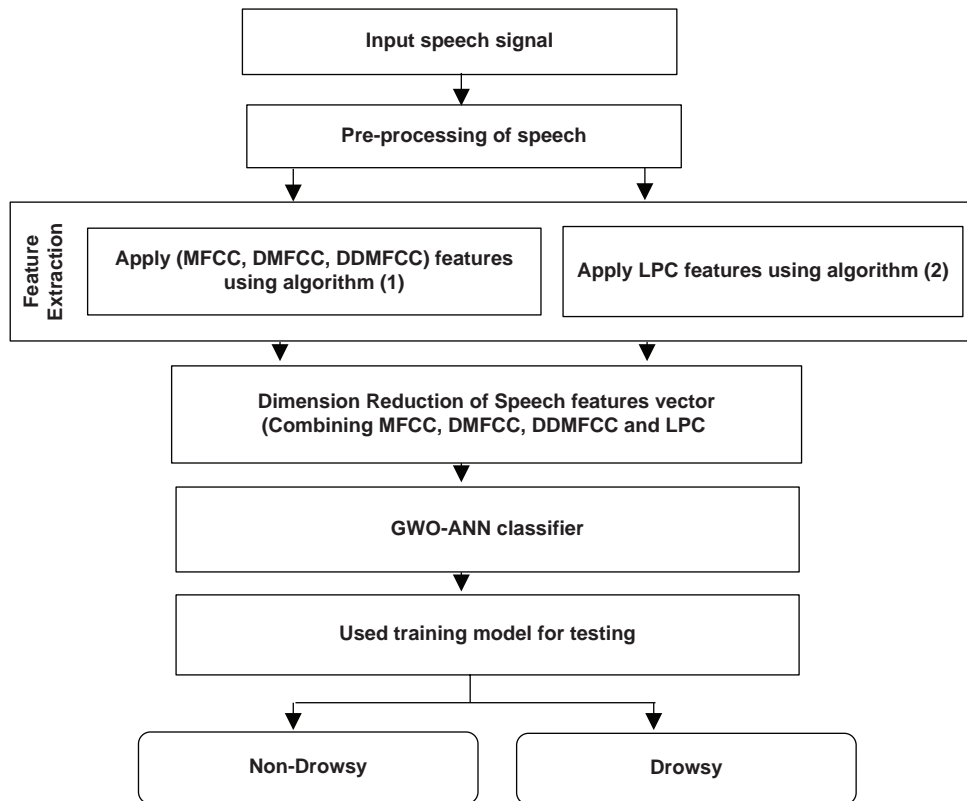
Fig. 1. The proposed speech drowsiness detection sub-system.

Fig. 2. (a) The samples of normal voice signal. (b) The samples of abnormal voice signal.

sample was squared, and the total of all squared samples was calculated using a specific threshold limit chosen by the experiment to reduce noise from a signal.

The speech stream is transformed into measurable values with differentiating characteristics. In this approach, two features are extracted from each utterance frame: Mel-frequency cepstrum coefficients (MFCCs) and linear predictive coding (LPC). The MFCCs characteristics exhibit promising performance in speech-related challenges, due to their capacity to accurately represent the speech spectrum's

amplitude. On the other hand, LPCs exhibit qualities comparable to formant frequencies and are resistant to noise. These two groups of characteristics were combined because LPCs features are intended to be complementary to MFCCs features. Each frame will produce two distinct groupings of features: 42 MFCCs with their first and second derivatives and 12 LPCs. The number of MFCC features is chosen to be 14 dimensions for the first derivative and 14 dimensions for the second derivative. These dimensions provide the best results found through trial and error. The algorithms (1 and 2) show how this feature is extracted. The samples of these features explain in Tables II and III, respectively.

Dimension reduction of features

The total features of speech signals are represented by MFCC, delta of MFCC (1st derivative), delta-delta of MFCC (2nd derivative), and LPC. This feature is a total of 54 features for each frame. The dimension reduction is applied by getting each 54 features per frame together and applying statistical values for each block, such as mean, standard deviation, skewness, kurtosis, root mean squared, energy, power, peak amplitude, and crest factor, for all these features obtained the mean and standard deviation.

Algorithm (1): The MFCC feature extraction

Input: Speech (phonocardiogram) signal
Output: Total MFCC features, total delta MFCC features, total delta-delta MFCC features, and total class

Begin
Step 1: Assign no of speech files (normal and abnormal)
Step2: For each speech file
  - Split into frames, phonocardiogram signals
  - Apply humming windowing to frames
  - Get spectrum by applying FFT to all frames
  - Determine matrix for a Mel spaced filter bank
  - Transform spectrum to Mel spectrum
  - Obtain the MFCC vector for each frame by applying DCT (a feature I)
  - Obtain delta MFCC (Features II)
  - Obtain delta-delta MFCC (features III)
Step 3: Assign a class label for each frame
Step 4: Append all features into a new matrix called total MFCC features
Step 5: Append all features into a new matrix called total delta MFCC features
Step 6: Append all features into a new matrix called total delta-delta MFCC features
Step 7: Append class label into a new vector called total class
        End Algorithm

Algorithm (2): The LPC feature extraction
Input: Speech (phonocardiogram) signal
Output: Total LPC features and total class
Begin
Step 1: Assign no of speech files (normal and abnormal)
Step 2: For each speech file
  - Split into frames, phonocardiogram signals
  - Apply humming windowing to frames
  - Predicate frame and compute the error
  - Obtain LPC coefficients
Step 3: Assign a class label for each frame
Step 4: Append all features into a new matrix called total LPC features
Step 5: Append class label into a new vector called total class
        End Algorithm

The samples of these features are explained in Table IV; the remaining features are reduced to 18, replaced by the total features.

Classification of drowsiness

Only two states, namely, the state of sleepiness (1) and the state of alertness (0), are used to classify binary drowsiness. They employ machine learning technologies, for example, to identify driver drowsiness detection using machine learning methods (SVM, KNN, ANN, DT, etc.). The proposed approach (GWOANN) was developed for this study because it successfully solved problems of non-linearity, high-dimensional classifications, and even classification in speech signals.

TABLE II
(42) Features of MFCCs, DMFCC, and DDMFCC

#	fet1	fet2	fet3	fet4	....	fet13	fet14
1	3.577749	−9.94938	3.016517	0.643449	....	−0.26558	1.048812
2	4.207627	−9.54046	2.638826	0.571662	....	−0.30679	1.009354
3	4.412491	−8.99577	2.50445	0.42222	....	−0.21795	0.994256
4	4.342696	−8.77655	2.650918	0.226735	....	−0.2938	1.037622
5	4.164026	−8.66088	2.47254	0.201869	....	−0.36561	0.933478
.	.	.	.	.	....	.	.
.	.	.	.	.	....	.	.
.	.	.	.	.	....	.	.
868	2.445446	−10.4559	3.48321	0.270007	....	0.203008	−0.411
#	fet1	fet2	fet3	fet4	....	fet13	fet14
1	0.477033	−1.32658	0.402202	0.085793	....	−0.03541	0.139842
2	0.918792	−2.267	0.653495	0.140566	....	−0.06746	0.239462
3	1.247611	−2.81677	0.79891	0.156359	....	−0.07744	0.303424
4	1.420042	−3.03746	0.880373	0.132012	....	−0.09027	0.340026
5	1.423893	−2.95017	0.849688	0.096793	....	−0.10288	0.328155
.	.	.	.	.	....	.	.
.	.	.	.	.	....	.	.
.	.	.	.	.	....	.	.
868	−0.1662	0.245301	−0.14333	−0.11122	....	−0.00524	−0.14969
#	fet1	fet2	fet3	fet4	....	fet13	fet14
1	0.063604	−0.17688	0.053627	0.011439	....	−0.00472	0.018646
2	0.170209	−0.43492	0.127353	0.027322	....	−0.01254	0.045912
3	0.29003	−0.69071	0.198684	0.040624	....	−0.01943	0.073725
4	0.391254	−0.88202	0.254247	0.045468	....	−0.02546	0.096305
5	0.445657	−0.96045	0.276373	0.041216	....	−0.03016	0.105967
.	.	.	.	.	....	.	.
.	.	.	.	.	....	.	.
.	.	.	.	.	....	.	.
868	−0.11303	0.109999	−0.08671	−0.03681	....	−0.00865	−0.06352

TABLE III
(12) Features of LPC

#	fet1	fet2	fet3	fet4	....	fet11	fet12
1	−1.56953	0.375995	0.196997	0.124915	....	0.097282	−0.01144
2	−1.65702	0.435827	0.259976	0.134094	....	0.060285	−0.01609
3	−1.54567	0.317511	0.239578	0.13383	....	0.066661	0.009763
4	−1.81261	0.584168	0.331195	0.122206	....	−0.01225	−0.02677
5	−1.77142	0.546968	0.331306	0.098397	....	0.014741	−0.03164
.	.	.	.	.	....	.	.
.	.	.	.	.	....	.	.
.	.	.	.	.	....	.	.
868	−1.5446	0.514186	0.085213	−0.07562	....	0.093095	−0.02346

TABLE IV
DIMENSION REDUCTION OF FEATURES

#	fet1	fet2	fet3	fet4	….	fet17	fet18
1.	−0.04673	1.484917	0.392288	0.391199	….	0.298653	0.070647
2.	−0.04347	1.498126	0.27122	0.225178	….	0.330448	0.135811
3.	−0.05821	1.513109	0.305602	0.274714	….	0.322942	0.14225
4.	−0.02647	1.520109	0.24549	0.214877	….	0.340287	0.167434
5.	−0.03354	1.499883	0.251718	0.197142	….	0.346273	0.163717
.	.	.	.	.	….	.	.
.	.	.	.	.	….	.	.
.	.	.	.	.	….	.	.
1840	−0.04655	1.638488	0.167963	0.167685	….	0.27592	0.102743

Support vector machine classifier

SVM is a prevalent machine learning technique that produces a high degree of accuracy that was used successfully in many areas (Okfalisa, 2021; Hassan and Alawi, 2017). An efficient classifier changes the "kernel" function used to carry out the classifications in a linear and non-linear manner (Zeng, et al., 2018) (Tao, Sun and Sun, 2018). The fundamental principle of SVM is to create separate hyperplanes for classification in high-dimensional spaces. The hyperplane that is the most distant from each class with a kernel function to the closest training data point achieves optimal separation. SVM has, therefore, been widely used in the field of EOG in recent years (Chen, Wang, and Hua, 2018). In SVM, four standard kernels are used, linear, polynomial, RBF, and sigmoid. The proposed method uses the first three, and specific results are obtained after the classification. Then, recognition (drowsy or non-drowsy) is performed (Salam and Hassan, 2019).

Artificial neural networks (ANNs)

A neural network is a collection of processing units designed to produce human-like outcomes (ADWAN, et al., 2022; Hassan and Hadi, 2016). On the other hand, one subdivision performs its computations and sends them to a second (Yusiong, 2012; Abdulwahed, 2018; Rashid and Abdullah, 2018). The network has three layers: Input, hidden, and output (Huang, Cheng, and Zhang, 2022; Nwobi-Okoye and Ochieze, 2018). ANN models have been developed for a long time. Even when people or other computer algorithms cannot find patterns and trends in complex or confusing data, neural networks can disclose them; ANN has good classification and performance approximation performance (Abed, 2019). As a result, a qualified neural network can be seen as an "Expert" in the knowledge category for which it has received only assessment. This expert can forecast new situations of concern and respond to "what if" queries (Bati and Adam, 2006; Hassan and Jasim, 2010).

The proposed Gray Wolf Optimizer with artificial neural network (GWO-ANN)

In the starting, GWO algorithms are used to train ANN to identify the best weights and biases, as shown in Algorithms 3. An effective back-propagation network is then used to equip the neural network. Finally, check to see if the network has reached the correct error rate or if the number of generations required to finish the algorithm

---

Algorithm (3): Find the best ANN weight using GWO

Input: Extracted speech feature vector (18 features) for normal/abnormal and class label
Output: Best ANN weight

Begin
Step 1: Initialized GWO parameters (pop size and initial weights)
Step 2: Specifying the number of hidden layers of ANN
Step 3: Specifying the same number of each node in each layer in the algorithm ANN classifier
Step 4: For I = 1 to pop size
   - Randomly generating initial weights for all neurons in ANN
   - Convert weights into vector
   - Find the fitness function for all populations that MSE represents
   - Sort MSE of pops in ascending order
   - Find best weights to population (minimum)
Step 4: Find the alpha (α) wolf, beta (β) wolf, and delta (δ) wolf (value and position)
Step 5: Find the average X av of $X_α$, $X_β$, and $X_δ$,
Step 6: Update ($X_α$, $X_β$, and $X_δ$)
Step 7: Depending on the new values ($X_α$, $X_β$, and $X_δ$), update all weights in pops
Step 8: Reorder the weights of neurons
Step 9: Find the MSE of inputs with new weights as a fitness function
Step 10: Sort the fitness function in ascending order
Step 11: Find new best for (α), (β), and (δ) wolves (X_1, X_2, and X_3, respectively)
Step 12: If the new best is less MSE than the old best, replace; otherwise, continue
Step 13: Applying steps from 8 to 13 until Max-iteration No.
Step 14: Return the best ANN weights (optimum weights)
   End

---

has been surpassed. Fig. 1 depicts the proposed method. For the depiction of the ANN, consider using a two-layered network:

$$\sum_{k=1}^{N} w_k f\left(\sum_{i=1}^{m} w_i x_i + b\right) \tag{7}$$

Where, N represents the number of neurons in the hidden layer, w represents the weight of a net, b represents the bias value, and f represents the activation function of each neuron in this example. The following is the mean squared error (MSE):

$$MSE = \frac{\sum_{t}^{z}\left(d^t - y^t\right)^2}{z} \tag{8}$$

If d is the desired output and y is the actual output, z is the number of testing outcomes, T is the goal value, and Y is the projected output, then Equation (3) indicates a superior model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(T_i - Y_i\right)^2} \tag{9}$$

IV. EXPERIMENTS RESULTS AND DISCUSSION

The recorded dataset has been used in current experiments to detect driver drowsiness. The following stages are included in each experiment:
- Pre-processing.
- Features extraction.

- Normalization features.
- Classification by GWOANN.

Discovering the parameters that work well for each classifier is critical. There were a set of predetermined values used. Three trials were run for each classifier, with a different training-testing proportion. These were 90%-10%; 80-20%; and 70-30%. This research tested two classification categories (drowsy or not).

A. Results for SVM classifier

The SVM classifier score is expected to have higher when the following parameters are used, outlier = (0.001), K-Fold = (2), TrPer = (90%, 80%, and 70%), counter = 1000, and kernel = Linear. Table V shows the results of applying the SVM classifier. It can be noticed that the SVM classifier achieved the best accuracy when 90-10% training-testing percentages.

Table VI where shows that the accuracy decreased when using the same parameters outlier = (0.001), K-Fold = (2), TrPer = (90%, 80%, and 70%), Counter = 1000, but the kernel is used polynomial.

B. Using ANN Classifier

It is used to classify the driver's sleepiness using MLP. Unsupervised learning is used to build an ANN that can classify driver drowsiness. The parameters used in Table VII determine the standalone ANN classifier.

Tables VIII and IX illustrate the results of utilizing a standalone ANN classifier with 90-10%, 80-20%, and 70-30% training-testing percentages. The ANN classifier had the best accuracy when the training-testing percentages were 70-30%. Table III shows that the accuracy percentages have reduced to 90-10% and 80-20%.

C. Using GWO-ANN

GWO swarm is used to find an accurate weights of ANN. The suggested method involves first training the network with starting weights and biases and then updating the findings. Hence, the global optima back-propagation is sped up. The suggested technique includes weights and biases. It is based on RMSE. This classification is more accurate than the traditional ANN classifier, as shown in Table X.

Tables XI and XII demonstrate the outcomes of implementing the proposed hybrid approach (GWO-ANN) with 90-10%, 80-20%, and 70-30% training-testing

TABLE VII
PARAMETERS USED FOR A STANDALONE ANN MODEL

Parameters	Value
Max iteration	15,000
Number of (neurons in i/p layer)	18
Number of (hidden layer)	4
Number of (neurons in each hidden layer)	(13, 9, 7, 5) and (30, 20, 13, 7)
Number of (neurons in o/p layer)	2

TABLE V
ACCURACY OF THE SVM MODEL TRAINING DATA (90%, 80% AND 70%) WITH KERNEL OF LINEAR

Run number	Training 90%, testing 10%	Training 80%, testing 20%	Training 70%, testing 30%
	Acc.	Acc.	Acc.
1	72.83%	73.37%	71.56%
2	68.48%	70.65%	74.28%
3	69.57%	66.58%	70.83%
4	71.20%	69.84%	71.38%
5	71.74%	71.20%	69.38%
6	71.20%	69.02%	72.28%
7	72.83%	77.17%	69.57%
8	73.91%	72.28%	74.64%
9	70.11%	71.74%	71.92%
10	76.46%	71.20%	71.20%

TABLE VIII
THE RESULTS PROVIDED BY STANDALONE ANN IN CASE THE NO. OF NEURONS IN THE HIDDEN LAYER ARE (13, 9, 7, 5)

No. of hidden of neurons	13, 9, 7, 5		
Run number	Acc. (90-10%)	Acc. (80-20%)	Acc. (70-30%)
1	80.43%	80.98%	77.72%
2	86.41%	80.98%	78.80%
3	83.70%	82.07%	80.62%
4	78.26%	80.16%	78.99%
5	86.41%	84.24%	78.80%
6	84.78%	77.45%	77.36%
7	84.24%	81.79%	79.17%
8	81.52%	82.88%	78.44%
9	79.35%	81.52%	82.61%
10	79.89%	81.25%	81.16%

TABLE VI
ACCURACY OF THE SVM MODEL TRAINING DATA (90%, 80% AND 70%) WITH KERNEL OF POLYNOMIAL

Run number	Training 90%, testing 10%	Training 80%, testing 20%	Training 70%, testing 30%
	Acc.	Acc.	Acc.
1	67.39%	65.49%	65.40%
2	60.33%	68.21%	64.49%
3	60.33%	59.51%	67.39%
4	68.48%	62.23%	66.67%
5	60.33%	66.30%	64.86%
6	65.76%	63.86%	64.49%
7	65.22%	68.75%	59.96%
8	59.24%	68.21%	69.20%
9	66.30%	64.40%	67.93%
10	63.59%	66.30%	58.88%

TABLE IX
THE RESULTS PROVIDED BY STANDALONE ANN IN CASE THE NO. OF NEURONS IN THE HIDDEN LAYER ARE (30,20,13,7)

No. of hidden of neurons	30,20,13,7		
Run number	Acc. (90-10%)	Acc. (80-20%)	Acc. (70-30%)
1	86.41%	82.61%	83.51%
2	86.96%	85.05%	79.35%
3	86.96%	85.60%	80.62%
4	83.70%	82.07%	82.25%
5	85.33%	79.08%	79.89%
6	82.07%	83.42%	83.33%
7	88.04%	86.14%	80.80%
8	83.15%	83.70%	84.24%
9	85.33%	85.05%	83.70%
10	84.78%	82.88%	79.71%

percentages. The GWO-ANN approach, once again, achieved balanced accuracy in all percentages of training and testing. To avoid overfitting, the algorithm's lower number of parameters was used; it achieved higher simplicity and reasonably lowered the risk of overfitting. They observed that the experiment determined the number of neurons in the hidden layers.

Tables XIII show the best, worst, mean, and standard deviation of three classifiers employing 90-10%, 80-20%, and 70-30% training-testing percentages with the number of neurons in hidden layers (13, 9, 7, 5) and (30,20,13,7). In 90-10% training-testing percentages, the proposed method's STD value (0.01650) was the lowest compared to the STD value of the ANN classifier (0.01793). On the other hand, compared to the ANN classifier and SVM, the value of best, worst, and mean (93.48%, 87.50%, and 90.05%) achieved the highest value. Thus, the proposed approach (GWO-ANN) is stable and specific.

In addition, all classifications considered in this work have reasonable precision, but the GWO-ANN classification reaches 90.05% with the highest rating accuracy. Therefore, the proposed solution may be essential for future studies or future "systems" vehicles as reference work. Table XIV

TABLE X
PARAMETERS BASED ON GWO AND ANN

Parameters based on GWO	
Parameters	Value
Iteration no.	100
Population size	100
Parameters based on ANN	
Parameters	Value
Number of (neurons in i/p layer)	18
Number of (hidden layer)	4
Number of (neurons in each hidden layer)	(13, 9, 7, 5) and (30, 20, 13, 7)
Number of (neurons in o/p layer)	2
Max iteration	15000

TABLE XI
THE RESULTS PROVIDED BY GWO-ANN IN CASE THE NO. OF NEURONS IN THE HIDDEN LAYER ARE (13, 9, 7, 5)

No. of hidden of neurons	13, 9, 7, 5		
Run number	Acc. (90-10%)	Acc. (80-20%)	Acc. (70-30%)
1	85.33%	87.50%	84.06%
2	87.50%	85.87%	85.51%
3	86.41%	86.14%	85.51%
4	88.04%	86.68%	86.41%
5	88.04%	87.23%	84.42%
6	85.87%	85.33%	82.61%
7	86.96%	85.33%	84.42%
8	86.96%	84.78%	86.96%
9	88.04%	81.25%	86.78%
10	86.41%	87.50%	84.06%

TABLE XII
THE RESULTS PROVIDED BY GWO-ANN IN CASE THE NO. OF NEURONS IN THE HIDDEN LAYER ARE (30,20,13,7)

No. of hidden of neurons	30,20,13,7		
Run number	Acc. (90-10%)	Acc. (80-20%)	Acc. (70-30%)
1	89.67%	90.22%	87.50%
2	90.76%	87.50%	87.50%
3	89.13%	88.59%	87.14%
4	88.04%	85.60%	87.50%
5	91.30%	89.13%	86.59%
6	89.67%	87.77%	86.59%
7	93.48%	88.32%	88.41%
8	91.30%	89.67%	89.31%
9	89.67%	87.23%	85.69%
10	87.50%	88.04%	87.32%

TABLE XIII
THE RESULTS OF CLASSIFIERS MODELS

No. of hidden of neurons	Name classifier	(90% training, 10% testing)			
		Best Acc.	Worst Acc.	Mean	STD.
13, 9, 7, 5	ANN	86.41%	78.26%	82.50%	0.02832
	GWO-ANN	88.04%	85.33%	86.96%	0.00908
30,20,13,7	ANN	88.04%	82.07%	85.27%	0.01793
	GWO-ANN	93.48%	87.50%	90.05%	0.01650
------------	SVM	76.50%	68.50%	71.80%	0.02188
No. of hidden of neurons	Name classifier	(80% training, 20% testing)			
		Best Acc.	Best Acc.	Best Acc.	Best Acc.
13, 9, 7, 5	ANN	84.24%	77.45%	81.33%	0.01684
	GWO-ANN	87.50%	81.25%	85.76%	0.01758
30,20,13,7	ANN	86.14%	79.08%	83.56%	0.01974
	GWO-ANN	90.22%	85.60%	88.21%	0.01252
------------	SVM	77.20%	66.60%	71.30%	0.02645
No. of hidden of neurons	Name classifier	(70% training, 30% testing)			
		Best Acc.	Worst Acc.	Mean	STD.
13, 9, 7, 5	ANN	82.61%	77.36%	79.37%	0.01541
	GWO-ANN	86.96%	82.61%	85.07%	0.01327
30,20,13,7	ANN	84.24%	79.35%	81.74%	0.01772
	GWO-ANN	89.31%	85.69%	87.36%	0.00948
------------	SVM	74.60%	69.40%	71.70%	0.01635

TABLE XIV
IMPROVEMENT RATE OF GWO-ANN OVER STANDALONE ANN

Classifier	90-10%		
	IR of best	IR of worst	IR of mean
Hidden neurons in ANN (13, 9, 7, 5)	2%	9%	5%
Hidden neurons in ANN (30,20,13,7)	6%	7%	6%
SVM	22%	28%	25%
Classifier	80-20%		
	IR of best	IR of worst	IR of mean
Hidden neurons in ANN (13, 9, 7, 5)	4%	5%	6%
Hidden neurons in ANN (30,20,13,7)	5%	8%	6%
SVM	17%	29%	24%
Classifier	70-30%		
	IR of best	IR of worst	IR of mean
Hidden neurons in ANN (13, 9, 7, 5)	5%	7%	7%
Hidden neurons in ANN (30,20,13,7)	6%	8%	7%
SVM	20%	23%	22%

Method	Accuracy %
Krajewski, Batliner and Golz, 2009 SVM	86.1
Ooi, et al., 2016 SVM	85
Greco, et al., 2019 SVM-RFE	+11.64
Martin, et al., 2021 PCA+LR	74.2
Proposed ANN	85.27
Proposed GWO-ANN	90.05

compares the proposed method's improvement rate with the ANN classifier in 90-10%, 80-20%, and 70-30% training-testing percentages.

$$Improvement\ rate = \frac{A1 - A2}{A2} \qquad (10)$$

Where, A1 represents the accuracy of the proposed algorithm and A2 represents the accuracy of another classifier.

D. Comparative analysis of results

The results of the ANN and GWOANN algorithms are compared to the findings of previous works utilizing the recorded dataset for each work. Table XV displays the outcomes of this evaluation. The proposed paper's results were very accurate compared to those of other papers.

V. CONCLUSION

This paper presents a system for drowsiness detection of vehicle drivers based on voice analysis by monitoring the alertness level periodically. The drowsiness detection is less accurate in a noisy environment, and it is a challenge to get high accuracy in detecting the abnormal status. The performance of the ANN classification method is enhanced by the swarm intelligent GWO. The method is based on speech signals for drowsiness detection through voice. The results of the experiments showed that efficiency is achieved when using a neural network with (4) hidden layer networks of (13, 9, 7, 5) and (30,20,13,7) neurons. The GWOANN technique had an 86.96% and 90.05% accuracy. The method is flexible in combining with other detection methods to build an integrated system for actual use. A discrete wavelet transform could be a future direction to consider.

REFERENCES

Abdul-Hassan, A.K. and Hadi, I.H., 2020. A proposed authentication approach based on voice and fuzzy logic. In: *Recent Trends in Intelligent Computing, Communication and Devices.* Springer, Berlin, Heidelberg.

Abdulwahed, M.N., 2018. Analysis of image noise reduction using neural network. *Engineering and Technology Journal,* 36, pp.76-87.

Abed, I.S., 2019. Lung Cancer Detection from X-ray images by combined Backpropagation Neural Network and PCA. *Engineering and Technology Journal,* 37, pp.166-171.

Adwan, I., Milad, A., Abdullah, N.H., Widyatmoko, I., Mubaraki, M., Yazid, M.R. and Yusoff, N.I., 2022. Predicting asphalt pavement temperature by using neural network and multiple linear regression approach in the Eastern Mediterranean region. *Journal of Engineering Science and Technology,* 17, pp.0015-0032.

Alzu'bi, H.S., Al-Nuaimy, W. and Al-Zubi, N.S., 2013. EEG-based driver fatigue detection. In: *2013 Sixth International Conference on Developments in eSystems Engineering.* IEEE, New Jersey, United States. pp.111-114.

Badr, A.A. and Abdul-Hassan, A.K., 2020. A review on voice-based interface for human-robot interaction. *Iraqi Journal for Electrical and Electronic Engineering,* 16, pp.91-102.

Bati, A.F. and Adam, N.E., 2006. Hybrid neuro-genetic based controller of power system. *Iraqi Journal of Computers, Communication, Control and Systems Engineering,* 6, pp.1-115.

Chen, J., Wang, H. and Hua, C., 2018. Assessment of driver drowsiness using electroencephalogram signals based on multiple functional brain networks. *International Journal of Psychophysiology,* 133, pp.120-130.

Dasgupta, A., Kabi, B., George, A., Happy, S. and Routray, A., 2015. A drowsiness detection scheme based on fusion of voice and vision cues. *arXiv preprint arXiv:1509.*

Gamit, M.R. and Dhameliya, K., 2015. Isolated words recognition using MFCC, LPC and neural network. *International Journal of Research in Engineering and Technology,* 4, pp.146-149.

Greco, A., Marzi, C., Lanata, A., Scilingo, E.P. and Vanello, N., 2019. Combining electrodermal activity and speech analysis towards a more accurate emotion recognition system. In: *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* IEEE, New Jersey, United States, pp.229-232.

Hassan, A. and Hadi, M., 2016. Sense-based information retrieval using artificial bee colony approach. *International Journal of Applied Engineering Research,* 11, pp.8708-8713.

Hassan, A.K. and Alawi, M., 2017. Proposed handwriting Arabic words classification based on discrete wavelet transform and support vector machine. *Iraqi Journal of Science,* 58, pp.1159-1168.

Hassan, A.K. and Jasim, S.S., 2010. Integrating neural network with genetic algorithms for the classification plant disease. *Engineering and Technology Journal,* 28, pp.686-702.

Hassan, A.K. and Mohammed, S.N., 2020. A novel facial emotion recognition scheme based on graph mining. *Defence Technology,* 16, pp.1062-1072.

Heidari, A.A. and Pahlavani, P., 2017. An efficient modified grey wolf optimizer with Lévy flight for optimization tasks. *Applied Soft Computing,* 60, pp.115-134.

Huang, X., Cheng, C. and Zhang, X.B., 2022. Machine learning and numerical investigation on drag reduction of underwater serial multi-projectiles. *Defence Technology,* 18, pp.229-237.

Huo, X.Q., Zheng, W.L. and Lu, B.L., 2016. Driving fatigue detection with a fusion of EEG and forehead EOG. In: *2016 International Joint Conference on Neural Networks (IJCNN),* IEEE, New Jersey, United States. pp.897-904.

Jasim, S.S. and Hassan, A.K., 2022. Modern drowsiness detection techniques: A review. *International Journal of Electrical and Computer Engineering,* 12, pp.2986-2995.

Krajewski, J., Batliner, A. and Golz, M., 2009. Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. *Behavior Research Methods,* 41, pp.795-804.

Martin, V.P., Rouas, J.L., Boyer, F. and Philip, P., 2021. Automatic Speech Recognition systems errors for accident-prone sleepiness detection through voice. In: *2021 29th European Signal Processing Conference (EUSIPCO),* IEEE, New Jersey, United States. pp.541-545.

Nwobi-Okoye, C.C. and Ochieze, B.Q., 2018. Age hardening process modelling and optimization of aluminium alloy A356/Cow horn particulate composite for brake drum application using RSM, ANN and simulated annealing. *Defence Technology,* 14, pp.336-345.

Okfalisa, Handayani, L., Juwita, P.D., Affandes, M., Fauzi, S.S. and Saktioto., 2021. Coronary heart disease using support vector machine. *Journal of Engineering Science and Technology,* 16, p.16.

Ooi, J.S., Ahmad, S.A., Chong, Y.Z., Ali, S.H., Ai, G. and Wagatsuma, H., 2016. Driver emotion recognition framework based on electrodermal activity measurements during simulated driving conditions. In: *2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, IEEE, New Jersey, United States. pp.365-369.

Pane, E.S., Hendrawan, M.A., Wibawa, A.D. and Purnomo, M.H., 2017. Identifying rules for electroencephalograph (EEG) emotion recognition and classification. In: *2017 5th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME),* IEEE, New Jersey, United States. pp.167-172.

Rashid, T.A. and Abdullah, S.M., 2018. A hybrid of an artificial bee colony, genetic algorithm, and neural network for diabetic Mellitus diagnosing. *ARO-The Scientific Journal of Koya University,* 6, pp.55-64.

Salam, M. and Hassan, A.A., 2019. Offline isolated Arabic handwriting character recognition system based on SVM. *International Arab Journal of Information Technology,* 16, pp.467-472.

Tao, P., Sun, Z. and Sun, Z., 2018. An improved intrusion detection algorithm based on GA and SVM. *IEEE Access,* 6, pp.13624-13631.

Wankhade, S.B. and Kharat, P.A., 2017. A novel two-tier classifier based on K-nearest neighbour and neural network classifier for emotion recognition using EEG signals. *International Journal of Latest Technology in Engineering,*

Management and Applied Science (IJLTEMAS), 6, p.7.

Xu, L., Wang, H., Lin, W., Gulliver, T.A. and Le, K.N., 2019. GWO-BP neural network-based OP performance prediction for mobile multiuser communication networks. *IEEE Access,* 7, pp.152690-152700.

Yoshida, R., Nakayama, T., Ogitsu, T., Takemura, H., Mizoguchi, H., Yamaguchi, E., Inagaki, S., Takeda, Y., Namatame, M., Sugimoto, M. and Kusunoki, F., 2014. Feasibility study on estimating visual attention using electrodermal activity. *Proceedings of the International Conference on Sensing Technology*, 2014, pp.589-595.

Yu, X., Wang, S.H. and Zhang, Y.D., 2021. CGNet: A graph-knowledge embedded convolutional neural network for detection of pneumonia. *Information Processing and Management,* 58, p.102411.

Yusiong, J.P., 2012. Optimizing artificial neural networks using cat swarm optimization algorithm. *International Journal of Intelligent Systems and Applications,* 5, p.69.

Zeng, N., Qiu, H., Wang, Z., Liu, W., Zhang, H. and Li, Y., 2018. A new switching-delayed-PSO-based optimized SVM algorithm for diagnosis of Alzheimer's disease. *Neurocomputing,* 320, pp.195-202.

Zhang, F., Su, J., Geng, L. and Xiao, Z., 2017, Driver fatigue detection based on eye state recognition. In: *2017 International Conference on Machine Vision and Information Technology (CMVIT)*, IEEE, New Jersey, United States. pp.105-110.

Zhang, L., 2019. *Analysis of Machine Learning Algorithms for the Recognition of Basic Emotions: Data Mining of Psychophysiological Sensor Information*. Ulm Universität, Germany.

Modern and Lightweight Component-based Symmetric Cipher Algorithms: A Review

Samar A. Qassir[1], Methaq T. Gaata[1], and Ahmed T. Sadiq[2]

[1]Department of Computer Science, College of Science, Mustansiriyah University,
Baghdad, Iraq

[2]Department of Computer Science, University of Technology - Iraq,
Baghdad, Iraq

Abstract—Information security, being one of the corner stones of network and communication technology, has been evolving tremendously to cope with the parallel evolution of network security threats. Hence, cipher algorithms in the core of the information security process have more crucial role to play here, with continuous need for new and unorthodox designs to meet the increasing complexity of the applications environment that keep offering challenges to the current existing cipher algorithms. The aim of this review is to present symmetric cipher main components, the modern and lightweight symmetric cipher algorithms design based on the components that utilized in cipher design, highlighting the effect of each component and the essential component among them, how the modern cipher has modified to lightweight cipher by reducing the number and size of these components, clarify how these components give the strength for symmetric cipher versus asymmetric of cipher. Moreover, a new classification of cryptography algorithms to four categories based on four factors is presented. Finally, some modern and lightweight symmetric cipher algorithms are selected, presented with a comparison between them according to their components by taking into considerations the components impact on security, performance, and resource requirements.

Index Terms—Information security, Lightweight Symmetric Cipher, Modern Symmetric Cipher, Symmetric cipher components.

I. INTRODUCTION

The collection of the steps and processes designed to protect and secure data which is one of the most important commodities of this era from any unauthorized access or alterations whether in the more static storage status or in the more dynamic transmission phase between one location to another, is defined collectively as Information security (abbreviated to "InfoSec"). Because knowledge has become one of the most valuable commodities in the 21st century,

efforts to keep data secure have become increasingly vital (Soomro, Shah, and Ahmed, 2016).

The phrase "information security" is usually used to refer to data secrecy, this confidentiality can be achieved with help of cryptography. Cryptography as a term came from the Greek word "kryptos" which defines "Hidden" or "Secrets" (Anand, et al., 2016; Pachghare, 2019). Cryptography is skill/discipline of transforming a plain/original data into coded data (Lakhtaria, 2011) and again reconverting that plain text into its original data, providing protection for this data from unauthorized access and manipulation both in storage and transfer circumstances. Fig. 1 shows cryptography process which involves two processes: Encryption and decryption (Stallings, et al., 2012; Bagane and Sirbi, 2021).

Plain text refers to the original information, whereas cipher text refers to the coded form. Cipher is an algorithm for converting original text to coded text. Key is information which is used in cipher known only to sender/recipient to encode or decode the information (Qadir and Nurhayat, 2019; Forouzan and Mukhopadhyay, 2015). Encryption which is also known as an encipherment is a process of converting original data into coded one. Nowadays, encryption is available by default often without the user even being aware of it. Decryption which is also known as decipherment can be defined as the recovery system of cipher text from plain text. The cryptography can be summarized as learning of encryption theory or techniques (Stallings, et al., 2012; Sharma, et al., 2022).

This paper presents a review of cipher categories, symmetric main components units, explanation for their properties and effect on security, and the symmetric advantages versus asymmetric cipher category, also comparison among the two types of the symmetric cipher category, in addition to gathering and discussing some common types of stream, block, and lightweight ciphers algorithm.

Due to the importance of encryption algorithms in the field of information security, many of researches and studies are being conducted and presented. Certain of these studies have been done for performance evaluation (Sallam and Beheshti, 2018), cipher design analysis (Hamza and Kumar, 2020), and summarization the various weaknesses of a particular
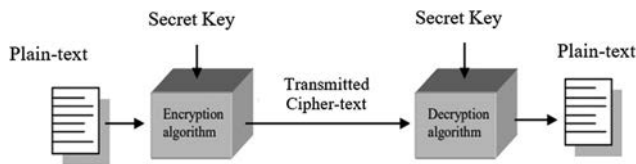
Fig. 1. General Model of Cryptography Process.

algorithm (Brahmjit, 2015). In the next section of this review, some of these studies will explain. For this review, which represent the first analysis study for the components that used in symmetric cipher algorithms design, the new contribution explains through these three points: First, a new classification of encryption is presented. This comprehensive classification is the first of its kind that give the insight idea about all types of encryption according to their chronological sequence and components, which shows how the development of encryption algorithms done depended mainly on the types and size of components used in algorithm design. Second, analysis of the design of modern and lightweight symmetric ciphers based on their main components used, focusing on the impact of each component on cipher design; the component that appears primarily in symmetric cipher design shows how these components give the properties of symmetric versus asymmetric cipher characteristics. Third, by selecting and analyzing 13 symmetric cipher algorithms of both types (stream and block), it was shown the lightweight that used today in smart devices they result from a reduction in the number and in the size of essential cipher components, which shows the importance of reviewing and analyzing all cipher components and knowing the basic one, in addition of its impact on the strength of the cipher algorithm design. The objectives of this analytical study for the basic symmetric cipher components are giving a more insight and comprehensive to developers in the field of encryption for designing new cipher algorithms.

The next sections of this review presents related work and new taxonomy of cryptography categories respectively; whereas section 4 discusses the theoretical background that demonstrates the categories of symmetric key cipher; additionally, the main components used in symmetric key cipher's types have explained. Some selected modern and lightweight cipher algorithms are explained in sections 5 and 6, whereas the last 2 sections (7 and 8) involve the discussion and conclusion.

II. Related Work

Sharma, et al. (2022) introduced a review study about the concept of cryptography, beginning with its history, definitions, security functions, and historical methods for symmetric and asymmetric encryption, as well as homomorphic encryption, and the Goldwasser_Micali encryption scheme. This investigation focused on several different kinds of cryptography computations, including AES, DES, 3DES, IDEA, RSA, ECC, homomorphic, and blowfish. This study has demonstrated that cryptography plays a crucial

role in achieving the main goals of security objectives, such as acceptance, rectification, safety, and repudiation. Cryptographic calculations are performed to accomplish these goals. To provide reliable, amazing, and strong association and data protection, cryptography is strongly motivated.

A comparative study between conventional cryptographic algorithms and evolutionary algorithms for cyber security was presented by Bagane and Sirbi (2021). Based on Darwinian evolutionary theories of natural selection and genetics, genetic algorithm (GA) is organized as random search algorithm. Moreover, the skill of creating secret codes is known as cryptography. This study comes to the conclusion that using a genetic algorithm-based cryptosystem is more safe and impenetrable than a conventional cryptosystem. The recommendation was to experiment with several sophisticated ciphers, such as AES and DES, to improve the performance of genetic algorithm.

Hamza and Kumar (2020) proposed a review for three algorithms DES, AES, and RSA. This study evaluated and listed the aforementioned algorithms in a sequential order, making clear how they relate to one another. For instance, how symmetric and asymmetric algorithms, those with secret keys, and those with key pairs relate to one another by examining the DES, AES, and RSA algorithms, they were able to highlight the advantages and disadvantages of both symmetric and asymmetric encryption techniques. This study might reveal some of the areas for future scholarly research in the area of cryptography.

The evolution of stream ciphers, their classification, and the design concepts of this cipher type were reviewed by Jiao, Hao and Feng (2020). In addition, a brief analysis of their advantages and shortages was done to stimulate additional study on this cipher type's implementation and security. This study shown that this cipher design focuses primarily on the cipher structure and the fundamental operations used, which must ensure the security that resists all categories of current attacks.

A survey of recent developments in the field of lightweight cryptographic algorithms was given by George Hatzivasilis, et al. (2018). Hardware and software designs for symmetric key block ciphers'' lightweight implementations have been studied. Three hundred and sixty implementations and 52 block ciphers were compared in terms of their security, cost, and adaptability to various types of embedded devices. The most significant cryptanalysis relating to these ciphers was also mentioned. The research offered, in the authors' opinion, will help designers create reliable systems and architectures, enabling a safe transition to this new world and the Internet of Things.

III. Cipher Classification

The cryptography fundamentals can be categorized in four main ways/categories according to several factors including number of keys, type of cipher operation, way of processing plain text, and, finally, the chronological order; as shown in Fig. 2. The first categorization method depends on the number
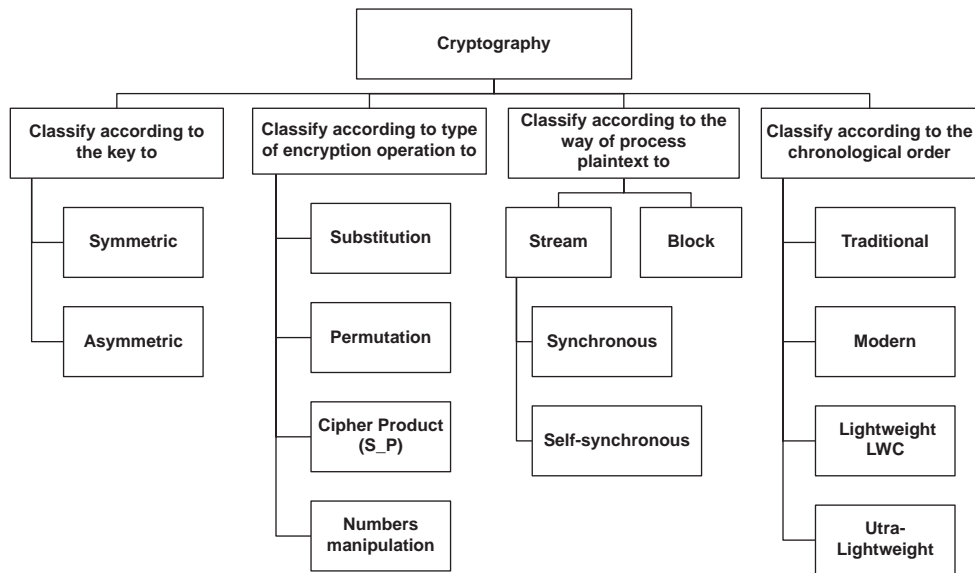
Fig. 2. The cipher classifications.

of keys: Symmetric versus asymmetric (Wahid, et al., 2018), whereas the second categorization method involves the type of cipher operation: Substitution (S-Box), permutation (P-Box), cipher product (S-P), and numbers manipulation. The third categorization method takes into consideration the way of processing plain text into two main categories: Stream (which is further sub-categorized into synchronous and self-synchronous) and block ciphers (Mewada, Sharma and Gautam, 2016). The last categorization method depends on the chronological order: Beginning with the traditional, into modern, the most recent lightweight and ultra-lightweight (Hasan, et al., 2021).

The cipher algorithm can involve more than 1 of the above-mentioned categorization methods; for example, the algorithm may be symmetric key using (s-p) in encryption, processing plain text as blocks, and lightweight type (Chiadighikaobi and Katuk, 2021).

IV. Basic Principles

In this section, the basic principles of symmetric cipher of its two types (stream and block) are covered. The main components that used in symmetric cipher algorithm design are explained in details and finally the different between symmetric ciphers algorithms according to their chronological order are presented.

A. Symmetric Cipher

As explained in the previous section, there are two cipher types based on number of key used: Symmetric and asymmetric. Table I shows direct comparison between these two categories. On a basis of processing way of plain text, a symmetric encryption schemes classified into main categories: The first one is stream cipher and the second one is block cipher (Hussain and Shah, 2013), (Forouzan and Mukhopadhyay, 2015). The first cipher, as explained in Fig. 3, is the most important encryption system. It is generally used for its speed
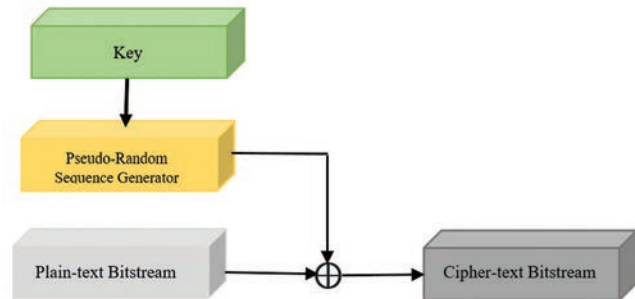
Fig. 3. Stream cipher encryption process.

and simplicity of implementation in hardware, they have less complex hardware circuitry. The characteristic feature of this cypher system is the continue change of the keys for every bit of the plain text resulting in the production of different cipher text even for plain text of repetitive blocks, this becomes very applicable and handy for unknown length plain text like a secure wireless connection, limited buffering, when individual processing of characters is a must on reception, in addition to, military applications, and in strategic sectors.

Due to merit of having limited or no error propagation, and where transmission errors are highly probable, this stream cipher has a prominent advantage. It is classified into synchronous stream ciphers (SSC) and self-synchronizing stream cipher (SSSC) where due to design problems, SSSC is less used in practice (Sharma, et al., 2022; Jindal and Singh, 2015).

Random key streaming that ensures the computational security of the cipher is the major selling point for this cipher system where its solidity and robustness against attacks and unauthorized access is based on complex keystream created to secure text and images transmitted digitally through open networks such as phones and emails. This defines its efficacy as a major property of this cipher system.

Many studies aiming to improve the stream ciphers took interest on the randomness and complexity of the keystream (Bagane and Sirbi, 2021).

The second category, based on how plain text is processed, is block type, in which the original text is divided into chunks of bits at a time (Hussain and Shah, 2013). High versatility, high diffusion, and robust resistance to attempts of concealed tampering attempts, and possessing very similar and identical encryption and decryption methods, all those merits grant this cipher system its advantages making its implementation with lesser resources is a genuine possibility. However, speaking about advantages cannot be completed without mentioning possible disadvantages such as slower encryption speed as a result of need to capture the entire block for encryption/decryption. Another negative point is the possibility of breeding errors due to the fact that a mistake in just one symbol could alter the whole block. As shown in Fig. 4, this cipher contains a set number of bits and multiple stages of transformation that is defined by a symmetric key (De Canniere, Biryukov and Preneel, 2006). Table II shows a comparison of block and stream cipher types.

B. Components Used in Symmetric Cipher

There are many properties of symmetric cipher such as its faster speed compared with asymmetric cipher, the need to use minimal resources only, suitability for today's constrained environmental resources such as low chip area or low-power supply, and being ideal for most of today's applications and for encoding large amounts of data.

The reason behind all those above-mentioned properties of the symmetric cipher is its design which is based on components that are different from the asymmetric cipher. Whereby, the design of symmetric cipher is based on substitution, permutation, XOR, and many other components which are much faster compared to the asymmetric cipher's design due to the application of mathematical functions to numbers (Kumar, Suneetha, and Chandrasekhar, 2012; Szaban and Seredynski, 2011).

Fig. 5 shows that illustration of the components of the two types of the symmetric cipher is shown in addition to

TABLE I
COMPARISON FOR THE TWO MAIN CIPHERS CATEGORIES SYMMETRIC AND ASYMMETRIC (DUTTA, GHOSH AND BAYOUMI, 2019)

Criteria	Symmetric – Cipher	Asymmetric –Cipher
No. of key	Sharing secrecy (one key)	Personal secrecy (two key)
Operation type	Based on substation, permutation	Numbers manipulation
Algorithm speed	Faster	Requires more time
Main ciphering aim	Provides confidentiality and authentication	Provides confidentiality
Algorithm utilization	For ciphering huge amounts of data, files and also for communication paths	For all key operations like ciphering, distributing
Algorithm design based on	Arithmetic and logic components	Mathematical computation
Hardware implementation difficulty	Less difficulty	More difficulty
For lightweight applications	Used a lot	Less used
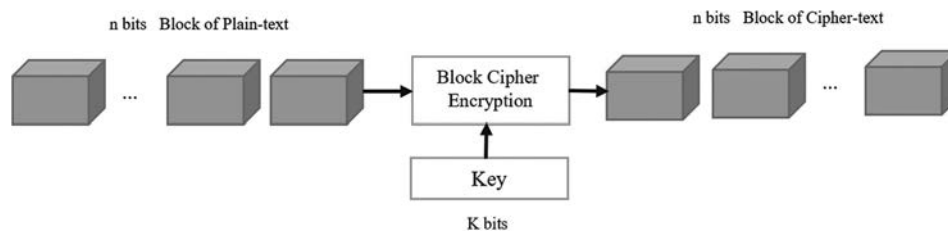
Fig. 4. Simple diagram for block cipher.
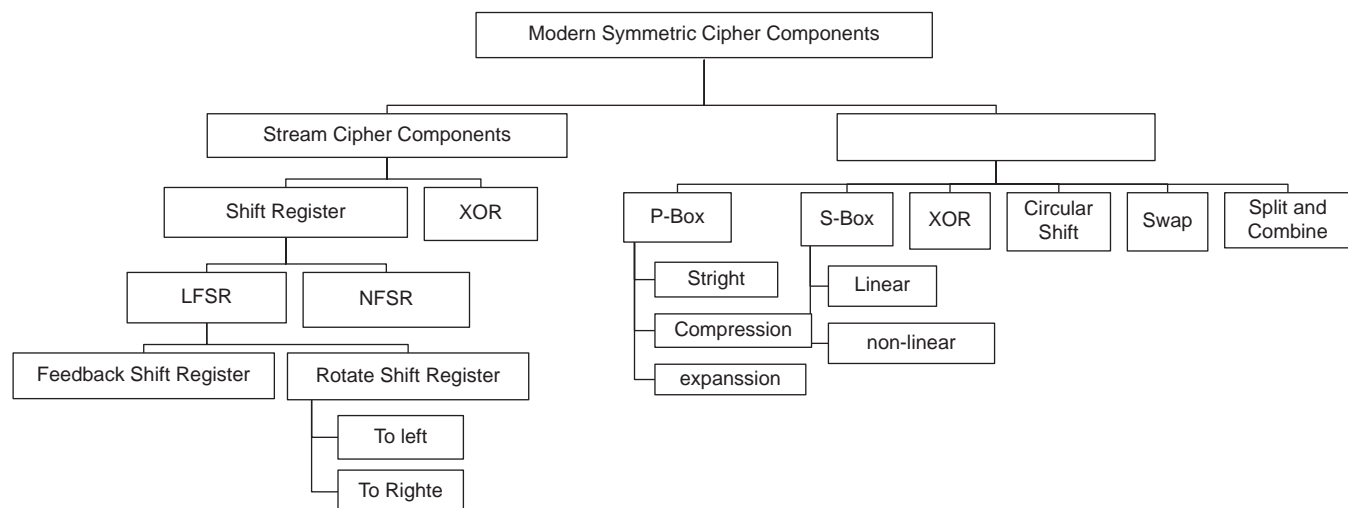
Fig. 5. Modern symmetric cipher components.

Criteria	Block type	Stream type
The length of data	Big (64,128,256) bit	Little (1) bit
The length of key	Fixed	Variable
Count of rounds	Big	-
Usage of resources	Big	Little
Implementations speed	Less than stream	More rapid and easier
Security	Less than stream	Maximum security
Error-resistant robustness	Little	Big
Applications	For big message	For small message
Shannon principles	Confusion and diffusion	Only confusion
Algorithm design based on	S-Box, P-Box, XOR, swap, circular shift, split, and combined	All types of shift registers and XOR

Fig. 6. S-Box in DES algorithm.

Fig. 7. The three types of P-Boxes.

the difference between the components used in each type of symmetric cipher: Stream and block. In the modern stream symmetric cipher, the algorithm is based on two main components only; feedback shift register (FSR) and XOR. They effectively utilize linear and non-linear feedback shift register (LFSR and NFSR). Whereas in modern block symmetric cipher, the algorithms are very versatile as a result of being designed to provide properties of modern cipher, such as diffusion and confusion by being constructed not as a single unit.

This modern cipher is made of a combination of different components such as S-Box, P-Box, Exclusive-Or, circular shift, swap, split, and combine. Some of these components will be explained in the next subsections (Easttom, 2021), (Bardis, Markovskyy and Andrikou, 2004).

1. S-Box and P-Box

Substitution box (S-Box) and permutation box (P-Box) are crucial components of every modern block cipher. The prime purpose of an S-Box is to prevent the output from being easily transformed back into the input by creating confusion between the cipher text and the secret key, this purpose of this component making it as the heart of every block cipher cryptosystem (Hussain and Shah, 2013). P-Box, on the other hand, is responsible for diffusion. To further maximize the difficulty of analyzing the cipher and increase its robustness, modern block ciphers actually use several different S-Boxes. Fig. 6 illustrates the S-Box in DES algorithm.

According to the number of output bits compared to those of the input bits, P-Boxes are classified as compression (less), expansion (greater), and straight (equal), where only the later ones are invertible (Chugunkov, et al., 2020), as shown in Fig. 7.

2. Exclusive_Or

Ciphers that are more sophisticated have a very common component operator considered the to be the mysterious recipe behind modern encryption, known as Exclusive_Or "XOR," where a cipher text is created by combining a key with a plain text through this operator, while restoring the original plain text by XOR-ing to same key to the cipher text. This operator is based on what is known as a truth table, in which 0 represents "false" and 1 represents "true."
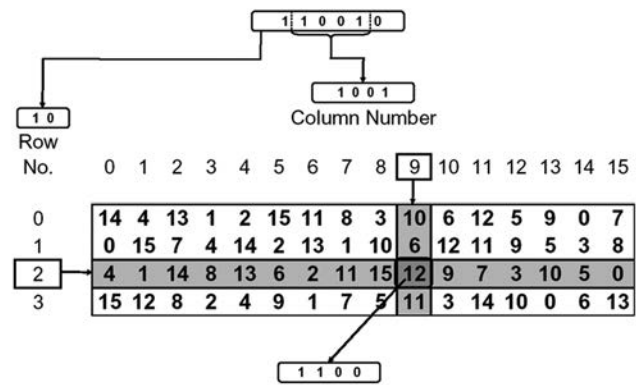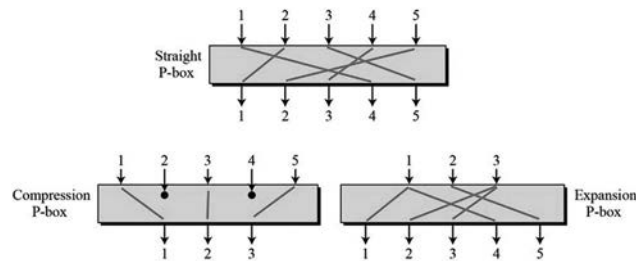
When one of two bits is 1, the other two bits are regarded true (or 1), however, when both bits are 0 or 1, they XOR to 0. For a really random key bit, this operation creates an ideal equilibrium in which the cipher text outcome is equally likely to be 0 or 1 for a given plain text input. This ideal/ perfect balance is the main reason why this operator is so useful in cryptography. In this section, the five essential properties of XOR (Dreier, et al., 2018) that make use of it are presented in Table III:

3. Circular shift

When parts fall off at one end and are put back at the other end, this is defined as "circular shift" with bits falling off at the left end being put back at the right end in a left rotation and bits falling off at the right end being put back at the left in a right rotation (Muchsin, Sari and Rachmawanto 2019). Fig. 8 depicts a circular shift to the left.

C. Symmetric Cipher Algorithms depending on Chronological Order

In classification of cipher algorithms as shown in section and Fig. 2, according to the chronological order, there are traditional, modern, lightweight, and ultra-lightweight algorithms ciphers. In the traditional, which is the oldest type, character oriented, can be reasonably computed by hand and based on using two only of operations types: Substitutions and permutation. In modern and lightweight ciphers, which are bit oriented, they are usually easy to decode using recent technology with product cipher (S-P) and many other components such as S-Box, P-Box, XOR, swap, circular shift, split, and combined. One of the main different between modern, lightweight, and ultra-lightweight cipher is in the number and size of components used in cipher algorithm design. This lightweight referred to as lightweight

TABLE III
THE ESSENTIAL PROPERTIES OF XOR

S. No.	Property	Its symbolization	Its description
1	Associative	$U \oplus (V \oplus Z) = (U \oplus V) \oplus Z$	The operations can be chained together without effect on the result
2	Closure	-	It is guarantees exclusive-or's result of two n-bits is completely different n-bit
3	Commutative	$U \oplus V = V \oplus U$	The sequence in which the two inputs are entered is irrelevant
4	Identity element	$U \oplus 0 = U$	Any value that XOR'd with zero is still the same unaffected
5	Self-inverse	$U \oplus U = 0$	Any value XOR'd with itself the result is zero
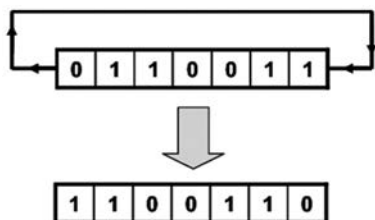
Fig. 8. Left circular shift.

cryptography (LWC), it is considering as a new cryptography type, to give adequate secure algorithms for restricted devices by reducing the length of key, the rate of each the cycle and the throughput, also the power consumption. Thus, for dealing with all parameters of computing devices to a minimal level, in lightweight and ultra-lightweight cipher system is strongly desired (Kousalya and Kumar, 2019; Sallam and Beheshti, 2018; Sliman, et al., 2021; Raza, et al., 2020). Table IV illustrates the main differences between the modern and the lightweight block ciphers:

V. Modern Symmetric Cipher Algorithms

A quick overview of common modern symmetric key (stream and block) cipher algorithms is presented. Three stream cipher algorithms and four vital block cipher are presented with concentration on basic components used in cipher algorithm design without describing the key schedule.

A. Modern Stream Cipher Algorithms

1. Rivest Cipher 4 (RC4)

Designed in 1984 and containing a basic component XOR, this modern stream cipher is one of the simplest and widely adopted ciphers, yet it is this simplicity that make this cipher covulnerable to security attacks (Brahmjit, 2015). As per functionality, for key creation, this cipher has two essential components: The Key Scheduling Algorithm (KSA) and the Pseudo Random Number Generator (PRGA), the latter of which generates a pseudorandom output sequence (bytes) from the permuted internal state, which is a random sequence. For the cryptanalyst, the statistical inadequacies of the output sequence are the focus of their investigation (Madarro-Capó, et al., 2021). Fig. 9 depicts functionality.

2. Salsa20

Being a family of 256 bits ciphers, this cipher is designed as another modern stream cipher and recommended for typical cryptographic applications due to its speed (faster than AES) while for applications where speed is more important than confidence, the faster, more reduced-round ciphers
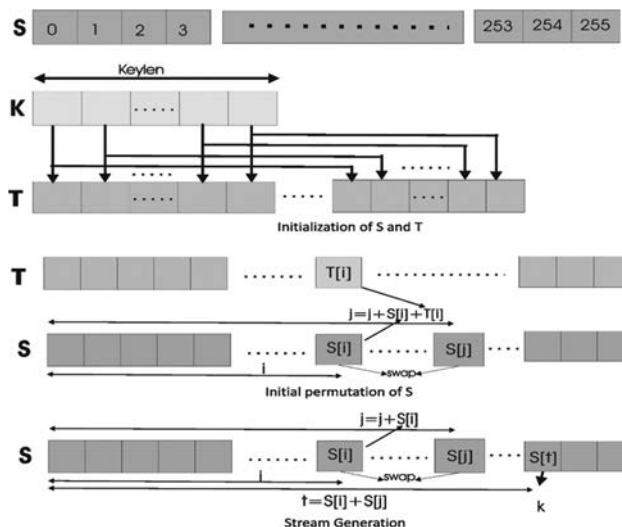
Fig. 9. Fundamental of RC4.

Salsa20/12 and Salsa20/8 are recommended. Functionality wise, this cipher encryption consists of a long chain of three simple operations on 32-bit words: Modular addition, XOR, and left shift rotation. Fig. 10 shows the quarter-round (Afdhila, Nasution and Azmi, 2016; Fukushima, et al., 2017).

3. A5/1

This cipher algorithm having two main components: The first is LFSR and the second is XOR, this algorithm is considered one of very fast stream cipher algorithms. The cipher design comprises three LFSRs: The first is R1 with length 19 bits, the second is R2 with length 22 bits, and the last is R3 with length 23 bits, this cipher designed for yield a randomly binary stream of bits with along cycle (Ekdahl and Johansson, 2003; Jiao, Hao and Feng, 2020).

The tap bits that select to achieve primitive polynomial are: 18, 17, 16, and 13 for the first register; 21 and 20 for the second one, and 22, 21, 20, and 7 for the last register (Amiri, Mahdavi and Mirzakuchaki, 2009; Sadkhan and Jawad, 2015). Fig. 11 provides illustration for the cipher design.

B. Modern Block Cipher Algorithms

1. Data encryption standard (DES)

This algorithm cipher is the first one of modern symmetric of block type. Its key and block length are (56, 64-bits), respectively, put by the NSA (Patil, et al., 2016). The design of this cipher based on 16 rounds, the main components that used in the design of each round are: S-Box, permutation, XOR, swap, split, and combined (Yihan and Yongzhen,

Criteria	Modern block cipher	Lightweight block cipher
Components number and its size	More than lightweight	Less than modern
Number of round	Less	More
The block size	64, 128, and 256 bits	32, 48, or 64 bits
Key sizes	128 bits or more	80 or 128 bits
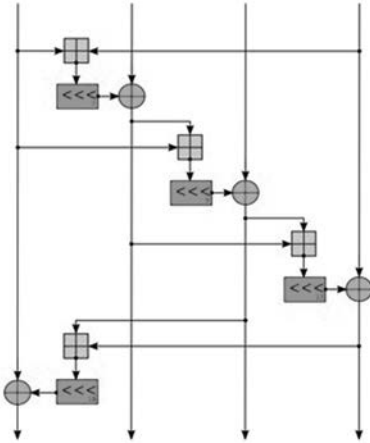
Fig. 10. Salsa20 quarter-round.
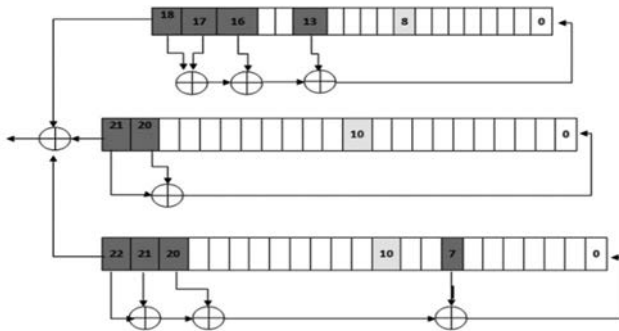
Fig. 11. The basic of A5 cipher algorithm that based on three LFSRs.

2021); the algorithm general structure with its round and its function are illustrated in the Fig. 12 below:

2. Advanced Encryption Standard (AES)

This cipher algorithm is designed as a solution for previous one. It is also modern symmetric of block type and its standard version has simple structure, based on substitution-permutation network (SPN) (Zhao, Ha and Alioto, 2015) its design comprises: Non-linear substitution step; permutation step; XOR; split; and combined components. Its key length is: (128 or 192, or 256-bits), respectively, and can be encrypted/decrypted for 128 bits of data length, as shown in Fig. 13.

The cipher design also having numbers round of plain text transformation, the number of round in this cipher depends on key length, the longer the key length, the more number of round used (10 for the 128 bits, 12 for the192 bits, and finally 14 for the 256 bits). It is flexible for hardware

implementations, also this cipher algorithm can be tailored for low-volume loads or can be improved for applications required high throughput (Kumar and Rana, 2016; Qiao, El-Assad and Taralova, 2020).

3. Blowfish

With variable key length from (32 to 448 bits) and block size of (64 bits), this modern symmetric cipher algorithm of block type is designed by Brice Schneier in 1993, with 16-round Feistel structure, intended as a replacement for cipher algorithm like DES or IDEA (faster than DES when implemented on 32 bits microprocessors). It is suitable for both domestic and commercial use due to its variable length key, most notably for applications such as communication links or file encryptors where the key does not change frequently, whereas it is unsuitable for other applications such as packet switching or smart cards that require even more compact ciphers (Schneier, 1993; Hussaini, 2020; Ghosh, 2020). The main components that used in each round are: S-Box, XOR, swap, split, and combined, as shown in Fig. 14.

4. International Data Encryption Algorithm (IDEA)

Operating on 64 bits data blocks length with 128 bits long key, the IDEA is another modern symmetric block cipher, the design principle of which depends on mixing of arithmetical operations (the three of which are: XOR, addition modulo 2^{16}, and multiplication modulo $2^{10}+1$) that are easily implemented both in hardware and software (Basu, 2011; Patil and Bhusari, 2014). The much needed non-linearity of this cipher is derived from the addition modulo 2^{16} and multiplication modulo $2^{10}+1$ arithmetical operation, whereas explicit S-Box is not used. This cipher designed base on modular additions, modular multiplications, and XOR, swap, split, and combined, as shown in Fig.15.

VI. LIGHTWEIGHT AND ULTRA-LIGHTWEIGHT SYMMETRIC CIPHER ALGORITHMS

This section presents the main of lightweight and ultra-lightweight symmetric cipher algorithms for two types (stream and block), three lightweight stream ciphers algorithms and 10 lightweight block ciphers, all these cipher algorithms are covered from many points of view (number and size of components, security, and performance).

A. Lightweight Stream Cipher Algorithms

1. Fruit-v2

It introduced informally in 2016 on the web page of IACR. This cipher was designed as ultra-lightweight symmetric stream cipher with smaller state, introduced for hardware applications in the eSTREAM project (Wang, et al., 2019), new ideas were introduced to show the possibility of shortening the size of the internal state size by exploiting a secret key not only in the initialization but also in the keystream generation (Naser and Naif, 2022), thus succeeding in reaching a formula that is able to resist attacks such as the classical time-memory data trade-off attack.
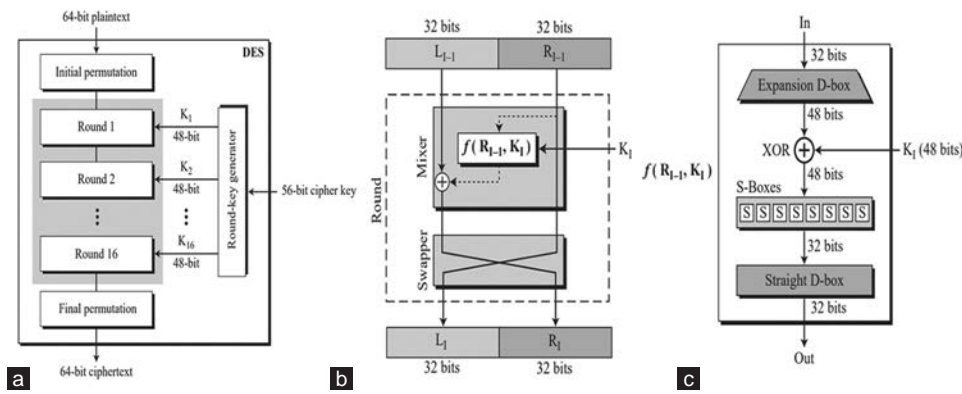
Fig. 12. (a) The general structure of DES, (b) A round in DES (encryption site), and (c) DES function.
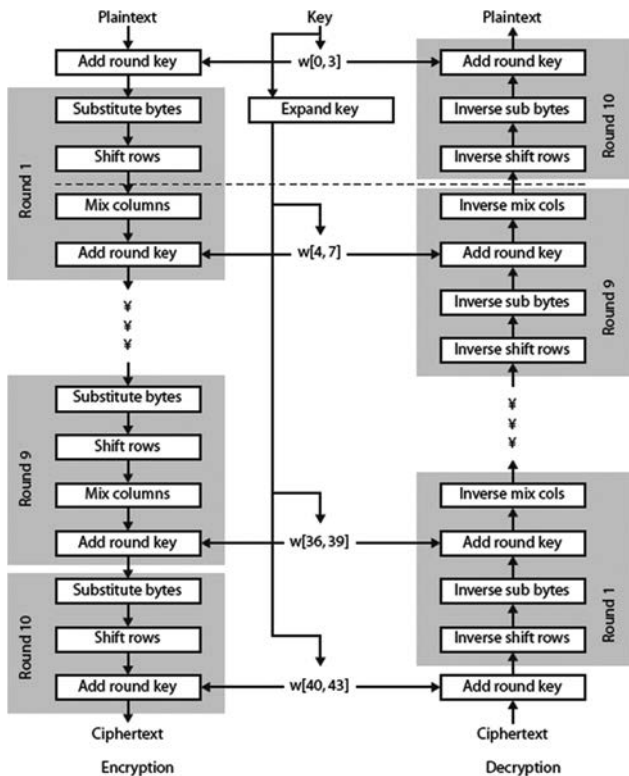
Fig. 13. AES encryption and decryption process.
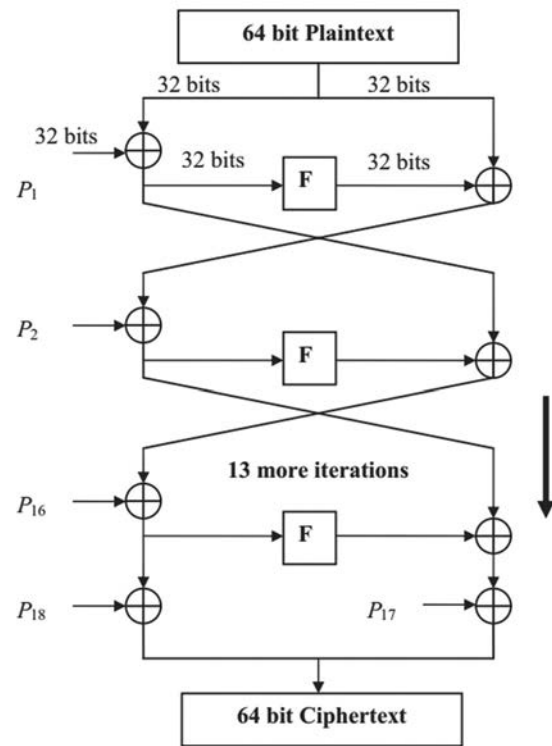
Fig. 14. A round structure.

Its structure, as shown in Fig. 16, involved of LFSR and NFSR, the size of these both FSR in Fruit-v2 is only 80 bits (Wang, et al., 2022).

2. LOGIC

Published in 2019, this cipher algorithm was introduced as a lightweight symmetric stream cipher behind a main idea of combining a chaotic system and two NFSRs. This cipher design is leading to the formation of a new cipher that is hardware-oriented and can be used efficiently in certain circumstances such as in resource constrained devices or environments (Ding, et al., 2019). As shown in Fig. 17, this algorithm has a secret key of 80 bits (40 bits for each one of the NFSR) also with three multiplexers; a filter function and logistic chaotic system. Regarding security of this cipher is considered good in resisting many essential attacks due to the high nonlinearity and the good elasticity

of the function used beside the existence of the two NFSRs (Ding, et al., 2019).

3. A4

This algorithm is lightweight symmetric stream cipher, published in 2020, its cipher design depend on two kinds of FSR: The first is LFSR and the second is FCSR (Mohandas, et al., 2020). This algorithm can be describing as two parts: The first one is encryption procedure and the second is decryption procedure, and a seed box exists between these two procedures with 128 bits key length. The characteristic of this algorithm is ensuring security to big extent and easy to implement. The LFSR acts as a clock thereby ensuring primary level of security. Robustness of this cipher against attacks like the algebraic, brute force, and differential is merited to the arrangement of LFSR and FCSR. Its time and space
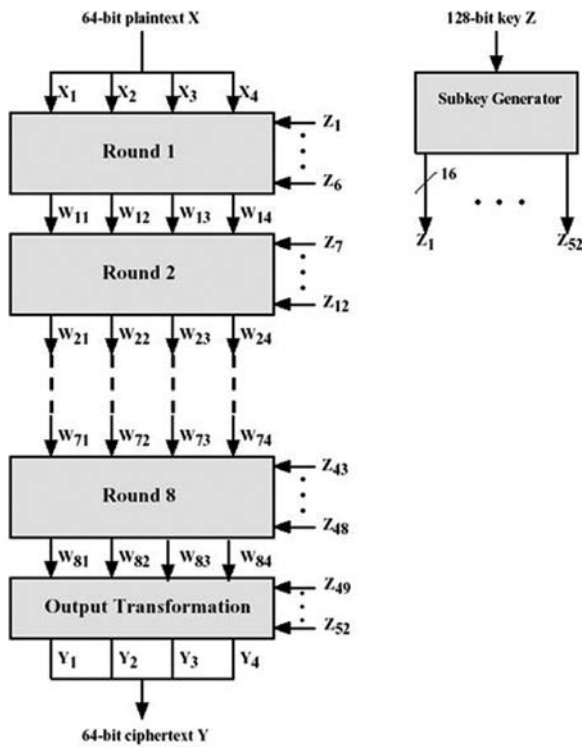
Fig. 15. IDEA encryption structure.

Fig. 16. The FRUIT-v2 cipher design.

consuming consider minimal if compared to other ciphers (Jassim and Farhan, 2021).

B. Lightweight Block Cipher Algorithms

1. Lightweight Data Encryption Standard (DESL)

Presented in 2006, this cipher, being a descendant from DES design, is the first DES variant (Poschmann, et al., 2006). By replacing the eight original S-Boxes of DES with a single new one (eliminating seven S-Boxes as well as the multiplexer implementation) achieving a reduction in the DES's gate complexity, through decreasing the chip by about 20% (1850 [GE] vs. 2310 [GE]).

Another leverage of the DESL on DES is being lighter merited to discarding the initial and final permutation of

Fig. 17. LOGIC cipher algorithm design.

DES (Kitsos, et al., 2012). The ability of this DESL to resist common attacks such as linear and differential cryptanalysis and the Davies-Murphy attack is attributed to the careful selection and prime optimization of the S-Box, reaching a level of security that is acceptable and suitable for many applications. Further robustness of the cipher can be achieved by applying key whitening, making way to the formation of the "DESXL" cipher, with a security level of approximately 118 bits. DESXL requires 2168 area (GE) compare to the DES 2309, to make it lighter.

2. Lightweight Advanced Encryption Standard (AES)

On top of being the most well-studied block cipher algorithm (Dutta, Ghosh and Bayoumi, 2019), this algorithm continues to the be subject of further studying past, present, and future optimizing its applicability for today's needs based on the intensely growing environmental demands in every single aspect of modern life with an ultimate goal of achieving sustainability. Moreover, it is through reducing the consumption of power and energy, a compact AES circuit design was proposed by Lu, et al., 2018. Only one S-Box is used in the cipher design, this one S-Box is utilized in key expansion process and data encryption process. This reduction in components achieves less latency in encryption process; and small area with high energy and throughput as a circuit design. In Mathew, et al., 2015, this cipher design based on one 8-bit S-Box with shift rows to compute all rounds, its aim to reduce the power of AES in encryption and decryption processes.

3. LBlock

Presented in 2011, and with (64, 80 bits) block size and key size, respectively, this is an ultra-lightweight symmetric block cipher algorithm (Wu and Zhang, 2011). In hardware, it occupies 1320 (GE), requiring 3955 clock cycles to encrypt a single block (Aljazeera, Nandakumar and Ershad, 2016). This cipher can be implemented effectively in hardware and software. As shown in Fig. 18, the structure of its round function based on substitution-permutation network (SPN), in which a small 4×4 S-Boxes is used with simple 4 bits word for permutation. Decent level of security in the face of attacks such as differential cryptanalysis and linear cryptanalysis can be established through the full 32 round LBlock.

Fig. 18. LBlock Block Cipher.

Fig. 19. The cipher design of Twine.

4. TWINE

Presented in 2012 by Suzaki, et al., 2011, this cipher algorithm is a lightweight symmetric block cipher with 64 bits block size along with variable key length of 80 and 128 bits. The hardware implementation is 1866 (GE). It is a Classical Feistel Network (GFN) and carries out encryption and decryption operations simultaneously. It was subjected to numerous comparable choices to LBlock previous algorithm, aiming to achieve equilibrium performance on both states (hardware and software), nevertheless, the number of S-Boxes and the permutation process remain two of the most important differences between the two ciphers, whereby, the cipher based on using one 4 bits S-Box and a 4 bits XOR, as shown in Fig. 19. Security wise, this cipher algorithm claims a high level of protection, significantly against differential and saturation attacks (Ménétrey, et al., 2021). Whereas, in Yoshikawa, Nozaki and Asahi, 2016, where an electromagnetic analysis attack method directed experimentally toward TWINE, results exposed TWINE's vulnerability when it is implemented as hardware.

5. Simon and Speck

Proposed publicly in 2013 by a group of researchers in the US National Security Agency's Research Directorate (Beaulieu, et al., 2017), Simon and Speck are two ciphers lightweight symmetric block algorithms.

With an n-bits word, the Simon cipher algorithm is used, where n need to be 16, 24, 32, 48, or 64. Simon2n/mn will be referred to as Simon2n with an m-word (mn-bits) key. Simon64/128, for example, is a variation of Simon that works with 64 bits plain text blocks and a 128 bits key. XOR, AND, mod 2n addition, and left circular shift are the Speck component units. For security, Simon and Speck employ simple round functions that are iterated as many times as necessary. In comparison, other algorithms (like as AES) use more sophisticated round functions but require fewer rounds. Because the algorithms use basic round functions, they have small realizations and are well suited for application on limited platforms. The general round of speck (Beaulieu, et al., 2013; Beaulieu, et al., 2015) is depicted in Fig. 20:
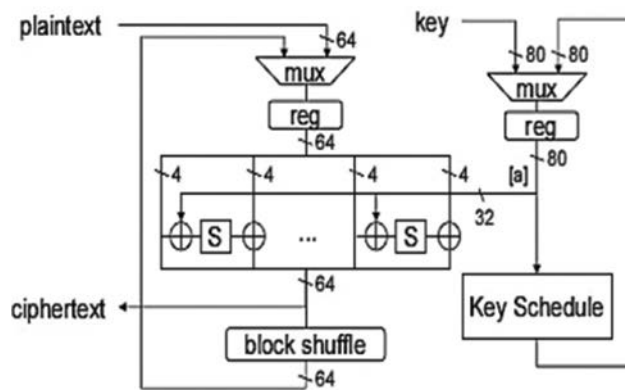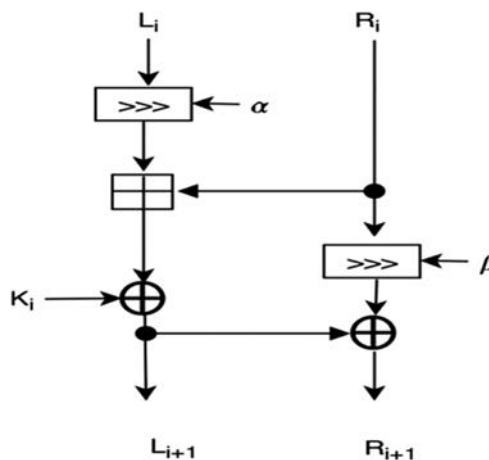
Fig. 20. General round of speck cipher.

6. RECTANGLE

This cipher algorithm, proposed by Zhang, et al., 2015, in 2014, is an ultra-lightweight block cipher with a block size of 64 bits, a variable key of 80 or 128 bits, and iterating based on SPN, as shown in Fig. 21, with a substitution mostly have of 16 (4×4) S-Boxes connected in parallel and a permutation layer consisting of three rotations. It has been demonstrated to have excellent hardware and software performance, enabling for a wide range of IOT applications (Philip, 2017).

This cipher algorithm, proposed by Zhang, et al., 2015, in 2014, is an ultra-lightweight block cipher with a block size of 64 bits, a variable key of 80 or 128 bits, and iterating based on SPN, as shown in Fig. 21, with a substitution mostly have of 16 (4×4) S-Boxes connected in parallel and a permutation layer consisting of three rotations. It has been demonstrated to have excellent hardware and software performance, enabling for a wide range of IOT applications (Philip, 2017; Philip, et al.,2018).

7. QTL

Proposed by Li, Liu and Wang 2016, in 2016, structurally as Generalized Feistel Network (GFN), this cipher algorithm is an ultra-lightweight block cipher designed for gadgets with restricted resources, with 64 bits of block size, variable key length of 64 or 128 bits iterating through 16, 20 rounds, respectively, and having a lot of S-Boxes in the encryption
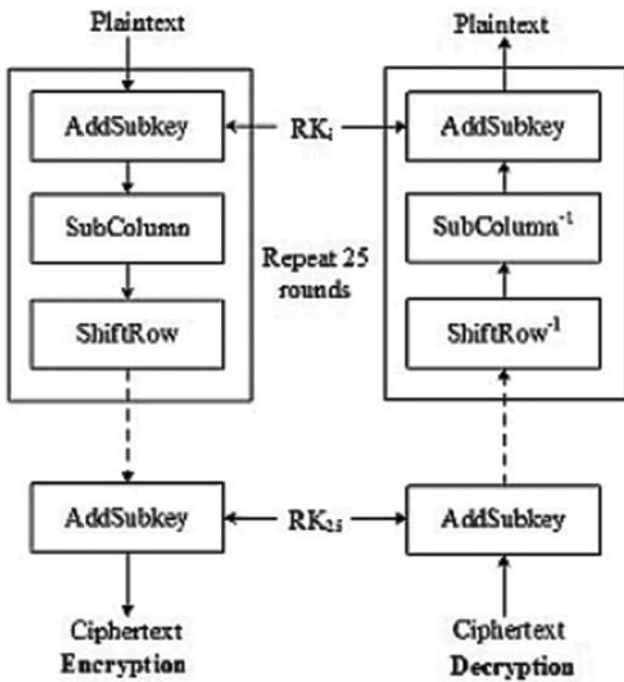
Fig. 21. The cipher design of RECTANGLE algorithm.

Fig. 22: QTL block cipher.

Fig. 23. BORON block cipher block diagram.

process. Improving the security of this cipher that has a Feistel structure was done using SPNs diffusion in the design, whereas not using a key schedule was intended to minimize hardware implementation's energy consumption (Shrivastava, Singh and Acharya, 2020). Compact HW implementation and high security as in Çoban, Karakoç and Özen, 2016, both can be achieved with QTL, a diagram of which is shown in Fig. 22.

8. BORON

Proposed by Bansod, Pisharoty and Patil, 2017, in 2017, this cipher algorithm that uses SPN, operating on a 64 bits block, supporting two key length of 80 and 128 bits, respectively, and having a total of 25 rounds, it is an ultra-lightweight block cipher with a compact structure, with higher throughput, and less power compared to other existing SPN ciphers (Sutar, 2018), and good performance concerning hardware and software platforms. It has three important components: Shift register, round permutation layers, and XOR component, as illustrated in Fig. 23 (Sutar, 2018). Opposing and deterring linear and differential attacks with this cipher are merited to the production of a high number of active S-Boxes in a small number of rounds that are possible due to its unique architecture.

9. NLBSIT

This cipher which was presented by Alahdal, et al., 2020, in 2020, designed as a new lightweight block cipher for resource restricted IOT computers due to consumption of less energy than other algorithms (encryption/decoding cycles) and less memory. Through integrating the benefits of both Feistel and SPN designs with an extra linear box idea, an increase in data security was possible to be achieved in this cipher. Structurally, it uses a key of 64 bits to encode 64 bits block size, it also uses uncomplicated mathematical operations with fewer rounds XOR, XNOR, shifting, and swapping, as illustrated in Fig. 24. High speed and high level of security resisting all known attacks achieved by the usage of both the SPN and Feistel architectures in the creation of NLBSIT are considered among the several advantages of this cipher, in addition to the less memory and less energy consumption mentioned earlier (Alahdal, et al., 2020).
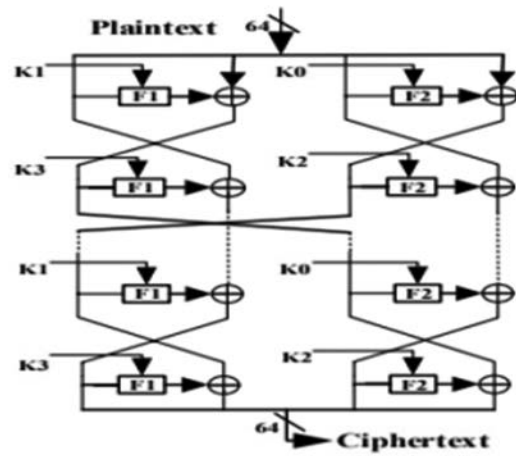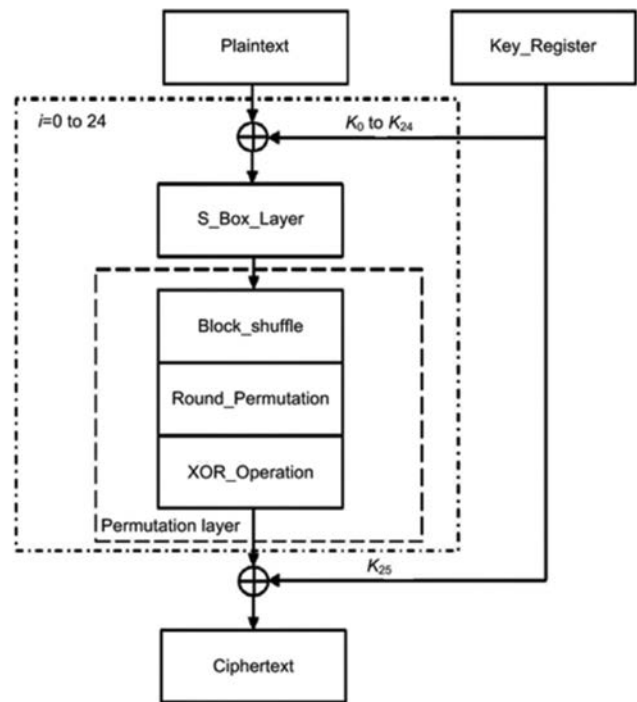
10. SAND

SAND-64 and SAND-128 are two AND-RX block ciphers having a Feistel structure. They accept 128 bits keys and have block sizes of 2n, where n refers to the branch length (n = 32 for SAND-64 and n = 64 for SAND-128). Strong security bounds and competitive performance viewed as merits of its novel design, are achieved by admitting an equivalent S-Box-based representation and thus supporting S-Box-based

TABLE V
MODERN SYMMETRIC CIPHER ALGORITHMS

Cipher name	Cipher classification	Block size	Key size	Rounds numbers	Cipher structure	S-Box
RC4	Modern stream cipher	-	-	-	-	-
Salsa20	Modern stream cipher	-	-	-	-	-
A5/1	Modern stream cipher	-	-	-	-	-
DES	Modern block cipher	64 bits	56 bits	16	Feistel based	8 S-Box
AES	Modern block cipher	128 bit	128, 192, 256 bits	10, 12, 14	SPN based	8 S-Box
Blowfish	Modern block cipher	64 bits	32–448 bits	16	Feistel based	4 S-Boxes
IDEA	Modern block cipher	64 bits	128 bit	8	Mixing of arithmetical operations	-

TABLE VI
LIGHTWEIGHT AND ULTRA-LIGHTWEIGHT SYMMETRIC CIPHER ALGORITHMS

S. No.	Cipher name	Cipher classification	Block size	Key size	Rounds numbers	Cipher structure	S-Box	P-Box
1	Fruit-v2	Ultra-lightweight stream cipher	-	-	-	-	-	-
2	LOGIC	Lightweight stream cipher	-	-	-	-	-	-
3	A4	Lightweight stream cipher	-	-	-	-	-	-
4	DESXL	Lightweight block cipher of DESX	64 bits		16	Feistel		-
5	DESL	Lightweight block cipher of DES	64 bits	56 bits	16	Feistel	Single (6 * 4 bits) S-Box 8 times	-
6	LBlock	Lightweight block cipher	64 bits	80 bits	32	Based on both Feistel and SPN	4×4 S-boxes	Permutations operate on 32bit
7	TWINE	Lightweight block cipher	64 bits	80 and 128-bit	36	Type-2 GFS	nonlinear substitution layer One 4-bit	Permutation operate on 4bits
8	Simon	Lightweight block cipher	32	64	32	Feistel Based	-	-
9	Speck	Lightweight block cipher	32	64	22	Feistel Based	-	-
10	RECTANGLE	Lightweight block cipher	64 bits	80 or 128 bits	25	SPN	16 4 bits S-Boxes	(P-layer) is composed of 3 rotations
11	QTL	Lightweight block cipher	64 bits	64 bits 128 bits	16 for 64, 20 for 128	Feistel –Based and SPNs	4×4 S-Box	Permutation operate on 16bits
12	BORON	Lightweight block cipher	64	80 bits 128 bits	25	SPN	4×4 S-Box nonlinear layer S-Box	Three sub-permutation layers Operate on 16bit
13	NLBSIT	Lightweight block cipher	64	64	5	Feistel –Based and SPN	4 S-Box	-
14	SAND-64	Lightweight block cipher	64	128	48	AND-RX block ciphers	4×8 synthetic	-

security evaluation approaches. Seeking for an easier S-Box-based cryptanalysis, the AND-RX-based design with several 4×8 S-Boxes provides a new way for the designing of AND-RX-based cipher in the future. The round function of this cipher algorithm (Chen, 2022) is explained in Fig. 25:

VII. DISCUSSION

Information security being subfield from computer science that comprises each of computer and network security, has attracted a lot of attention by several developers and scholars who's working on technologies that used every day, like smart mobile application, cloud service, and the new lifestyle using IOT.

This review presented a new classification of cryptographic, comparison between symmetric versus asymmetric cipher, comparison between stream versus block cipher, the main concepts about the components used in symmetric cipher, and finally some of selected modern and lightweight symmetric cipher algorithms are presented based on analysis its components. This review aims to illustrate how the cipher algorithms design and its properties are depending basically on the type, number, and size of components used in algorithm design.

In cryptographic fundamentals, after researched and studied many research work about cryptographic mechanism, this review present new classification that classify the cipher according to four criteria's: The number of keys that used,

TABLE VII
Chronological Order, Properties, Vulnerability, and Basic Components for All Cipher Algorithms presented in this Review

S. No.	Cipher name	Proposed in	Chronological order	Properties	Vulnerability	Basic components
1	RC4	1984	Modern	It is simple to use and fast, strong in coding and straightforward to implement, does not require extra memory, and can handle enormous streams of data when compared to other ciphers. Frequently used	A bit-flipping attack is possible	XOR
2	Salsa20	2005	Modern	Based on ARX for keystream generator, the ChaCha cipher is closely related.	There are no differential characteristics with a probability$>2^{-130}$.	Modular addition, XOR, Left shift rotation
3	A5/1	2000	Modern	Flexible, common algorithms, Widely used	Hardware-based attacks are possible	XOR and 3 LFSR
4	DES	1970	Modern	Common algorithm	Not particularly secure, but very adaptable	XOR, permutation, substitution
5	AES	1977	Modern	Flexible and common algorithm	Excellent security	XOR, addition, multiplication, mixing, substitution, shifting
6	Blowfish	1993	Modern	Flexible common algorithms	Excellent security	XOR, shifting, mixing, substitution, S-Boxes
7	IDEA	1991	Modern	Less common algorithm	Vulnerable against MITM Meet-in-the-middle attack.	XOR, swap, multiplication modulo, addition modulo 2^{16}, $2^{10}+1$, split, and combined
8	Fruit-v2	2016	Ultra-lightweight	The cipher design includes a new round key function as well as a new initialization strategy. Increase the LFSR's size.	More resistant to traditional time and memory-data trade-offs than other ciphers.	1 LSFR and 1 NSFR
9	LOGIC	2019	Lightweight	Suited for devices with a limited amount of resources.	a security margin that is too tiny	2 NFSRs
10	A4	2020	Lightweight	Low computational cost	Fundamental cryptographic attacks resistant.	1 LFSR and 1 FCSR
11	DESL	2006	Lightweight	DESL encryption is more power efficient than DES, requiring 20% fewer gate equivalences and 25% less DES implementation.	More secure against specific types of cryptanalyses, including linear and differential because of the S-Box's non-linearity, it is more resistant to the Davies-Murphy attack, more size optimized, and more power efficient than DES.	XOR, S-Box, permutation
12	LBlock	2011	Lightweight	Improve hardware performance while also increasing software efficiency.	More secure against differential and linear, related key, integral attack.	XOR, S-Box, permutations, left cyclic shift, right cyclic shift, concatenation
13	TWINE	2012	Lightweight	Its primarily intended to fit into very small hardware.	Its security naturally needs to be studied further.	XOR, S-Box, permutation, split, and combined
14	Simon	2013	Lightweight	Simplicity, ease of implementation, and efficient implementations, attractive for homomorphic encryption, as does support for a broad 12 range of block and key sizes.	a security margin that is too tiny	Exclusive-OR, AND, left circular shift
15	Speck	2013	Lightweight	Simplicity, easy to implement, efficient implementations, attractive for homomorphic encryption, support for a broad 12 range of block and key sizes	A security margin that is too tiny	Exclusive-OR, addition modulo 2^n, left and right circular shifts
16	RECTANGLE	2014	Lightweight	By leveraging bit-slice techniques, the primary purpose of this cipher design is to allow for lightweight and rapid implementations.	Differential and linear distinguisher; Statistical Saturation Attack are all impossible to build. Also offers adequate protection to withstand integral cryptanalysis and key schedule assaults	XOR, S-Box, row rotated, permutation

S. No.	Cipher name	Proposed in	Chronological order	Properties	Vulnerability	Basic components
17	QTL	2016	Lightweight	suited for devices with a limited amount of resources.	achieves a high level of security. Differential, linear, algebraic, and related-key attacks are all safe.	Permutation operate on 16 bits, S-Box layer, exclusive-OR
18	BORON	2017	Lightweight	Low-power encryption with excellent performance on both the software and hardware platforms.	Linear, differential, key scheduling, and key collision attacks are all resistant.	Shift, permutation layers, exclusive-OR
19	NLBSIT	2020	Lightweight	Using simple mathematics, requires less energy and memory.	Linear, differential, Weak Key, Related keys, and square attack resistance.	XOR, XNOR, shifting and swapping
20	SAND-64	2020	Lightweight	One of the most lightweight primitives available in software	All of its variants are resistant to MITM (Meet-in-the-Middle) attack	AND, rotation, and XOR
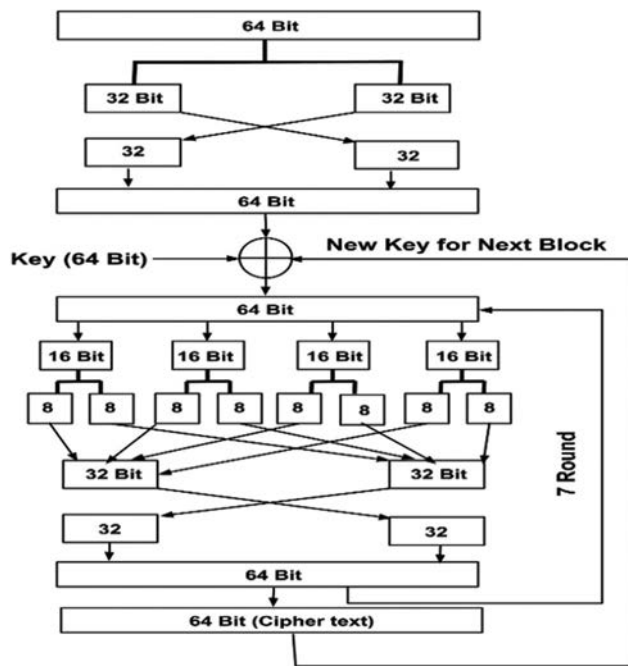
Fig. 24. The encipherment of NLBSIT algorithm.
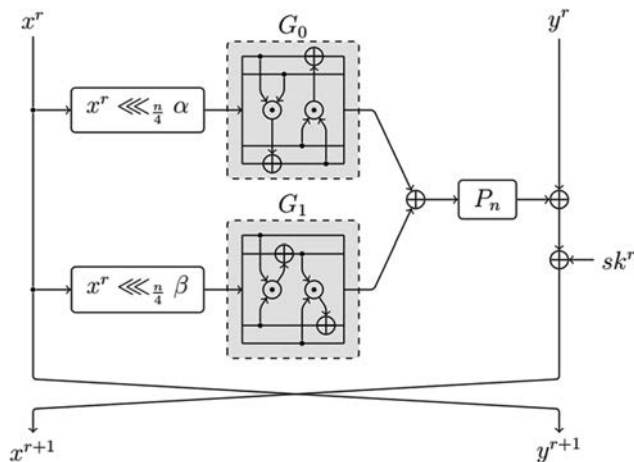
Fig. 25. Round function of SAND.

the type of encryption operation, the way of processing the plain text, and finally classifies according to chronological order. Farther more the cipher algorithm can classify to more than 1 type as illustrate in Section 2, these new classifications help to give fast and comprehensive understanding for cryptographic fundamentals.

In first comparison presented in this review between symmetric and asymmetric cipher, the conclusion was: That the two most important properties, the fast and less complex for symmetric cipher are due to the component units used in cipher algorithm design like XOR, S-Box, P-Box, circular shift, swap, split, and combined that used in cipher algorithm design, which different from asymmetric cipher that depending on numbering manipulation in cipher algorithm design. Furthermore, these properties make the symmetric cipher type the best solution for lightweight cipher applications. The second comparison presented is between the stream cipher and block cipher, the main types of symmetric cipher, the different of stream from block is not only in the way of processing the plain text but also the components that used in algorithm design are different, it is depends basically on FSR and XOR components, whereas in block cipher, the algorithm design depends on using other components such as S-Box, P-Box, split, and combine as mention previously. It is important to mention that these collection of components made the used of block cipher more versatile than stream ciphers, where used in many of lightweight cipher applications.

From the review of modern, lightweight and ultra-lightweight symmetric cipher algorithms in this research; 3 modern stream ciphers, 4 modern block ciphers, 3 lightweight stream ciphers, and 10 lightweight/ultra-lightweight block ciphers are presented, studied, and listed in Tables V and VI, respectively, based on chronological order classification. The conclusion from studying these cipher algorithms is: The type of cipher algorithm whether if it is modern or lightweight is basically depending on the type, the number, and size of components that used in cipher algorithm design. The number and size of components in lightweight type are less than those of conventional ciphers. Hence, the conclusion,

the type, number, and the size of component will effect in obtain new cipher algorithm that is more suitable solution for today information security applications. In addition, the final conclude is that the important and basic component that used in all types of symmetric cipher is XOR, due to the important properties of this component as illustrate in Section 3.2.2. Another important and basic component used in symmetric block cipher is S-Box that achieves confusion between the cipher text and the secret key.

In the end of this section, Table VII summarizes the major characteristics, chronological order, vulnerability, and basic components of each cipher algorithm presented in this review research.

VIII. Conclusion

This review is providing an overview of cryptography fundamental, theoretical background of symmetric ciphers with its two main types (stream and block), main concepts of the components used in symmetric cipher and finally some of selected modern and lightweight symmetric algorithms are presented. The present review was attempt to present the cipher algorithm depending on the basic components that used in cipher design, and explain how these components kind's, number, and size effect on the classification of cipher algorithm and on performance, area, and its weight. Better understating of various algorithms following their establishment, design, function, and field of application can be found in this study, with futuristic aspirations for more analysis and development that can lead to improvement of pre-existing new encryption algorithms or the evolution of new ones.

References

Afdhila, D., Nasution, S.M. and Azmi, F., 2016. Implementation of stream cipher Salsa20 algorithm to secure voice on push to talk application. In: *2016 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*. IEEE, United States. pp.137-141.

Alahdal, A., AL-Rummana, G.A., Shinde, G.N. and Deshmukh, N.K., 2020. NLBSIT: A new lightweight block cipher design for securing data in IOT devices. *International Journal of Computer Sciences and Engineering*, 8(10), p.13.

Aljazeera, K.R., Nandakumar, R. and Ershad, S.B., 2016. Design and characterization of L Block cryptocore. In: *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*. IEEE, United States. pp.166-172.

Amiri, M.A., Mahdavi, M. and Mirzakuchaki, S., 2009. QCA implementation of A5/1 stream cipher. In: *2009 Second International Conference on Advances in Circuits, Electronics and Micro-Electronics*. IEEE, United States. pp.48-51.

Anand, A., Raj, A., Kohli, R. and Bibhu, V., 2016, Proposed symmetric key cryptography algorithm for data security. In: *2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH)*. IEEE, United States. pp.159-162.

Bagane, P. and Sirbi, D.K., 2021. Comparison between traditional cryptographic methods and genetic algorithm based method towards Cyber Security. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 12(2), pp.676-682.

Bansod, G., Pisharoty, N. and Patil, A., 2017. BORON: An ultra-lightweight and low power encryption design for pervasive computing. *Frontiers of Information Technology and Electronic Engineering*, 18(3), pp.317-331.

Bardis, N.G., Markovskyy, A.P. and Andrikou, D.V., 2004. Method for designing pseudorandom binary sequences generators on Nonlinear Feedback Shift Register(NFSR). *WSEAS Transactions on Communications*, 3(2), pp.758-763.

Basu, S., 2011. International data encryption algorithm (Idea)-a typical illustration. *Journal of Global Research in Computer Science*, 2(7), pp.116-118.

Beaulieu, R., Shors, D., Smith, J., Treatman-Clark, S., Weeks, B. and Wingers, L., 2017. Notes on the design and analysis of SIMON and SPECK.

Beaulieu, R., Shors, D., Smith, J., Treatman-Clark, S., Weeks, B. and Wingers, L., 2013. *Implementation and Performance of the Simon and Speck Lightweight Block Ciphers on ASICs*. Unpublished Work.

Beaulieu, R., Shors, D., Smith, J., Treatman-Clark, S., Weeks, B. and Wingers, L., 2015. The SIMON and SPECK lightweight block ciphers. In: *Proceedings of the 52nd Annual Design Automation Conference.* pp.1-6.

Chen, S., Fan, Y., Sun, L., Fu, Y., Zhou, H., Li, Y., Wang, M., Wang, W. and Guo, C., 2022. SAND: An AND-RX Feistel lightweight block cipher supporting S-box-based security evaluations. *Designs, Codes and Cryptography*, 90(1), pp.155-198.

Chiadighikaobi, I.R. and Katuk, N., 2021. A scoping study on lightweight cryptography reviews in IOT. *Baghdad Science Journal*, 18(2), pp.989-1000.

Chugunkov, I.V., Kliuchnikova, B.V., Riakhovskaia, I.S., Chernikova, E.A. and Chugunkov, V.I., 2020. Improvement of P-box efficiency. In: *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. IEEE, United States. pp.274-276.

Çoban, M., Karakoç, F. and Özen, M., 2016. Cryptanalysis of QTL Block cipher. In: *International Workshop on Lightweight Cryptography for Security and Privacy*. Springer, Cham. pp.60-68.

De Canniere, C., Biryukov, A. and Preneel, B., 2006. An introduction to block cipher cryptanalysis. *Proceedings of the IEEE*, 94(2), pp.346-356.

Ding, L., Liu, C., Zhang, Y. and Ding, Q., 2019. A new lightweight stream cipher based on chaos. *Symmetry*, 11(7), p.853.

Dreier, J., Hirschi, L., Radomirovic, S. and Sasse, R., 2018. Automated unbounded verification of stateful cryptographic protocols with exclusive OR. In: *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, United States. pp.359-373.

Dutta, I.K., Ghosh, B. and Bayoumi, M., 2019. Lightweight cryptography for internet of insecure things: A survey. In: *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, United States. pp.0475-0481.

Easttom, W., 2021. S-box design. In: *Modern Cryptography*. Springer, Cham. pp. 187-204.

Ekdahl, P. and Johansson, T., 2003. Another attack on A5/1. *IEEE Transactions on Information Theory*, 49(1), pp.284-289.

Forouzan, B.A. and Mukhopadhyay, D., 2015. *Cryptography and Network Security*. Vol. 12. McGraw Hill Education (India) Pvt Ltd., New York, NY, USA.

Fukushima, K., Xu, R., Kiyomoto, S. and Homma, N., 2017. Fault injection attack on Salsa20 and ChaCha and a lightweight countermeasure. In: *2017 IEEE Trustcom/BigDataSE/ICESS*. IEEE, United States. pp.1032-1037.

Ghosh, A., 2020. Comparison of encryption algorithms: AES, Blowfish and Twofish for security of wireless networks. *International Research Journal of Engineering Technology*, 7, pp.4656-4658.

Hamza, A. and Kumar, B., 2020, A review paper on DES, AES, RSA encryption standards. In: *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*. IEEE, United States. pp.333-338.

Hasan, M.K., Shafiq, M., Islam, S., Pandey, B., Baker El-Ebiary, Y.A., Nafi, N.S., Rodriguez, R.C. and Vargas, D.E., 2021. Lightweight cryptographic algorithms

for guessing attack protection in complex internet of things applications. *Complexity*, 2021, 5540296.

Hatzivasilis, G., Fysarakis, K., Papaefstathiou, I. and Manifavas, C., 2018. A review of lightweight block ciphers. *Journal of Cryptographic Engineering*, 8(2), pp.141-184.

Hussain, I. and Shah, T., 2013. Literature survey on nonlinear components and chaotic nonlinear components of block ciphers. *Nonlinear Dynamics*, 74(4), pp.869-904.

Hussaini, S., 2020. Cyber security in cloud using blowfish encryption. *International Journal of Information Technology*, 6(5).

Jassim, S.A. and Farhan, A.K., 2021. A survey on stream ciphers for constrained environments. In: *2021 1ˢᵗ Babylon International Conference on Information Technology and Science (BICITS)*. IEEE, United States. pp.228-233.

Jiao, L., Hao, Y. and Feng, D., 2020. Stream cipher designs: A review. *Science China Information Sciences*, 63(3), pp.1-25.

Jindal, P. and Singh, B., 2015. RC4 encryption-a literature survey. *Procedia Computer Science*, 46, pp.697-705.

Kitsos, P., Sklavos, N., Parousi, M. and Skodras, A.N., 2012. A comparative study of hardware architectures for lightweight block ciphers. *Computers and Electrical Engineering*, 38(1), pp.148-160.

Kousalya, R. and Kumar, G.S., 2019. A survey of light-weight cryptographic algorithm for information security and hardware efficiency in resource constrained devices. In: *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*. IEEE, United States. pp.1-5.

Kumar, D.S., Suneetha, C.H. and Chandrasekhar, A., 2012. A block cipher using rotation and logical XOR operations. *arXiv preprint arXiv:1202.1898*.

Kumar, P. and Rana, S.B., 2016. Development of modified AES algorithm for data security. Optik, 127(4), pp.2341-2345.

Lakhtaria, K.I., 2011. Protecting computer network with encryption technique: A study. In: *International Conference on Ubiquitous Computing and Multimedia Applications*. Springer, Berlin, Heidelberg. pp.381-390.

Li, L., Liu, B. and Wang, H., 2016. QTL: A new ultra-lightweight block cipher. *Microprocessors and Microsystems*, 45, pp.45-55.

Lu, M., Fan, A., Xu, J. and Shan, W., 2018. A compact, lightweight and low-cost 8-bit datapath AES circuit for IOT applications in 28nm CMOS. In: *2018 17ᵗʰ IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12ᵗʰ IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE)*. IEEE, United States. pp.1464-1469.

Madarro-Capó, E.J., Legón-Pérez, C.M., Rojas, O. and Sosa-Gómez, G., 2021. Information theory based evaluation of the RC4 stream cipher outputs. *Entropy*, 23(7), p.896.

Mathew, S., Satpathy, S., Suresh, V., Anders, M., Kaul, H., Agarwal, A., Hsu, S., Chen, G. and Krishnamurthy, R., 2015. 340 mv-1.1 v, 289 gbps/w, 2090-gate nanoaes hardware accelerator with area-optimized encrypt/decrypt GF (2 4) 2 polynomials in 22 nm tri-gate CMOS. *IEEE Journal of Solid-State Circuits*, 50(4), pp.1048-1058.

Ménétrey, J., Pasin, M., Felber, P. and Schiavoni, V., 2021. Twine: An embedded trusted runtime for webassembly. In: *2021 IEEE 37ᵗʰ International Conference on Data Engineering (ICDE)*. IEEE, United States. pp.205-216.

Mewada, S., Sharma, P. and Gautam, S.S., 2016. Classification of efficient symmetric key cryptography algorithms. *International Journal of Computer Science and Information Security*, 14(2), p.105.

Mohandas, N.A., Swathi, A., Abhijith, R., Nazar, A. and Sharath, G., 2020. A4: A lightweight stream cipher. In: *2020 5ᵗʰ International Conference on Communication and Electronics Systems (ICCES)*. IEEE, United States. pp.573-577.

Muchsin, H.N., Sari, D.E. and Rachmawanto, E.H., 2019. Text encryption using extended bit circular shift cipher. In: *2019 Fourth International Conference on Informatics and Computing (ICIC)*. IEEE, United States. pp.8138-8143.

Naser, N.M. and Naif, J.R., 2022. A systematic review of ultra-lightweight encryption algorithms. *International Journal of Nonlinear Analysis and Applications*, 13(1), pp.3825-3851.

Pachghare, V.K., 2019. *Cryptography and Information Security*. PHI Learning Pvt. Ltd., New Delhi.

Patil, P., Narayankar, P., Narayan, D.G. and Meena, S.M., 2016. A comprehensive evaluation of cryptographic algorithms: DES, 3DES, AES, RSA and blowfish. *Procedia Computer Science*, 78, pp.617-624.

Patil, S. and Bhusari, V., 2014. An enhancement in international data encryption algorithm for increasing security. *International Journal of Application or Innovation in Engineering and Management*, 3(8), pp.64-70.

Philip, M.A., 2017. A survey on lightweight ciphers for IOT devices. In: *2017 International Conference on Technological Advancements in Power and Energy (TAP Energy)*. IEEE, United States. pp.1-4.

Philip, M.A., Vaithiyanathan, V. and Jain, K., 2018. Implementation analysis of rectangle cipher and its variant. In: *2018 3ʳᵈ IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology (RTEICT)*. IEEE, United States. pp.474-479.

Poschmann, A., Leander, G., Schramm, K. and Paar, C., 2006. A Family of Light-Weight Block Ciphers Based on DES Suited for RFID Applications. Vol. 6. *Workshop on RFID Security-RFIDSec*.

Qadir, A.M. and Nurhayat, V., 2019. A review paper on cryptography. *International Symposium on Digital Forensics and Security (ISDFS)*. IEEE, United States. pp.1-6.

Qiao, Z., El Assad, S. and Taralova, I., 2020. Design of secure cryptosystem based on chaotic components and AES S-Box. *AEU-International Journal of Electronics and Communications*, 121, p.153205.

Raza, A.R., Mahmood, K., Amjad, M.F., Abbas, H. and Afzal, M., 2020. On the efficiency of software implementations of lightweight block ciphers from the perspective of programming languages. *Future Generation Computer Systems*, 104, pp.43-59.

Sadkhan, S.B. and Jawad, N.H., 2015. Simulink based implementation of developed A5/1 stream cipher cryptosystems. *Procedia Computer Science*, 65, pp.350-357.

Sallam, S. and Beheshti, B.D., 2018. A survey on lightweight cryptographic algorithms. In: *TENCON 2018-2018 IEEE Region 10 Conference*. IEEE, United States. pp.1784-1789.

Schneier, B., 1993. Description of a new variable-length key, 64-bit block cipher (Blowfish). In: *International Workshop on Fast Software Encryption*. Springer, Berlin, Heidelberg. pp.191-204.

Sharma, D.K., Singh, N.C., Noola, D.A., Doss, A.N. and Sivakumar, J., 2022. A review on various cryptographic techniques and algorithms. *Materials Today: Proceedings*, 51, pp.104-109.

Shrivastava, N., Singh, P. and Acharya, B., 2020. Efficient hardware implementations of QTL cipher for RFID applications. *International Journal of High Performance Systems Architecture*, 9(1), pp.1-10.

Sliman, L., Omrani, T., Tari, Z., Samhat, A.E. and Rhouma, R., 2021. Towards an ultra-lightweight block ciphers for Internet of Things. *Journal of Information Security and Applications*, 61, p.102897.

Soomro, Z.A., Shah, M.H. and Ahmed, J., 2016. Information security management needs more holistic approach: A literature review. *International Journal of Information Management*, 36(2), pp.215-225.

Stallings, W., Brown, L., Bauer, M.D. and Howard, M., 2012. *Computer Security: Principles and Practice*. Vol. 2. Pearson, Upper Saddle River.

Sutar, S.A., 2018. Differential power attack analysis of ultra-lightweight block cipher

BORON. In: *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, United States. pp. 365-370.

Suzaki, T., Minematsu, K., Morioka, S. and Kobayashi, E., 2011. Twine: A Lightweight, Versatile Block Cipher. Vol. 2011. In: *ECRYPT Workshop on Lightweight Cryptography*.

Szaban, M. and Seredynski, F., 2011. Designing cryptographically strong S-boxes with use of ID cellular automata. *Journal of Cellular Automata*, 6(1).

Wahid, M.N., Ali, A., Esparham, B. and Marwan, M., 2018. A comparison of cryptographic algorithms: DES, 3DES, AES, RSA and blowfish for guessing attacks prevention. *Journal Computer Science Applications and Information Technology*, 3(2), pp.1-7.

Wang, S., Liu, M., Lin, D. and Ma, L., 2019. Fast correlation attacks on grain-like small state stream ciphers and cryptanalysis of plantlet, fruit-v2 and fruit-80.

Wang, S., Liu, M., Lin, D. and Ma, L., 2022. On grain-like small state stream ciphers against fast correlation attacks: Cryptanalysis of plantlet, fruit-v2 and fruit-80. *The Computer Journal,* bxac016.

Wu, W. and Zhang, L., 2011. L Block: A lightweight block cipher. In: *International Conference on Applied Cryptography and Network Security*.

Springer, Berlin, Heidelberg. pp.327-344.

Yihan, W. and Yongzhen, L., 2021. Improved design of DES algorithm based on symmetric encryption algorithm. In: *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*. IEEE, United States. pp. 220-223.

Yoshikawa, M., Nozaki, Y. and Asahi, K., 2016. Electromagnetic analysis attack for a lightweight block cipher TWINE. In: *2016 IEEE/ACES International Conference on Wireless Information Technology and Systems (ICWITS) and Applied Computational Electromagnetics (ACES)*. IEEE, United States. pp.1-2.

Zhang, W., Bao, Z., Lin, D., Rijmen, V., Yang, B. and Verbauwhede, I., 2015. RECTANGLE: A bit-slice lightweight block cipher suitable for multiple platforms. *Science China Information Sciences*, 58(12), pp.1-15.

Zhao, W., Ha, Y. and Alioto, M., 2015. AES architectures for minimum-energy operation and silicon demonstration in 65nm with lowest energy per encryption. In: *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, United States. pp.2349-2352.

Burning Skin Detection System in Human Body

Noor M. Abdulhadi, Noor A. Ibraheem and Mokhtar M. Hasan

Department of Computer Science, College of Science for Women, University of Baghdad,
Baghdad, Iraq

Abstract—**Early accurate burn depth diagnosis is crucial for selecting the appropriate clinical intervention strategies and assessing burn patient prognosis quality. However, with limited diagnostic accuracy, the current burn depth diagnosis approach still primarily relies on the empirical subjective assessment of clinicians. With the quick development of artificial intelligence technology, integration of deep learning algorithms with image analysis technology can more accurately identify and evaluate the information in medical images. The objective of the work is to detect and classify burn area in medical images using an unsupervised deep learning algorithm. The main contribution is to developing computations using one of the deep learning algorithm. To demonstrate the effectiveness of the proposed framework, experiments are performed on the benchmark to evaluate system stability. The results indicate that, the proposed system is simple and suits real life applications. The system accuracy was 75%, when compared with some of the state-of-the-art techniques.**

Index Terms—**Skin burn, Clustering, Deep learning, Fuzzy c-means clustering, Image segmentation, Medical images.**

I. Introduction

It is possible to make an accurate diagnosis of burns in the early stages through computerized image processing. The differences in the degree of color between the affected and uninjured skin are the real criterion for distinguishing between them and for knowing the severity of the injury. Distinguishing the type of injury is based on the color analysis method according to criteria and statistical indications to find out the type of injuries if they are mild, moderate, or serious. This technique can be adapted to assess and track wound severity in humans in a clinical setting (Hao, et al., 2021). Regardless of this aspect, this review is for the burned skin area. Whereas different burn areas have different types of burns, the depth of the burn is usually maximum in a burn area. However, a series of tests can be carried out to determine the extent of humidity. An evaluation by a dermatologist after the examination cannot

be ruled out. Expert opinions remain to date the most reliable assessment of injury (Malini, Siva, and Niranjan, 2013). An overview of recent techniques for burn images is presented in (Domagoj and Damir, 2020). This research proposed an artificial intelligence system that uses unsupervised deep learning, to examine the affected area and recognize it, as well as to give an accurate diagnosis.

The skin is the largest organ in the body and is mostly damaged during burn accidents. The skin accounts for 15% of the total weight of an adult human being. The primary functions of the skin are protection, sensation, temperature regulation, and Vitamin D synthesis. The main components of human skin are the epidermis and dermis. The epidermis is the thin outer part of the skin, whereas the dermis is the inner thick layer of connective tissue made of elastic fibers, Burns are characterized by their size, area, and depth. Burns are classified as first-, second-, third-degree, or fourth-degree depending on how deeply and severely they penetrate the skin's surface (Tina, et al., 2015; Kuan, et al., 2018).

A. First-degree (Superficial) Burns

First-degree burns influence just the epidermis, or external layer of skin. The consume site is red, difficult, dry, and without any rankles. Gentle burn from the sun is a model. Long-term tissue harm is uncommon and as a rule comprises an increment or decline in the skin tone (Papini, 2004; Miller, et al., 1998), as shown in Fig. 1.

B. Second-degree (Partial Thickness) Burns

Second-degree burns include the epidermis and a piece of the dermis layer of skin. The consumption site seems red, rankled, and might be enlarged and painful (Hao, et al., 2021; Papini, 2004). Fig. 2 is an example of partial thickness burns.

C. Third-degree (Full Thickness) Burns

Third-degree burns devastate the epidermis and dermis. Third-degree burns may likewise harm the hidden bones, muscles, and ligaments. The affected site seems white or burned. There is no sensation nearby since the sensitive spots are obliterated. Fig. 3 is an example of full-thickness burns.

The rest of the paper is organized as follows. Related work is presented in Section II. Section III introduced proposed system methodology. The experimental results and comparisons with other systems were demonstrated in Section IV. Section V concluded the paper by supplying a summary of the steps taken in this study.

Fig. 1. Superficial dermal burn.

Fig. 2. Partial thickness burn.

Fig. 3. Full thickness burn.

II. Related Works

The current standard for evaluating burn wounds is based on digital photography of wounds examined by a burn specialist. Researchers are developing automated burn wound analysis systems due to the subjectivity of this approach. These systems should include three major components: burn image segmentation, feature extraction (Hasan, Ibraheem and Abdulhadi, 2022), and classification of segmented regions into healthy skin, and burned skin. (Liu, et al., 2021) introduced a review of some common ideas and characteristic features of medical image segmentation based on deep learning algorithms by applying the segmentation method for different human organs areas such as brain tumors, lung nodules, eyes, chest, heart, abdomen, and some other parts besides, whereas (Liangrui, Zhichao and Shaoliang, 2022) reviewed an investigation on various methods in pathology in a pathological and technical perspective.

According to the study results, image acquisition type has a significant impact on segmentation and classification performance. The fuzzy c-means algorithm for segmentation and a multilayer feed-forward ANN trained by the back-propagation algorithm for classification were found to be the best combination for successfully categorizing the images into the skin, burn, and background regions (Ugur, et al., 2019).

The deep learning-based image segmentation method achieves excellent results, for example, in (Badrinarayanan, Kendall, and Cipolla, 2017), researchers compared the performance of the segmentation approach versus deep learning, SegNet-based segmentation process was implemented as a deep learning approach (Badrinarayanan, Kendall, and Cipolla, 2017). Mask R-CNN is a method for utilizing the cutting-edge deep learning framework, researchers have modified the Mask R-CNN to adapt the dataset for more refined result and faster training speed (Chong, et al., 2019) (Prabhakar, Gauraveand and Kailesh, 2020). Over-redundant vocabulary is used to detect and classify burnt areas using sparse representations of feature vectors. Feature vectors are extracted from image patches and assigned to a class representing a burning degree for each patch. Color and texture information are used as features (Brenda and Roberto, 2021). Various segmentation algorithms have been proposed to improve classification performance, to extract the wound from an image, the DNN is trained to segment the testing data (Fangzhao, et al., 2018). Authors (Hao, et al., 2021), built a framework that segmented the burn area in burn images and calculated the percentage of burn area in total body surface area (TBSA) to extend the network output structure and label the burn dataset. The framework is then applied to segment multiple burn depth areas. The most important step in image classification is feature extraction. As a result, many works proposed various types of color and/or texture-based features. Alcala-Fdez, and Alonso (2016) gave a survey of fuzzy systems software in means of taxonomy, recent researchers issues. The features generated by (Deepak, Antony, and Niranjan 2012) are the mean and (2, 1) the coefficient of the discrete cosine transform (DCT) function of the V1 chrominance plane of the L*a*b* color space were extracted as features. Fully convolutional networks (FCN) have been applied by (Despo, et al., 2017), whereas researchers in (Sabeena and Raj Kumar, 2019) have improved the feature extraction of GLCM and further to the SVM method. An automated system for categorizing healthy and burned skin in images of trauma patients was developed. A database was created using 105 images gathered from Turkish burn centers. The effects of color spaces and feature matrices on the results of segmentation-based classification methods were investigated. The FCM algorithm for segmentation and a multilayer feed forward ANN trained by the back-propagation algorithm for classification were found to be the best combination for successfully classifying

burn images. The a* and b* channels of the L*a*b* color space were used as input to the FCM algorithm, whereas the artificial neural network was trained using selected Haralick features. The average F-score obtained was 0.7428. Deep learning semantic segmentation was also used to classify burn images. The architecture trained on 64 pixel sized blocks analyzed from training datasets achieved an F-score of 0.805 (Ugur, et al., 2019). Fuzzy c means algorithm based on morphological reconstruction was applied by (Rahman and Islam, 2021), whereas researchers in (Wang, et al., 2021) incorporated a residual-related regularization term derived from the distribution properties of various types of noise into FCM. An image mining approach was used to conduct a comparative study of various classification algorithms on the classification of different burn depths as explained in (Kuan, et al., 2018), authors (Hansen, et al., 1997) developed several approaches for noninvasive burn evaluation in humans in a clinical environment employing a color imaging system for monitoring and tracking wound severity, In (Nameirakpam, Khumanthem, and Yambem, 2015), authors applied K-means clustering algorithm with a subtractive clustering algorithm to segment 5 images. A comprehensive overview of current deep learning-based medical picture segmentation approaches is presented by (Liu, et al., 2021). Wagh, et al., (2020) used smartphone wound picture analysis as a practical tool to monitor the healing process and give patients and caregivers useful feedback in between visits to the hospital.

III. System Methodology

The main objective of this study is to segment different kinds of medical burn images using deep learning. The training and testing set of images were used in their entirety without cropping or any preprocessing detection of the main injured area. Manual input images, and (*Stephen*, 2020) database has also been used in our work (Health Encyclopedia Site).

The following subsections explain briefly Fuzzy C-means, deep learning algorithm, and proposed work.

A. Fuzzy C-mean Clustering (FCM)

Fuzzy C-means (FCM) is the most widely used approach for pixel categorization. The layout of partition space for clustering algorithms is as follows:

Let c be an integer, and x_1, x_2, x_3,x_n signify a set of n unlabeled vectors in the space Rp, where p is the number of features in each vector. The standard FCM objective function for partition $(x_k)_{k=1}^{n}$ into c clusters is given by equations (1), (2), and (3):

$$\text{Minimize } j_m(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m(d_{ik})^2 \tag{1}$$

$$\text{Subject to } \sum_{j=1}^{c} u_{ik} = 1, 1 \le i \le n \tag{2}$$

$$u_{ik} \ge 0, 1 \le i \le n, 1 \le j \le c \tag{3}$$

The exponent m > 1 is used to adjust the membership value's weighting impact. Where $d_{ik} = \| x_k - v_i \|$. U is the fuzzy partition matrix, which provides the membership of each cluster's feature vectors (Bezdek, 1981) (Bekir. 2016) (Yeganejou and Dick, 2018).

The cluster center matrix is denoted by the letter v., $\{v_1, v_2., v_c\}$ where vi is a cluster center of features. The updated cluster center v_i can be obtained from the following equations (4), (5) and (6):

$$v_i \frac{\sum_{k=1}^{n} u_{ik}{}^m x_k}{\sum_{k=1}^{n} u_{ik}{}^m} \text{ for each i} \tag{4}$$

The objective function U can be derived from:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}(\frac{d_{ik}}{d_{jk}})^{\frac{2}{(m-1)}}} \text{ for each i and k} \tag{5}$$

$$u_{ik} = \begin{cases} 1 & if\ d_{ik} = 0 \\ 0 & otherwise \end{cases} \tag{6}$$

In the FCM clustering algorithm, as demonstrated in Fig. 4, each cluster's data points are represented with a different color, and each cluster's center is represented with a black X. For all feature vectors, the cluster center matrix, v, is randomly initialized, and the fuzzy partition matrix U is constructed. The main FCM algorithm's outline is explained as follows:

Algorithm: Fuzzy Partitioning Matrix U Construction
Input: Feature vectors
Output: Partition matrix U
Data: Initialize cluster center matrix, v. Set k = 0.

Begin
  Step 1: Initialize the fuzzy clustering matrix, U (k = 0) with (5), (6), m is set to be 2.0.
  Step 2: Increment k=k+1. Compute v using step 3.
  Step 3: Compute U (k) using step 4.
  Step 4: If ‖U (k) -U (k −1) ‖<ε then stop, else repeat steps from 1 to 4.
End

B. Deep Learning

Deep learning (DL) is a type of machine learning and training that uses a significant quantity of data to create an informed model. It is one of the most effective methods for allowing a machine to learn various levels of data properties (e.g., images). The success of every intelligent machine learning algorithm is determined by its features. Clustering algorithms are unsupervised machine-learning algorithms that collect data based on their similarity such as FCM.

C. Proposed System

In this work, we applied the FCM algorithm to deal with the burn area, the burn area is labeled as "white" and the background is marked as "black" in a binary image if the body part outline detection results are accurate. Fig. 5 shows the flow diagram of the suggested system. The system first

converts the input image into a gray image and then entered into the FCM algorithm to segment the injured area, after that an identification of the burn area is obtained. Fig. 5 shows the flow of the steps.

Fig. 6 shows the stages of the proposed system, which mainly consists of three stages: The first stage: is the stage that includes input image, and then converting the picture into a gray color system to facilitate the correct extraction of data. In this work, second and third-degree burns to the skin are taken into consideration.

The second stage: It is the stage through which the data used from the first stage is entered into the directed deep learning algorithm, and through it, the infected part is separated from the healthy part to be transferred to the third stage, which is the stage of data extraction for diagnosis to help the doctor expedite the process processing. Fig. 7 explains the details of the deep layers with fuzzy clustering for each layer until we reach the final result. Fig. 7 demonstrates that the framework contains three parts. The input burn image part, and the image. The second part is the proposal network that extracts the regions of interest, and the third part is the network that output result.
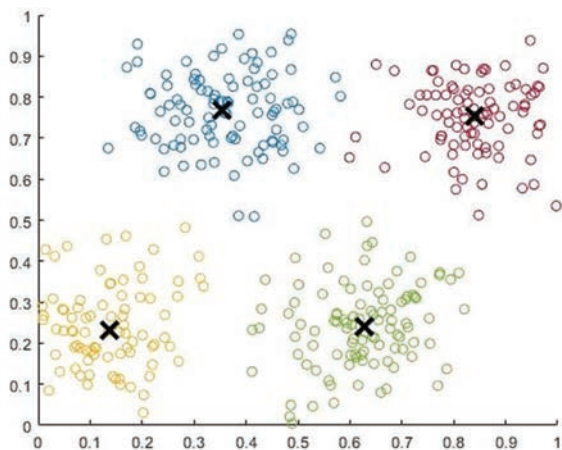
IV. Experimental Results

First, the used database included real burn images under hospital conditions that contain superficial and full-thickness burns, and some input images are from Google search as well. Manual input images, and (Stephen, 2020) database has also been used in our work.

The initial number of segments of the burn images under consideration is determined by experimental results and it decided to set up at 2 clusters in which the required result is to detect the burned area from the unbury areas, as shown in Table I.

System performance measures could be investigated by several parameters. In this work, two sets of metrics are taken into consideration to evaluate the accuracy of the suggested method, the first metric is sensitivity, precision, accuracy, and precision, and the second metric is a calculating mean square error (MSE), and peak sign to noise ratio (PSNR).

Sensitivity and precision were used as performance measures for burn categorization Sensitivity is the rate

Fig. 4. Fuzzy c-means clustering algorithm. Each cluster's data points are represented by a different color. Each cluster's center is represented with a black X.

Fig. 5. Block diagram of the proposed system.
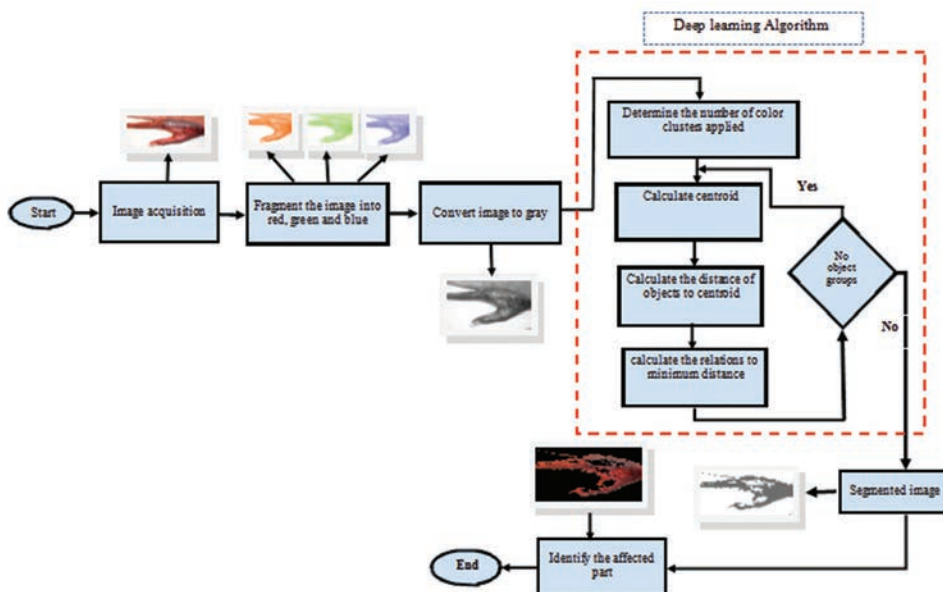
Fig. 6. System flow diagram.
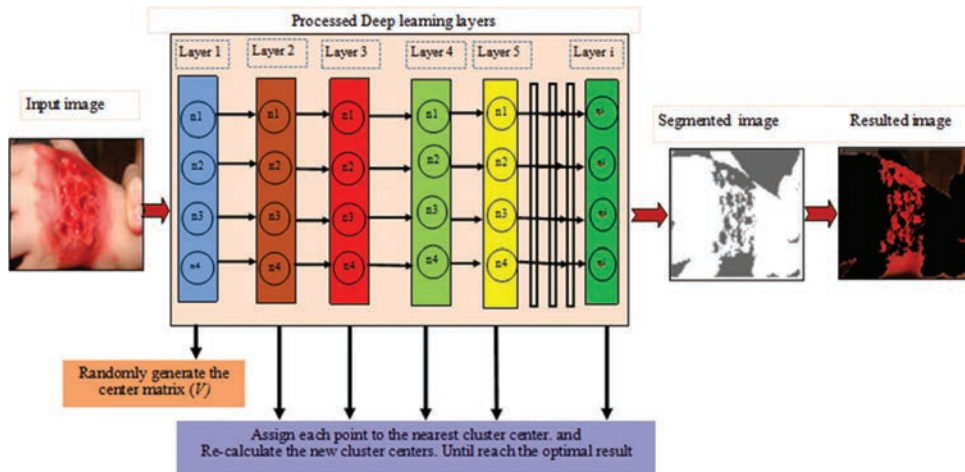
Fig. 7. Graphical representation of deep learning fuzzy c-means layers proposed.
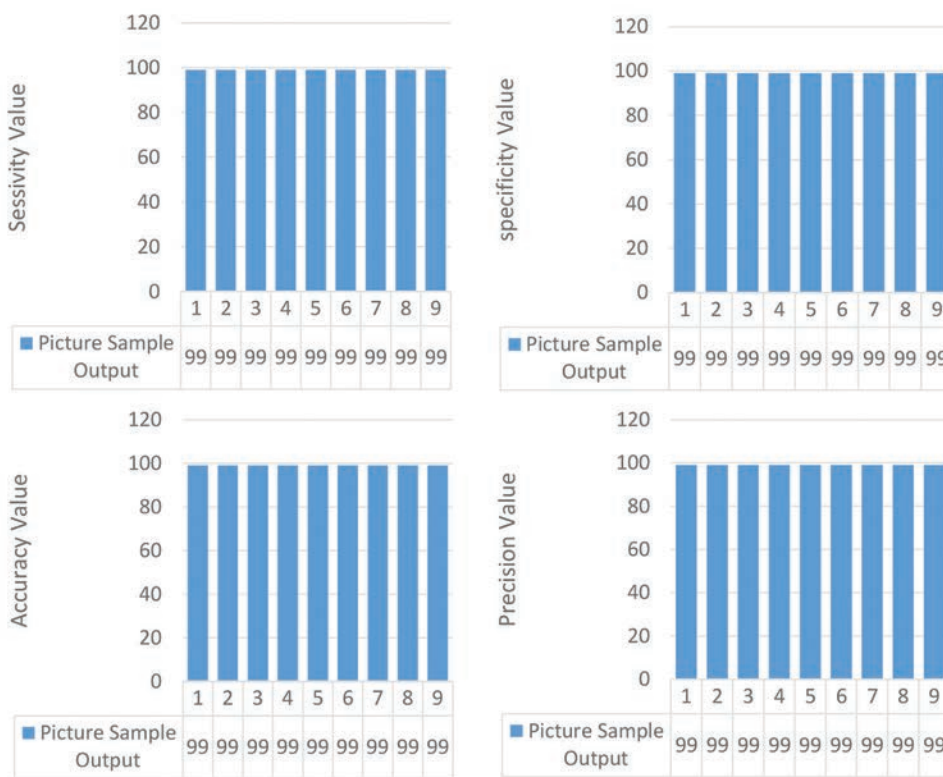
Fig. 8. Performance metric representation, sensitivity, specificity, accuracy, and precision.

<table>
TABLE I
INITIAL ESTIMATES OF K

Input image	Estimation of K
All input images	2
</table>
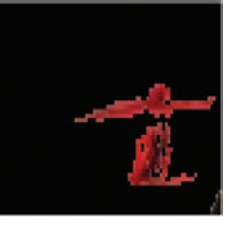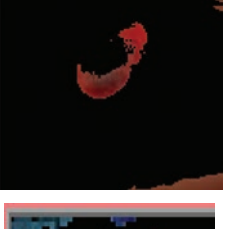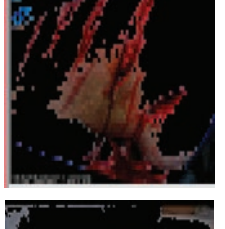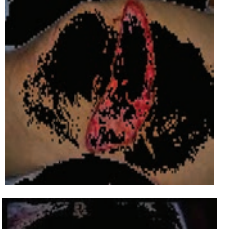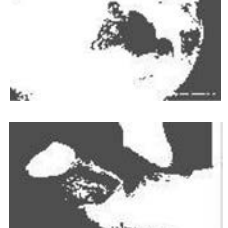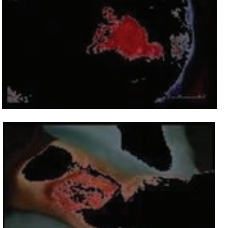
of correctly classifying a burn, sensitivity = TP/(TP + FN); where true positives (TP) are burns images correctly identified, false negatives (FN) are burns wrongly classified. The other performance metric was precision, which is the rate of correctly detected burns, precision = TP/(TP + FP); where false positives (FP) are the area that has been wrongly labeled as burns such as in the darkened skin case, besides

TABLE II
PERFORMANCE METRIC REPRESENTATION, AS SENSITIVITY, SPECIFICITY, ACCURACY, AND PRECISION FOR SOME BURN IMAGES

Image No.	Sensitivity	Specificity	Accuracy	Precision
1	63	99	70	99
2	55	98	59	99
3	61	99	67	99
4	51	90	51	99
5	64	99	72	99
6	60	99	67	99
7	58	98	64	99
8	59	99	65	99
9	65	98	61	99

TABLE III
Results of Applying Deep Fuzzy C-means on Some Burn Images

Image No.	Original image	Segmented image	Resulted image
1			
2			
3			
4			
5			
6			
7			
8			

(Contd...)

TABLE III
(Continued)

Image No.	Original image	Segmented image	Resulted image
9			

TABLE IV
MSE and PSNR Performance Measure for the Segmented Image

Image No.	MSE	PSNR
1	52.2266	39.5413
2	38.8464	40.1840
3	46.0421	39.8150
4	30.6418	40.6992
5	51.5427	39.5699
6	45.8580	39.8237

TABLE V
MSE and PSNR Performance Measure for the Resulted Image

Image No.	MSE	PSNR
1	49.2013	39.6709
2	37.103	40.2837
3	46.2525	39.8051
4	29.9125	40.7515
5	49.3238	39.6655
6	45.1898	39.8555

accuracy, specificity metrics, where accuracy = (TP + TN)/(TP + TN + FP + FN), and specificity = TN/(TN + FP) as shown in Table II. Where the first column represents the number of images.

The accuracy of the system is considered as the maximum accuracy achieved in the system, which is approximately to 75%. In Fig. 8, the four evaluation metrics are depicted for details clarifying.

While the second metric, MSE is used to measure the error between the original input image and the processed image. MSE between two different images such as g (x, y) and ĝ(x, y) is defined as in equation (7):

$$MSE = \frac{1}{MN}\sum_{n=0}^{M}\sum_{m=0}^{N}\left[\hat{g}(n,m) - g(n,m)\right]^2 \tag{7}$$

Sign to noise ratio is a designing term for the power proportion between a sign (significant data) and the background noise as in equation (8):

$$SNR = \left[\frac{\sum_{i=0}^{N-1} wi^2}{MSE}\right] \tag{8}$$

The PSNR addresses the proportion between the peak signal and the mean square mistake, usually calculated to evaluate the objective quality of the processed image after
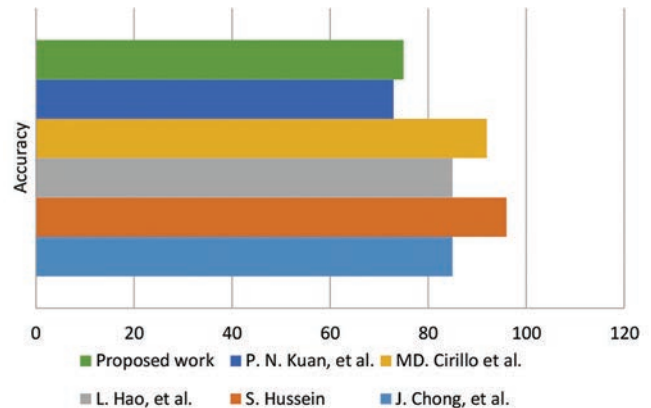

Fig. 9. Accuracy of different approaches.

performing the segmentation process, and it is determined by the equation (9):

$$PSNR(dB) = 10 log_{10}\left[\frac{255^2}{MSE}\right] \tag{9}$$

Typically, PSNR value range from 20 to 40, and they are usually reported up to two decimal points. The actual value is not meaningful, but the image segmentation between the two values for different reconstructed images gives more indications about the quality (Chong, et al., 2019; Umme, Morium and Mohammed, 2019).

Table III shows the segmentation result of the proposed segmentation method for some burn images. The first column represents the original burn image, the second column represents the segmentation result using the deep FCM technique, and the third column represents the segmentation result.

Tables IV and V show the calculation of MSE and PSNR performance measure for the segmented image, and resulted image, respectively.

The categorization of the previously discussed state-of-the-art research in terms of scope of the system, frameworks used, applied techniques, dataset utilized, training and evaluating the number of images, and accuracy, as explained in Table VI.

To evaluate the performance of the proposed method, we compare the segmentation results achieved with our method, Table VII introduces a comparison of the proposed work and some other methods in terms of accuracy, whereas Fig. 8 shows the graphical representation of Table VII.

TABLE VI
COMPARATIVE ANALYSIS OF DIFFERENT RELATED WORKS RESEARCHES

Author name	Scope of the system	Frameworks used	Applied techniques	Data set	Training/Evaluating	Accuracy
Chong, et al.	Classification of burn images	Segmenting burn images	Convolutional neural network (Mask R-CNN)	Common Objects in Context (COCO) data set	1000 burn images for training and 150 for evaluating	84.51%
Badea, et al.	Classification of burn images	Identification of burn areas image patches	Convolutional neural network	Manual dataset, total 611 images depicting burn wounds.	The training and validation sets contain a total of 74763 patches of different images	75.91%
Prabhakar, Gauraveand and Kailesh	Classification of images with color loss	Segmenting of burn photos	Mask Regions of the R-CNN	Wuhan Hospital Burn Department, used own smartphone to collect fresh burn injuries	1000 burn pictures for training and another 150 for assessment	84.5%
Brenda, and Roberto	Classification of burn images	identification and classification of burns images	dictionary learning accompanied by K-singular value decomposition	Public websites	Dictionaries without K-SVD training (L = 1, 2, 3), or K-SVD-trained dictionaries with 30 iterations and various factor values (L = 1, 2, 3) for testing	90%
Murfi, Rosaline and Hariadi	Topic detection for textual data	Topic detection method that combines deep auto encoder (DAE) for representation learning and FCM for fuzzy clustering	Deep auto encoder and fuzzy c-means	Enron – an English email dataset. The first dataset is Enron consisting of 500,000 emails generated by employees of the Enron Corporation	Use pre-trained word2vec model trained on the Google News dataset for the English email dataset/The model contains 300-dimensional vectors for 3 million words and phrases	N/A
				Berita – An Indonesian news dataset. consisting of 50,304 digital Indonesian news articles shared online	Construct the word2vec model using a corpus consisting of 750,000 Indonesian documents from wiki, news, and tweets	
Kuan, et al.	Classification of burn injury images.	both Discrete wavelet transform (DWT) and principle component analysis (PCA), Gray Level Co-occurrence Matrix (GLCM) method	the Waikato Environment for Knowledge Analysis (WEKA)	Both the supplied test set and 10-fold cross validation methods	The dataset splitting into two sets, 70% and 30% for training and testing, respectively	73.2%
Hussein	Segmentation of mammalian body skin (Mice)	Segmenting mice skin layers from images into epidermis, dermis, and adipose layers	Combined color de-convolution method with fuzzy C-mean (FCM) clustering	A dataset of 7,000 mice skin images taken at ×20 magnification	100 images were randomly selected from 1,000 images.	96%
Wang, et al.	Classification of burn images.	A convolutional neural network (CNN), MobileNetV219	Deep convolutional neural networks	The foot ulcer dataset	Training set containing 3645 images and a testing set containing 405 images	90.47%
Despo, et al.	Classification of burn images.	Predicting different burn depths	Applied FCN with a CRF layer	749 images from the Santa Clara Valley Medical Center and 180 images from Google search to create BURNED dataset	334 out of the 396 images in the training set, testing set is 62 images	N/A
Hao, et al.	Segmentation and diagnosis of burn images.	Deep learning used to automate segmentation of burn area and diagnosis.	Fully convolutional Networks	Jianggan District Hospital, Hangzhou. no. of burn images is 1200	960 images are used for training the model and 240 images for testing.	85%
Cirillo, et al.	Segmentation and diagnosis of burn images.	Used burn-depth assessment using semantic segmentation	Convolutional neural network, based on the U-Net	100 burn images	17 images to train the network, and 83 images to evaluate the network.	92%

TABLE VII
COMPARISON OF THE PROPOSED WORK AND SOME OTHER METHODS IN TERM OF ACCURACY

Author names	Accuracy
Chong, et al.	84.51%
Hussein	96%
Hao, et al.	85%
Cirillo, et al.	92%
Kuan, et al.	73.2%
Proposed work	75%

From Fig. 9, we can notice that despite the benefits of using an autonomous wound detection stage preceding the segmentation has its advantages, cases of erroneous wound detection must also be taken into consideration.

IV. CONCLUSION

Image segmentation is regarded as a crucial stage in the processing of images. One of the popular techniques for

segmenting images is the clustering algorithm. However, there are problems with this approach, including sensitivity to initial parameters, getting stuck in local optima, and being unable to identify among objects with identical luminescence. This research proposed a deep learning bases FCM approach to detect the injured area. Experimental results showed that the proposed technique is simple, and effective in the real input image. Two types of metrics have been used to evaluate the system, and a comparison with various types of research is also presented to examine our system correctly. Our future work is to develop the work by combining other machine learning algorithms with this system to classify the degree of burn as superficial, partial thickness¸ or full thickness.

V. Acknowledgment

We would like to express our gratitude for the reviewers' efforts in reading and commenting on our paper. A great thanks my colleagues, for their patient instruction, passionate support, and constructive criticisms of this study effort.

References

Alcala-Fdez, J. and Alonso, J.M., 2016. A survey of fuzzy systems software: Taxonomy, current research trends and prospects. *IEEE Transactions on Fuzzy Systems*, 24(1), pp.40-56.

Badea, M.S., Vertan, C., Florea, C., Florea, L. and Bădoiu, S., 2016. Automatic Burn Area Identification in Color Images. In: *2016 International Conference on Communications* (*COMM*). pp.65-68.

Badrinarayanan, V., Kendall, A. and Cipolla, R., 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), pp.2481-2495.

Bekir, K., 2016. The positive effects of fuzzy C-means clustering on supervised learning classifiers. *International Journal of Artificial Intelligence and Expert Systems*, 7(1), pp.1-8.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Brenda, R.O. and Roberto, R.R., 2021. Detection and classification of burnt skin via sparse representation of signals by over-redundant dictionaries. *Computers in Biology and Medicine*, 132, p.104310.

Chong, J., Kehua, S., Weiguo, X. and Ziqing, Y., 2019. Burn image segmentation based on mask regions with convolutional neural network deep learning framework: More accurate and more convenient. *Burns and Trauma*, 7, p.6.

Cirillo, M.D., Mirdell, R., Sjöberg, F. and Pham, T.D., 2021. Improving burn depth assessment for pediatric scalds by AI based on semantic segmentation of polarized light photography images. *Burns*, 47(7), pp.1586-1593.

Deepak, L., Antony, J. and Niranjan, U.C., 2012. Hardware Co-simulation of Skin Burn Image Analysis. In: *19th IEEE International Conference in High Performance Computing* (HiPC-2012).

Despo, O., Yeung, S., Jopling, J., Pridgen, B., Sheckter, C., Silberstein, S., Fei-Fei, L. and Milstein, A., 2017. BURNED: Towards Efficient and Accurate Burn Prognosis Using Deep Learning. Avaliable from: http://cs231n.stanford.edu/reports/2017/pdfs/507.pdf [Last accessed on 2022 Feb 01].

Domagoj, M. and Damir, F., 2020. A systematic overview of recent methods for non-contact chronic wound analysis. *Applied Science*, 10, p.7613.

Fangzhao, L., Changjian, W., Xiaohui, L., Yuxing, P. and Shiyao, J., 2018. A composite model of wound segmentation based on traditional methods and deep neural networks. *Computational Intelligence and Neuroscience*, 2018,

p.4149103.

Hansen, G.L., Sparrow, E.M., Kokate, J.Y., Leland, K.J. and Iaizzo, P.A., 1997. Wound status evaluation using color image processing. *IEEE Transactions on Medical Imaging*, 16(1), pp.78-86.

Hao, L., Kepiang, Y., Siyi, C., Wenjun, L. and Zhihui, F., 2021. A framework for automatic burn image segmentation and burn depth diagnosis using deep learning. *Computational and Mathematical Methods in Medicine*, 2021, p. 5514224.

Hasan, M.M., Ibraheem, N.A. and Abdulhadi, N.M., 2022. 2D geometric object shapes detection and classification. *Webology*, 19(1), pp.1689-1702.

Health Encyclopedia Site. Available from: https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID=90&ContentID=P09575 [Last accessed on 2022 Oct 01].

Hussein, S., 2021. Automatic layer segmentation in H&E images of mice skin based on colour deconvolution and fuzzy C-mean clustering. *Informatics in Medicine Unlocked*, 25, p.100692.

Kuan, P.N., Chua, S., Safawi, E.B., Wang, H.H. and Tiong, W., 2018. A comparative study of the classification of skin burn depth in human. *Journal of Telecommunication Electronic and Computer Engineering*, 9(2-10), pp.15-23.

Liangrui, P., Zhichao, F. and Shaoliang, P., 2022. A review of machine learning approaches, challenges and prospects for computational tumor pathology. *arXiv*, 2206, p.01728.

Liu, X., Song, L., Liu, S. and Zhang, Y., 2021. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3), p.1224.

Malini, S., Siva, K. and Niranjan, U., 2013. Classification methods of skin burn images. *International Journal of Computer Science and Information Technology*, 5(1), pp.109-118.

Miller, S.J., Burke, E.M., Rader, M.D., Coulombe, P.A., and Lavker, R.M., 1998. Reepithelialization of porcine skin by the sweat apparatus. *The Journal of Investigative Dermatology*, 110(1), pp.13-19.

Murfi, H., Rosaline, N. and Hariadi, N., 2022. Deep autoencoder-based fuzzy c-means for topic detection, *Array*, 13, p.100124.

Nameirakpam, D., Khumanthem, M. and Yambem, J.C., 2015. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54, pp.764-771.

Papini, R., 2004. Management of burn injuries of various depths. *British Medical Journal*, 329(7458), pp. 158-160.

Prabhakar, K., Gaurave, S. and Kailesh, P., 2020. Burn image segmentation based on mask regions with convolutional neural network deep learning framework. *International Journal of Research in Engineering Science and Management*, 3(8), pp.478-482.

Rahman, T. and Islam, M.S., 2021. Image Segmentation Based on Fuzzy C Means Clustering Algorithm and Morphological Reconstruction. In: *International Conference on Information and Communication Technology for Sustainable Development* (ICICT4SD). pp.259-263.

Sabeena, B. and Rajkumar, P., 2019. Diagnosis and detection of skin burn analysis segmentation in colour skin images. *International Journal of Advanced Research in Computer and Communication Engineering*, 6(2), pp. 369-374.

Stephen, T., 2020. Stock Pictures of Wounds. Medetec Wound Database. Available from: https://www.medetec.co.uk/files/medetec-image-databases.html [Last accessed on 2022 Feb 01].

Tina, M., Uros, M., Karin, S.K., Raščan, I.M. and Dragica, M.S., 2015. Advanced therapies of skin injuries. *Wiener Klinische Wochenschrift*, 127(5), pp.187-198.

Ugur, U.K., Erdinç, K., Tolga, B., Yesim, A. and Serder, T., 2019. Automatic classification of skin burn colour images using texture-based feature extraction. *IET Image Processing*, 13(11), pp.2018-2028.

Umme, S., Morium, A. and Mohammed, S.U., 2019. Image quality assessment through FSIM, SSIM, MSE and PSNR-a comparative study. *Journal of Computer*

and Communications, 7(3), pp.8-18.

Wagh, A., Jain, S., Mukherjee, A., Agu, E., Pedersen, P.C., Strong, D., Tulu, B. Lindsay, C. and Liu, Z., 2020. Semantic segmentation of smartphone wound images: Comparative analysis of AHRF and CNN-based approaches. *IEEE Access*, 8, pp.181590-181604.

Wang, C., Anisuzzaman, D.M., Williamson, V., Mrinal, K.D., Behrouz, R., Jeffrey, G., Sandeep, G. and Zeyen, Y., 2020. Fully automatic wound segmentation with deep convolutional neural networks. *Scientific Reports*, 10(1), p.21897.

Wang, C., Pedrycz, W., Li, Z. and Zhou, M., 2021. Residual-driven fuzzy C-means clustering for image segmentation. *IEEE/CAA Journal of Automatica Sinica*, 8(4), pp.876-889.

Yeganejou, M. and Dick, S., 2018. Classification Via Deep Fuzzy C-means Clustering. In: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018. pp.1-6.

General Information

ARO's Mission: ARO seeks to publish those papers that are most influential in their fields or across fields and that will significantly advance scientific understanding. Selected papers should present novel and broadly important data, syntheses, or concepts. They should merit the recognition by the scientific community and general public provided by publication in ARO, beyond that provided by specialty journals.

We welcome submissions from all fields of natural science and technology, and from any source. We are committed to the prompt evaluation and publication of submitted papers. ARO is published biannually; selected papers are published online ahead of print.

Submission

Manuscripts should be submitted by the correspondent authors of the manuscript via the on-line submission page. Regardless of the source of the word-processing tool, only electronic Word (.doc, .docx, .rtf) files can be submitted on-line. There is no page limit. Only online submissions are accepted to facilitate rapid publication and minimize administrative costs. Submissions by any other one but the authors will not be accepted. The submitting author takes responsibility for the paper during submission and peer review. If for some technical reason submission through the email is not possible, the author can contact aro.journal@koyauniversity.org for support. Before submitting please check ARO's guide to authors thoroughly to avoid any delay in the review and publication process.

Authors are explicitly responsible for the language of their texts. Paper should be submitted in a well written in understandable English. Authors should not expect the editor or editorial board to rewrite their paper. Prior to submission, authors should have their paper proofread by a possible academic native speaker of English.

- Submit the Article with contact Information
- File name should be your article tittle
- Don't submit your article in multiple journal, we are taking only minimum time for review process. please don't waste our time
- Once the paper is accepted, it can't be withdrawn
- Please follow publication ethics and regulation
- Avoid plagiarism and copied material
- Strictly Follow ARO's template

Terms of Submission

Papers must be submitted on the understanding that they have not been published elsewhere and are not currently under consideration by another journal or any other publisher. ARO accepts original articles with novel impacts only. Post conference papers are not accepted "as is", however, regular papers on the same topic but with a different title can be submitted. The new paper should contain significant improvements in terms of extended content, analysis, comparisons with popular methods, results, figures, comments, etc. Please do not forget that the publication of the same or similar material in ARO constitutes the grounds for filing of an (auto) plagiarism case.

The submitting author is responsible for ensuring that the article's publication has been approved by all the other co-authors. It is also the authors' responsibility to ensure that the articles emanating from a particular institution are submitted with the approval of the necessary institution. Only an acknowledgement from the editorial office officially establishes the date of receipt. Further correspondence and proofs will be sent to the author(s) before publication unless otherwise indicated. It is a condition of submission of a paper that the authors permit editing of the paper for readability. All enquiries concerning the publication of accepted papers should be addressed to aro.journal@koyauniversity.org.

Peer Review

All manuscripts are subject to peer review and are expected to meet standards of academic excellence. Submissions will be considered by an editor and "if not rejected right away" by peer-reviewers, whose identities will remain anonymous to the authors.

Guide to Author

We welcome submissions from all fields of science and from any source. We are committed to the prompt evaluation and publication of submitted papers. Selected papers are published online ahead of print. Authors are encouraged to read the instructions below before submitting their manuscripts. This section arranged into an overview speedy guidelines below and more detailed at the bottom section of this page

Manuscript Preparation

Submitting your manuscript will be in two stages namely before final acceptance and after.

Stage One:

For the initial submission, the manuscript should be prepared electronically in Microsoft Word (.doc, .docx, .rtf) and PDF formats. Submit it through the online submission system after completing the registration. The Word file should be in a single-column format, double-spaced, with Times New Roman font, and 12-point font size. The authors' names and affiliations should be removed from the manuscript for the double-blind review process. Referencing and citation should follow the Harvard/ARO system. You can download the stage-one manuscript template by clicking here.

Stage Two:

Once the manuscript is accepted, the production team of ARO Journal will prepare the camera-ready paper.

Units of Measurement

Units of measurement should be presented simply and concisely using System International (SI) units.

Title and Authorship Information

The following information should be included;
- Paper title.
- Full author names.
- Affiliation.
- Email addresses.

Abstract

The manuscript should contain an abstract. The abstract should be self-contained and citation-free and should not exceed 250 words.

Introduction

This section should be succinct, with no subheadings.

Materials and Methods

This part should contain sufficient detail so that all procedures can be repeated. It can be divided into subsections if several methods are described.

Results and Discussion

This section may each be divided by subheadings or may be combined.

Conclusions

This should clearly explain the main conclusions of the work highlighting its importance and relevance.

Acknowledgements

All acknowledgements (if any) should be included at the very end of the paper before the references and may include supporting grants, presentations, and so forth.

References

References must be included in the manuscript and authors are responsible for the accuracy of references. Manuscripts without them will be returned. ARO is following Harvard System of Referencing. (Learn how to import and use Harvard Styling in your Microsoft Office by following this link:

http://bibword.codeplex.com/releases/view/15852)

Preparation of Figures

Upon submission of an article, authors are supposed to include all figures and tables in the PDF file of the manuscript. Figures and tables should be embedded in the manuscript. Figures should be supplied in either vector art formats (Illustrator, EPS, WMF, FreeHand, CorelDraw, PowerPoint, Excel, etc.) or bitmap formats (Photoshop, TIFF, GIF, JPEG, etc.). Bitmap images should be of 300 dpi resolution at least unless the resolution is intentionally set to a lower level for scientific reasons. If a bitmap image has labels, the image and labels should be embedded in separate layers.

Preparation of Tables

Tables should be cited consecutively in the text. Every table must have a descriptive title and if numerical measurements are given, the units should be included in the column heading. Vertical rules should not be used.

Copyright

ARO Reviewer/Associate Editor Application Form

ARO is a scientific journal of Koya University (p-ISSN: 2410-9355, e-ISSN: 2307-549X) which aims to offer a novel contribution to the study of Science. The purpose of ARO is twofold: first, it will aim to become an ongoing forum for debate and discussion across the sciences and Engineering. We hope to advance our problem solving capacity and deepen our knowledge regarding a comprehensive range of collective actions. Second, ARO accepts the challenges brought about by multidisciplinary scientific areas and aspires to expand the community of academics who are able to learn from and help to produce advances in a variety of different disciplines.

The Journal is seeking reviewers who can provide constructive analysis of papers thus enhancing overall reputation of the Journal. If any expert is interested in participating of the review process, we highly encourage you to sign up as a reviewer for our Journal and help us improve our presence in domain of your expertise. Appropriate selection of reviewers who have expertise and interest in the domain relevant to each manuscript are essential elements that ensure a timely, productive peer review process. We require proficiency in English.

How to apply

To apply for becoming a reviewer of ARO, please submit the application form by following the link:
https://aro.koyauniversity.org/user/register

To apply for becoming a member of the Editorial Board of ARO, please submit the application form by completing the application form.

Both Associate Editor and Reviewers should specify their areas of research and expertise. Applicants must have a doctorate (or an equivalent degree), and if Master degree they need to have significant publishing experience. Please note that;

- You will need to write your full official name.
- Please provide an email which reflects your official name, such as nameOne.NameTwo@... , or your institute's official email.
- All data need to be written in English.

Note: For more information, kindly visit the following websites:

1. aro.koyauniversity.org.
2. http://libweb.anglia.ac.uk/referencing/harvard.htm.
3. http://bibword.codeplex.com/releases/view/15852.

KOYA
UNIVERSITY

Koya University is a young University established in 2003 and it is located in the city of Koya (Koysinjaq), short distance to the East of regional capital city of Erbil (Arbil, Hewlêr) in Kurdistan Region of Iraq. It is on the foothills of beautiful High Mountain. Its campus has been carefully laid out to embrace the beautiful mountainous nature. The Koya University has a Faculty system which enhances the interactions between similar academic fields. Today, Koya University has four Faculties: Engineering, Science and Health, Humanities and Social Sciences and Education in addition to the School of Medicine, which all consist of twenty-five scientific departments in different fields, such as Petroleum Engineering, Geotechnical Engineering, Software Engineering, Physics, Chemistry, Clinical Psychology, Social Science, Medical Microbiology and Sport Education.

ARO-The Scientific Journal of Koya University is a biannual journal of original scientific research, global news, and commentary in the areas of Science and Technology. ARO is a Peer-reviewed Open Access journal with CC BY-NC-SA 4.0 license. It provides immediate, worldwide and barrier-free access to the full text of research articles without requiring a subscription to the journal, and has no article processing charge (APC). ARO Journal seeks to publish those papers that are most influential in their fields or across fields and that will significantly advance scientific understanding. ARO Journal is a member of ROAD and Crossref agencies and has got ESCI, DOAJ seal, SHERPA/RoMEO deposit policy, and LOCKSS archiving policy.

ARO

The Scientific Journal of Koya University

Koya University (KOU)
University Park
Danielle Mitterrand Boulevard
Koya KOY45, Kurdistan Region - Iraq