زانکۆی گەشەپێدانی مرۆیی
اقرأ
جامعة التنمية البشرية
UNIVERSITY OF HUMAN DEVELOPMENT

p-ISSN 2521-4209
e-ISSN 2521-4217

UHD Journal of Science and Technology

A Scientific periodical issued by University of Human Developement

Vol.4 No.(1) June 2020

2020 2720

www.jst.uhd.edu.iq

Kurdistan Regional Government
University of Human Development

UHD Journal of Science and Technology

A periodic scientific journal issued by University of Human Development

Editorial Board

Technical

Advisory Board

Introduction

UHD Journal of Science and Technology (UHDJST) is a semi-annual journal published by the University of Human Development, Sulaymaniyah, Kurdistan Region, Iraq. UHDJST member of ROAD, e-ISSN: 2521-4217, p-ISSN: 2521-4209 and a member of Crossref, DOI: 10.21928/issn.2521-4217. UHDJST publishes original research in all areas of Science, Engineering, and Technology. UHDJST is a Peer-Reviewed Open Access journal with Creative Commons Attribution Non-Commercial No Derivatives License 4.0 (CC BY-NC-ND 4.0). UHDJST provides immediate, worldwide, barrier-free access to the full text of research articles without requiring a subscription to the journal, and has article processing charge (APC). UHDJST applies the highest standards to everything it does and adopts APA citation/referencing style. UHDJST Section Policy includes three types of publications: Articles, Review Articles, and Letters.

By publishing with us, your research will get the coverage and attention it deserves. Open access and continuous online publication mean your work will be published swiftly, ready to be accessed by anyone, anywhere, at any time. Article Level Metrics allow you to follow the conversations your work has started.

UHDJST publishes works from extensive fields including, but not limited to:

- Pure Science
- Applied Science
- Medicine
- Engineering
- Technology

Scope and Focus

UHD Journal of Science and Technology (UHDJST) publishes original research in all areas of Science and Engineering. UHDJST is a semi-annual journal published by the University of Human Development, Sulaymaniyah, Kurdistan Region, Iraq. We believe that if your research is scientifically valid and technically sound then it deserves to be published and made accessible to the research community. UHDJST aims to provide a service to the international scientific community enhancing swap space to share, promote and disseminate the academic scientific production from research applied to Science, Engineering, and Technology.

SEARCHING FOR PLAGIARISM

We use plagiarism detection: detection; According to Oxford online dictionary, Plagiarism means: *The practice of taking someone else's work or ideas and passing them off as one's own.*

Section Policies

No.	Title	Peer Reviewed	Indexed	Open Submission
1	Articles: This is the main type of publication that UHDJST will produce	☑	☑	☑
2	Review Articles: Critical, constructive analysis of the literature in a specific field through summary, classification, analysis, comparison.	☑	☑	☑
3	Letters: Short reports of original research focused on an outstanding finding whose importance means that it will be of interest to scientists in other fields.	☑	☑	☑

PEER REVIEW POLICIES

At UHDJST we are committed to prompt quality scientific work with local and global impacts. To maintain a high-quality publication, all submissions undergo a rigorous review process. Characteristics of the peer review process are as follows:

- The journal peer review process is a "double-blind peer review".
- Simultaneous submissions of the same manuscript to different journals will not be tolerated.
- Manuscripts with contents outside the scope will not be considered for review.
- Papers will be refereed by at least 2 experts as suggested by the editorial board.
- In addition, Editors will have the option of seeking additional reviews when needed. Authors will be informed when Editors decide further review is required.
- All publication decisions are made by the journal's Editors-in-Chief on the basis of the referees' reports. Authors of papers that are not accepted are notified promptly.
- All submitted manuscripts are treated as confidential documents. We expect our Board of Reviewing Editors, Associate Editors and reviewers to treat manuscripts as confidential material as well.
- Editors, Associate Editors, and reviewers involved in the review process should disclose conflicts of interest resulting from direct competitive, collaborative, or other relationships with any of the authors, and remove oneself from cases in which such conflicts preclude an objective evaluation. Privileged information or ideas that are obtained through peer review must not be used for competitive gain.
- Our peer review process is confidential and the identities of reviewers cannot be revealed.

Note: UHDJST is a member of CrossRef and CrossRef services, e.g., CrossCheck. All manuscripts submitted will be checked for plagiarism (copying text or results from other sources) and self-plagiarism (duplicating substantial parts of authors' own published work without giving the appropriate references) using the CrossCheck database. Plagiarism is not tolerated.

For more information about CrossCheck/iThenticate, please visit
http://www.crossref.org/crosscheck.html.

OPEN ACCESS POLICY

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge. Open Access (OA) stands for unrestricted access and unrestricted reuse which means making research publications freely available online. It access ensures that your work reaches the widest possible audience and that your fellow researchers can use and share it easily. The mission of the UHDJST is to improve the culture of scientific publications by supporting bright minds in science and public engagement.

UHDJST's open access articles are published under a Creative Commons Attribution CC-BY-NC-ND 4.0 license. This license lets you retain copyright and others may not use the material for commercial purposes. Commercial use is one primarily intended for commercial advantage or monetary compensation. If others remix, transform or build upon the material, they may not distribute the modified material. The main output of research, in general, is new ideas and knowledge, which the UHDJST peer-review policy allows publishing as high-quality, peer-reviewed research articles. The UHDJST believes that maximizing the distribution of these publications - by providing free, online access - is the most effective way of ensuring that the research we fund can be accessed, read and built upon. In turn, this will foster a richer research culture and cultivate good research ethics as well. The UHDJST, therefore, supports unrestricted access to the published materials on its main website as a fundamental part of its mission and a global academic community benefit to be encouraged wherever possible.

Specifically:

- The University of Human Development supports the principles and objectives of Open Access and Open Science
- UHDJST expects authors of research papers, and manuscripts to maximize the opportunities to make their results available for free access on its final peer-reviewed paper
- All manuscript will be made open access online soon after final stage peer-review finalized.
- This policy will be effective from 17th May 2017 and will be reviewed during the first year of operation.
- Open Access route is available at http://journals.uhd.edu.iq/index.php/uhdjst for publishing and archiving all accepted papers,
- Specific details of how authors of research articles are required to comply with this policy can be found in the Guide to Authors.

ARCHIVING

This journal utilizes the LOCKSS and CLOCKSS systems to create a distributed archiving system among participating libraries and permits those libraries to create permanent archives of the journal for purposes of preservation and restoration.

LOCKSS: Open Journal Systems supports the LOCKSS (Lots of Copies Keep Stuff Safe) system to ensure a secure and permanent archive for the journal. LOCKSS is open source software developed at Stanford University Library that enables libraries to preserve selected web journals by regularly polling registered journal websites for newly published content and archiving it. Each archive is continually validated against other library caches, and if the content is found to be corrupted or lost, the other caches or the journal is used to restore it.

CLOCKSS: Open Journal Systems also supports the CLOCKSS (Controlled Lots of Copies Keep Stuff Safe) system to ensure a secure and permanent archive for the journal. CLOCKSS is based upon the open-source LOCKSS software developed at Stanford University Library that enables libraries to preserve selected web journals by regularly polling registered journal websites for newly published content and archiving it. Each archive is continually validated against other library caches, and if the content is found to be corrupted or lost, the other caches or the journal is used to restore it.

PUBLICATION ETHICS

Publication Ethics and Publication Malpractice Statement

The publication of an article in the peer-reviewed journal UHDJST is to support the standard and respected knowledge transfer network. Our publication ethics and publication malpractice statement is mainly based on the Code of Conduct and Best-Practice Guidelines for Journal Editors (Committee on Publication Ethics, 2011) that includes;

- General duties and responsibilities of editors.
- Relations with readers.
- Relations with the authors.
- Relations with editors.
- Relations with editorial board members.
- Relations with journal owners and publishers.
- Editorial and peer review processes.
- Protecting individual data.
- Encouraging ethical research (e.g. research involving humans or animals).
- Dealing with possible misconduct.
- Ensuring the integrity of the academic record.
- Intellectual property.
- Encouraging debate.
- Complaints.
- Conflicts of interest.

ANIMAL RESEARCHES

- For research conducted on regulated animals (which includes all live vertebrates and/or higher invertebrates), appropriate approval must have been obtained according to either international or local laws and regulations. Before conducting the research, approval must have been obtained from the relevant body (in most cases an Institutional Review Board, or Ethics Committee). The authors must provide an ethics statement as part of their Methods section detailing full information as to their approval (including the name of the granting organization, and the approval reference numbers). If an approval reference number is not provided, written approval must be provided as a confidential supplemental information file. Research on non-human primates is subject to specific guidelines from the Weather all (2006) report (The Use of Non-Human Primates in Research).
- For research conducted on non-regulated animals, a statement should be made as to why ethical approval was not required.
- Experimental animals should have been handled according to the highest standards dictated by the author's institution.
- We strongly encourage all authors to comply with the '*Animal Research: Reporting In Vivo Experiments*' (ARRIVE) guidelines, developed by NC3Rs.
- Articles should be specific in descriptions of the organism(s) used in the study. The description should indicate strain names when known.

ARTICLE PROCESSING CHARGES

UHDJST is an Open Access Journal (OAJ) and has article processing charges (APCs). The published articles can be downloaded freely without a barrier of admission.

Address

University of Human Development, Sulaymaniyah-Kurdistan Region/Iraq
PO Box: Sulaymaniyah 6/0778

Contact

Principal Contact

Dr. Aso Darwesh

Editor-in-Chief

University of Human Development –
Sulaymaniyah, Iraq

Phone: +964 770 148 5879

Email: jst@uhd.edu.iq

Support Contact

UHD Technical Support

Phone: +964 770 158 4888

Email: jst@uhd.edu.iq

Contents

FULL LENGTH ARTICLE | UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Quality Improvement for Exemplar-based Image Inpainting using a Modified Searching Mechanism

Mariwan Wahid Ahmed[1], **Alan Anwer Abdulla**[2]

[1]Department of Computer, College of Science, University of Sulaimani, Sulaymaniyah, Iraq, [2]Department of Information Technology, College of Commerce, University of Sulaimani, Sulaymaniyah, Iraq, [2]Department of Information Technology, University College of Goizha, Sulaymaniyah, Iraq

ABSTRACT

Digital image processing has a significant impact in different research areas including medical image processing, biometrics, image inpainting, object detection, information hiding, and image compression. Image inpainting is a science of reconstructing damaged parts of digital images and filling-in regions in which information are missing which has many potential applications such as repairing scratched images, removing unwanted objects, filling missing area, and repairing old images. In this paper, an image inpainting algorithm is developed based on exemplar, which is one of the most important and popular images inpainting technique, to fill-in missing area that caused either by removing unwanted objects, by image compression, by scratching image, or by image transformation through internet. In general, image inpainting consists of two main steps: The first one is the priority function. In this step, the algorithm decides to select which patch has the highest priority to be filled at the first. The second step is the searching mechanism to find the most similar patch to the selected highest priority patch to be inpainted. This paper concerns the second step and an improved searching mechanism is proposed to select the most similar patch. The proposed approach entails three steps: (1) Euclidean distance is used to find the similarity between the highest priority patches which need to be inpainted with each patch of the input image, (2) the position/location distance between those two patches is calculated, and (3) the resulted value from the first step is summed with the resulted value obtained from the second step. These steps are repeated until the last patch from the input image is checked. Finally, the smallest distance value obtained in step 3 is selected as the most similar patch. Experimental results demonstrated that the proposed approach gained a higher quality in terms of both objectives and subjective compared to other existing algorithms.

Index Terms: Image Inpainting, Sum of Square Difference, Image Quality, Peak Signal-to-noise Ratio, Position Distance

1. INTRODUCTION

The image processing techniques have been developed in the various of areas such as remote sensing, video processing, image inpainting, medical image processing, pattern recognition, biometrics, traffic systems, image compression, information hiding [1], and printing industry [2]. Image inpainting is the process of reconstructing the damaged image and/or removing an object in an image [3], [4]. The fundamental concept of image inpainting is to fill the missing regions of an image using the surrounding information of that region. Image inpainting techniques entail many applications such as automatic text removal and/or object removal in images/films for special effects [5], deleting of blurs of dust in image, red-eye correction, inventive effect by deleting object [6], [7], remove video logos [8], old image/film restoration [9], and medical image correction

Access this article online

DOI: 10.21928/uhdjst.v4n1y2020.pp1-8

E-ISSN: 2521-4217

P-ISSN: 2521-4209

Corresponding author's e-mail: Alan Anwer Abdulla, Department of Information Technology, College of Commerce, University of Sulaimani and University college of Goizha, Sulaimani, Iraq. E-mail: Alan.abdulla@univsul.edu.iq

Received: 24-12-2019 Accepted: 19-01-2020 Published: 21-01-2020

(image inpainting use to complete missing or distorted information in medical images) [10]. In general, image inpainting techniques are categorized into five major types which are: (1) Partial differential equation based inpainting, (2) texture synthesis based inpainting, (3) exemplar-based inpainting, (4) semi-automatic and fast inpainting, and (5) hybrid inpainting [11], [12]. Each of these types has its own advantages and limitations and each of them recovers the damaged regions in accordance with certain requirements of expectation of the repaired image content. This paper focuses on the third type, i.e., exemplar-based image inpainting.

The exemplar-based image inpainting is one of the most important types of inpainting techniques and it has been proved to be the most effective type of inpainting [13]. Exemplar-based image inpainting mainly includes two main steps. The first step is the priority assignment, in this step, one patch from the boundary of the missing region with the highest priority must be selected using one of the priority functions and it must be filled at first. The second step involves searching for the best-matched patch based on one of the searching mechanisms. In this area of research, researchers try to improve one or both of these steps to obtain better results in terms of the quality of the reconstructed image. This paper focuses on improving the searching mechanism by taking advantage from the distance between the patch to be inpainted and the other best match patches.

The rest of the paper is organized as follows: In section 2, the literature review is presented. Section 3 illustrates the statement of the problem. Section 4 presents the proposed inpainting approach. The experimental results are shown in Section 5. Finally, our conclusions are given in Section 6.

2. LITERATURE REVIEW

Different categories of image inpainting are mentioned in the previous section. Since the contribution in this paper concerns on the exemplar-based image inpainting, therefore this section is going to review some important approaches related to exemplar-based image inpainting.

The idea of exemplar-based inpainting is first invented by Perez *et al.* in 2003 [14]. The core of this algorithm is an isophote driven image sampling process, isophote refers to the direction and intensity of the patch center point. Patch can be defined as a small (generally rectangular) piece/block of an image. Essentially, in image inpainting, Ψp refers to the targeted patch in which some of its pixels values are missing and Ψq refers to the best match founded patch. Moreover, one of the most important steps in image inpainting is from which patch (out of all the patches located around the border of the missing area) should the system start to inpainting; this process refers to priority function and it denotes as P(p). In Perez *et al.* [14], the priority function of their algorithm is defined by a production of confidence term C(p) and data term D(p).

$$P(p)=C(p)\times D(p) \tag{1}$$

Moreover, the confidence term represents texture information and data term represents structure information. These two parameters decide which patch has the highest priority. Furthermore, Ψp will be filled by the Ψq, only the missing part of the Ψp will replace by the Ψq. In each iteration, once the best match patch Ψq is founded, it will immediately use to fill the missing part of the targeted patch Ψp and then the next iteration will start to find the best match patch Ψq for the next targeted patch Ψp. This process continues until all the targeted patches are filled completely [14].

The extended version of the previous work was published by Perez *et al.* in 2004 with a more detailed description of the algorithm and extensive experiments [15].

In general, each exemplar-based image inpainting approach includes two main steps. The first one is the priority function that uses to decide from which patch, out of all the patches located around the boundary of the missing area, the algorithm should start. On the other hand, the second phase is the searching mechanism using distance technique, for example, sum of square difference that is used to select the best-matched patch based on the difference between the targeted patch Ψp and the best match patch Ψq. Therefore, all the approaches proposed in this research area are either improve the priority function, improve the way of finding differences between Ψp and Ψq, or improve both steps. In addition, the essential purpose of the competition in this research area is to improve the quality of the reconstructed image.

Wen-Huang *et al.* highlighted the limitation of the previous algorithm in Perez *et al.* [14] and Perez *et al.* [15] in which the defined priority function quickly drops the value of the confidence term C(p) to zero [16]. Consequently, this leads to create a problem in which the priority function P(p) is gradually depend on only structure information D(p) without making the texture information C(p) into consideration;

therefore, the priority function is not able to find the highest priority patch correctly. This kind of drawback is known as the dropping effect. To overcome this drawback, Wen-Huang et. al. proposed an algorithm in which defined a regularized function Rc(p) instead of confidence term C(p) as follows:

$$R_c(p)=(1-\omega)\times C(p)+\omega \tag{2}$$

Where ω is the regularizing factor for controlling curve smoothness and takes a value between 0 and 1. In general, in this work, ω sets to 0.7. Finally, the equation changed as:

$$RP(p)=\alpha.R_c(p)+\beta.D(p) \tag{3}$$

Where $0\leq\alpha$, $\beta\leq 1$, and $\alpha+\beta=1$. Authors claimed that their proposed approach is improved the quality of the reconstructed image [16].

In the previously reviewed approaches, only one patch Ψq is used to reconstruct the inpainted patch Ψp. In 2008, Alexander *et al.* proposed a new mechanism to fill the inpainted patches using multiple samples within the image and weight their contribution according to their similarity to the neighborhood under evaluation. Using the weighted aggregation of multiple patches yields a much-improved result [17].

In 2015, Huang *et al.* defined a new priority function for exemplar-based image inpainting [18]. Different from the priority function defined in Perez *et al.* [14], Hsieh *et al.* [16], this proposed priority function first rely on the structure information D(p), and then used texture information C(p). In other words, this priority function works on the D(p) and C(p) separately. This approach works well for the case of the curve or cross-shaped structures. The presented results in this paper show that this approach is able to recover image geometry and texture well compared to other existing approaches [18].

A new mechanism was proposed by Liu *et al.*, in 2018, in which the proposed algorithm added boundary-constrained similar patches to find more suitable similar patches and boundary filling instead of patch filling. To measure the similarity between boundaries, they used boundary matching distance. This type of distance measurement was applied for the boundary matching algorithm. This proposed algorithm can repair structural and texture complex images very well [19].

Awati *et al.*, in 2018, proposed a new technique to enhance the quality of exemplar-based image inpainting [20]. This algorithm upgrades the image inpainting approaches mainly for these images that involve edges and corner information. This approach uses fractional derivative and additional curvature finding methods for finding data term in patch priority.

The proposed approach presents in this paper will modify the searching mechanism, while uses the same priority function presented in Hsieh *et al.* [16], more details will be given in the next section.

3. STATEMENT OF THE PROBLEM

The digital image may cause to lose some of it is regions affecting by removing unwanted objects, by image compression, by scratching image, or by image transformation through internet. Image inpainting is becoming a common tool for repairing the scratched or damaged area of an image. It entails reconstructing damaged parts and filling-in regions in which data/color information are missing.

4. PROPOSED APPROACH

This section gives the steps of the proposed approach in detail. It is assumed that the source region, target region, and the boundary denote by Φ, Ω, and δΩ, respectively. Furthermore, the target patch denotes by Ψp in which its center point is p δΩ [21]. The proposed approach consists of the following steps:

4.1. Manually Select the Target Region Ω

Selecting the target region means to select the degraded region or unwanted region to be filled using the surrounding information.

In this step, the target region is selected by marking this area using a specific color, for example, “black color,” and then the mask image is also created. Mask image is a binary image in which zero pixels indicate the target region and the ones are showing the remaining part of the image, as illustrated in Fig. 1.

4.2. Target Region’s Boundary Detection δΩ

One of the most important steps of developing the inpainting algorithm is the boundary detection of the target region because the selection of the highest priority patch from the boundary of its target region can be decided. The boundary of the target region should be detected using one of the edge detection techniques such as Prewitt, Sobel, Canny, or Laplacian edge detection see Fig. 2.

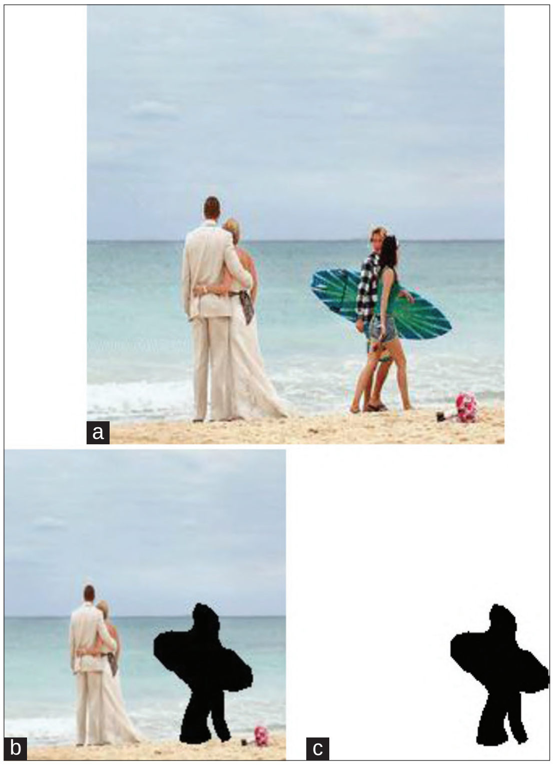


Fig. 1. (a) Original image, (b) mage with target region, (c) mask image.

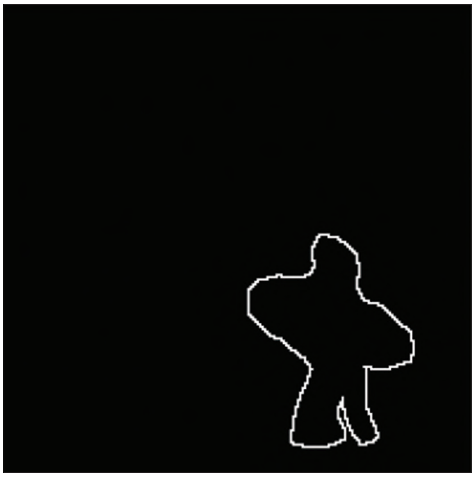


Fig. 2. Boundary Detection of the Target Region using Laplacian edge detection.

4.3. Patch Priority Function

It is useful to decide which patch on the boundary of the target region must be inpainted first. After patch priority function is computed, the patch Ψp centered at the point p for all p є δΩ with the highest priority is taken, see Fig. 3. In this proposed approach, the same patch priority equation presented in Hsieh *et al.* [16] is performed.

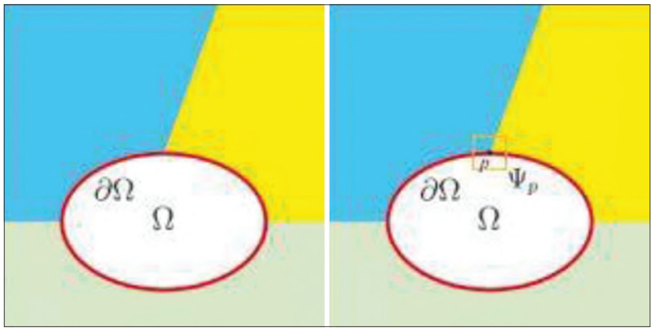


Fig. 3. Select Ψp from δΩ with the highest priority.

4.4. Searching for the Most Similar Patch

Once the highest priority patch Ψp is selected on the boundary of the target region δΩ in the previous step, the most similar patch to fill the unknown pixels of the Ψp needs to be found. For this sake, the proposed approach developed an improved searching mechanism in which the entire source region Φ of the image should be searched patch by patch from the upper left corner of the image. The source region Φ consists of the whole image without the target region (I-Ω). Now for each patch from the source region that is denoted by Ψq, two different distances between Ψp and Ψq_i need to be found. For the first one, the distance between the positions of each patch on the image is calculated.

$$X\_POS=|\Psi p_x - \Psi q_{xi}|$$

$$Y\_POS=|\Psi p_y - \Psi q_{yi}|$$

$$DISTANCE\_POS=\frac{X\_POS+Y\_POS}{img\_size/\lambda} \tag{4}$$

ʎ is a positive number of uses to make a balance between the position's distance and Euclidean distance. In this proposed approach ʎ is set to 10, and X_POS calculates a position distance between Ψp and Ψq on the X-axis and Y_POS for Y-axis.

For the second one, the Euclidean distance between these two patches is calculated as follows:

$$\hat{R}=(R_{\Psi p}-R_{\Psi qi})^2$$

$$\hat{G}=(G_{\Psi p}-G_{\Psi qi})^2$$

$$\hat{B}=\left(B_{\Psi p} - B_{\psi qi}\right)^2$$

$$ED=\sqrt{\left(\hat{R}+\hat{G}+\hat{B}\right)/3} \tag{5}$$

Where, R, G, and B represent the red, green, and blue channels of the image, respectively.

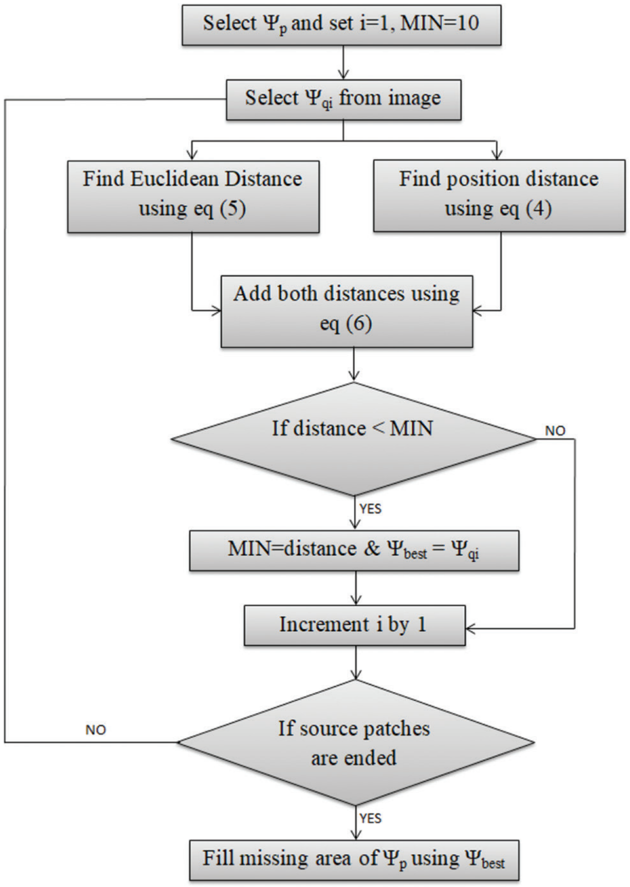

Select Ψ_p and set i=1, MIN=10
Select Ψ_{qi} from image
Find Euclidean Distance using eq (5)
Find position distance using eq (4)
Add both distances using eq (6)
If distance < MIN
NO
YES
MIN=distance & $\Psi_{best} = \Psi_{qi}$
Increment i by 1
NO
If source patches are ended
YES
Fill missing area of Ψ_p using Ψ_{best}

Fig. 4. Block diagram of the proposed searching mechanism.

Finally, we perform the summation between these two distances value to obtain the appropriate distance value.

$$\text{Distance=DISTANCE\_POS+ED} \qquad (6)$$

The above equations should repeat between Ψp and the remaining patches in the source region Ψq_i. In the end, the minimum distance value is chosen as the best similar patch. This process continues until the entire target region is filled.

The following diagram illustrates the proposed searching mechanism.

Fig. 4 is illustrated our proposed searching mechanism to fill the missing pixels in the highest priority patch Ψp that is selected in section 4.3. Consequently, the algorithm searches the entire image patch by patch. These patches are denoted by Ψq_i where *i* is a number of patches in the image that starts from one to the last patch number. After performing

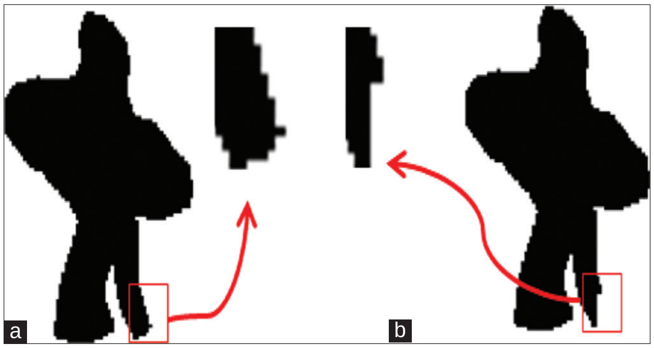


Fig. 5. (a) Mask image before the confidence value is updated, (b) mask image after the confidence value is updated at the first iteration.

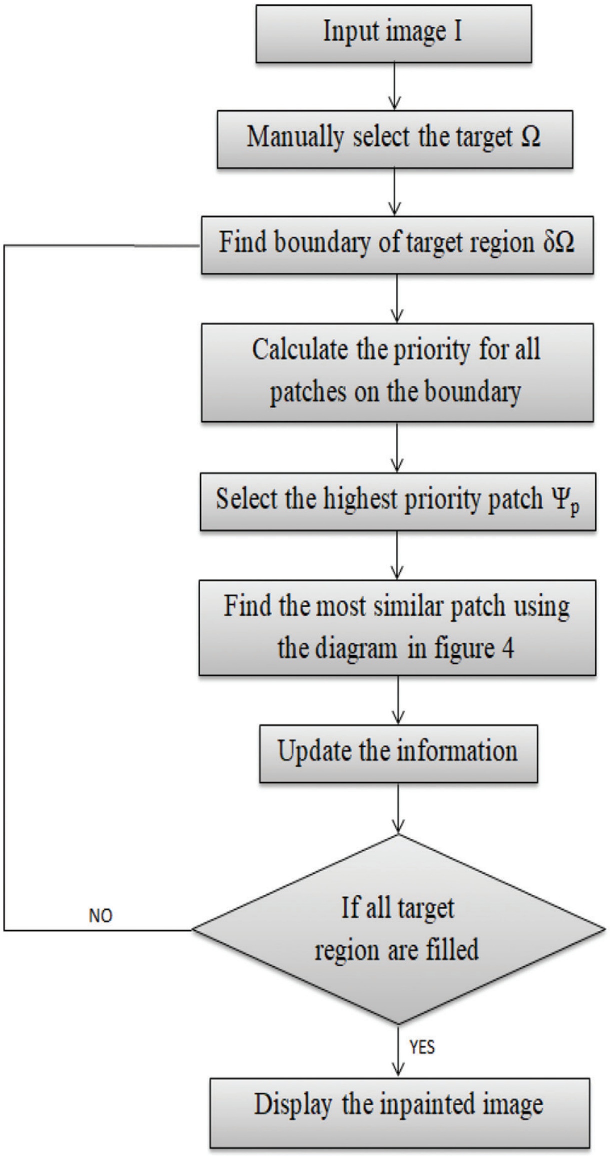

Input image I
Manually select the target Ω
Find boundary of target region δΩ
Calculate the priority for all patches on the boundary
Select the highest priority patch Ψ_p
Find the most similar patch using the diagram in figure 4
Update the information
NO
If all target region are filled
YES
Display the inpainted image

Fig. 6. Block diagram of the proposed approach.

equations of 4, 5, and 6, the result of equation 6 needs to be checked whether it is smaller than the MIN value or not (MIN is a variable used to take the smaller distance value in each iteration that is initially set to 10) if it is smaller than the MIN variable, it means that the current patch (Ψq_i) is the most similar to Ψp than the previous patch (Ψq_{i-1})? This process continues until the last *i*. Finally, we set Ψ_{best} to the most similar patch and replace the missing pixels of Ψp using Ψ_{best}.

4.5. Updating Confidence Value

After the patch Ψp has been filled as explained in the previous step, the region of Ψp in the confidence C(p) on the mask image is updated as follow:

C(p)=C(p') where C(p') is the region of the Ψp after it is filled.

Fig. 5 shows updating confidence value at the first iteration.

Fig. 6 shows the block diagram of the whole steps of the proposed approach.

5. RESULTS AND COMPARISON

To evaluate the performance of the proposed approach, experiments are conducted extensively using different natural and complex images of different file formats such as. JPG and.PNG. These images are common images that are usually used in image inpainting research area and they are from a dataset [22]. Fig. 7 shows some images which are used in the experiments: The second column includes original images and the third column contains images which marked the target region in black color.

The proposed approach is compared with algorithms in Perez *et al.* [14] and Huang *et al.* [18]. Two main kinds of quality assessment techniques are performed (objective and subjective). Fig. 8 illustrates the results of reconstructing the target region of the images in Fig. 7 using each of the proposed approaches as well as approaches in Perez *et al.* [14] and Huang *et al.* [18]. The reconstructed of the missing region is highlighted in a red circle.

In Fig. 8, some of them resulted images cannot be seen subjectively, i.e., visually. The peak signal-to-noise ratio (PSNR) is used as a quality assessment objectively to measure the quality of the reconstructed images. Results demonstrate that the proposed approach obtained the highest PSNR value compared to other tested approaches, Table 1.

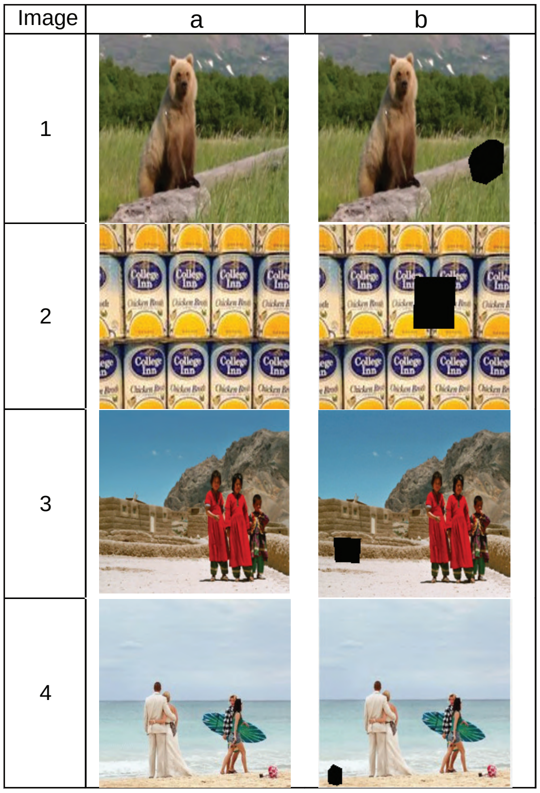


Fig. 7. Example of tested images, (a) original images, (b) original images with target region to be inpainted.

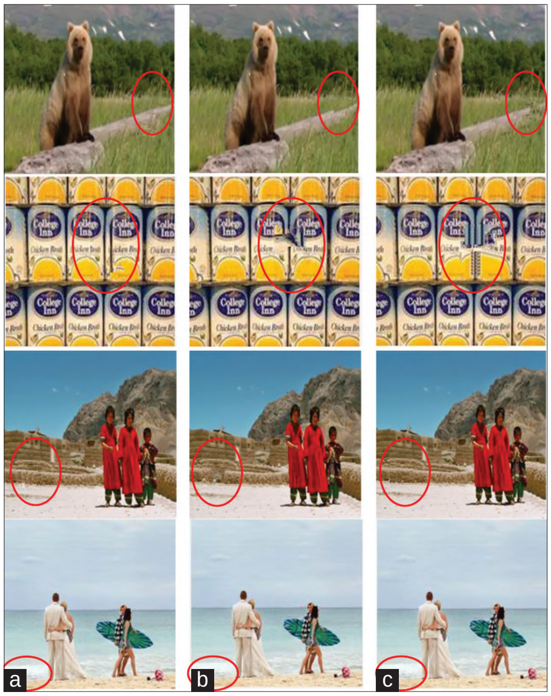


Fig. 8. Results of (a) proposed approach, (b) approach in Perez *et al*. [14], (c) approach in Huang *et al*. [18].

TABLE 1: Peak signal-to-noise ratio values

Tested images	Proposed	[14]	[18]
Image 1	36.5061	35.8802	36.0454
Image 2	27.6487	25.4769	23.7248
Image 3	35.4665	34.8515	34.0345
Image 4	47.2478	45.0843	46.7304

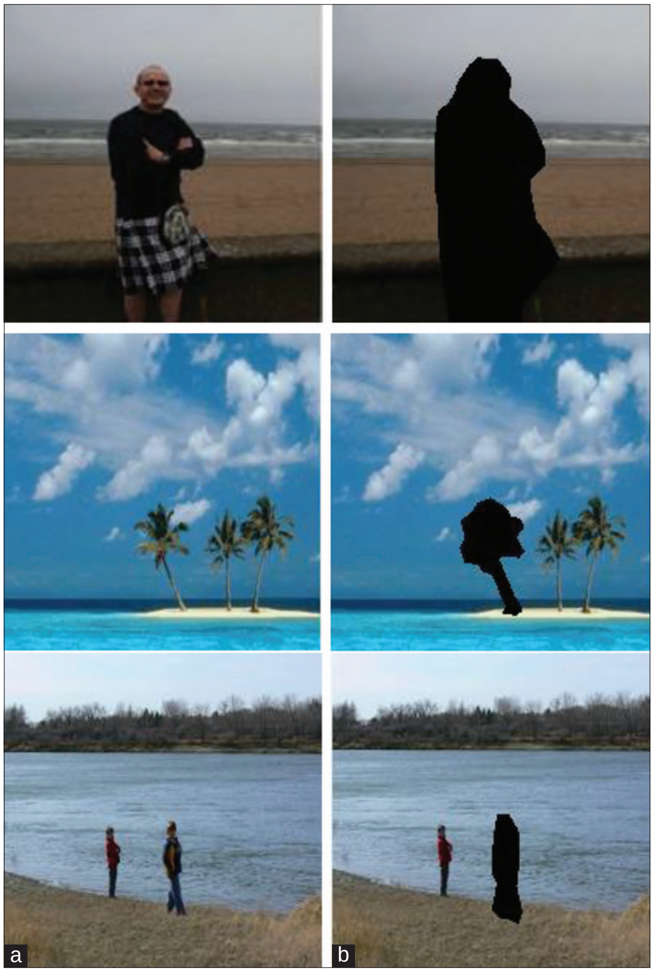


Fig. 9. (a) Original images (b) selecting unwanted objects.

Based on the results present in Table 1, one can see that sometimes the result of Perez *et al.* [14] is better than the result of Huang *et al.* [18] and vice versa, while for the proposed approach, for all tested images, the results are better than both Perez *et al.* [14] and Huang *et al.* [18]. This is because each of the tested algorithms has limitations for filling different regions such as shapes, edges, textures, and different backgrounds, while the proposed approach exceeds this limitation.

The previous results show how to fill the missing area and how to evaluate the result objectively using PSNR. Now another test is conducted by removing an object from some tested images that are evaluated subjectively, i.e., visually. Fig. 9 shows some images with selecting unwanted objects to be removed by black color.

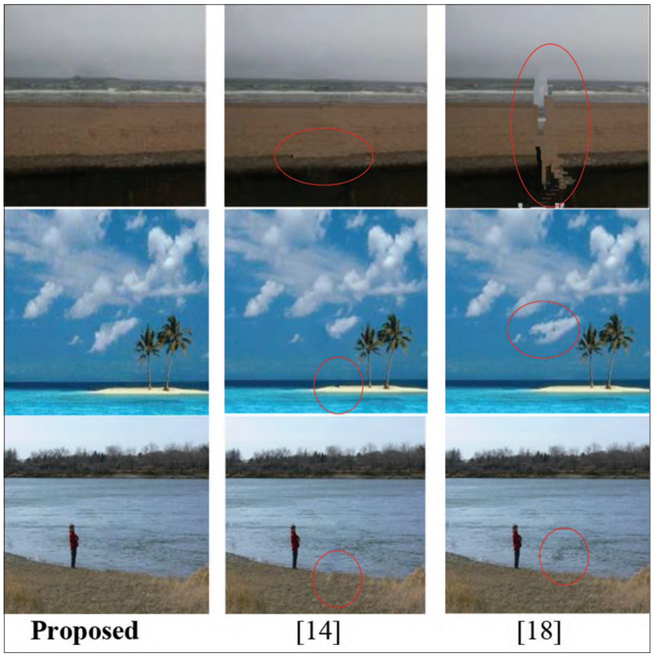


Fig. 10. Reconstructed images.

Fig. 10 presents the results of the reconstructed images in Fig. 9(b) after filling the region of the removed objects that are marked by black color. The limitation of the reconstructed images using approaches in Perez *et al.* [14] and Huang *et al.* [18] is highlighted in a red circle. Fortunately, the proposed approach does not contain such limitation, Fig. 10.

6. CONCLUSION

This paper is concerned with developing a new exemplar-based image inpainting algorithm by improving the searching mechanism to find the best similar patch. The proposed algorithm is performed two steps of distances to select the best-matched patch. The first step finds the distance between the position of the selected patches and the patch which needs to be filled. On the other hand, the second step calculates the Euclidean distance between them. Finally, the summation of these two distances is calculated to find the most similar patch. Experimental results demonstrate that the proposed approach is achieved a higher quality in terms of both objectives (i.e. PSNR) and subjective (i.e., visually) compared to other existing approaches.

REFERENCES

[1] H. Sellahewa, S. A. Jassim and A. A. Abdulla. “*Stego Quality Enhancement by Message Size Reduction and Fibonacci Bit-plane Mapping*”. United Kingdom, London, pp. 151-166, 2014.

[2] A. A. Abdulla. “*Exploiting Similarities between Secret and Cover Images for Improved Embedding Efficiency and Security in Digital Steganography*,” Department of Applied Computing, University of Buckingham, PhD Thesis, 2015.

[3] C.Guillemot and O. Meur. “Image Inpainting: Overview and Recent advances”. *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 127-144, 2014.

[4] L. Cai and T.Kim. “Context-driven hybrid image inpainting”. *IET Image Processing*, vol. 9, no. 10, pp. 866-873, 2015.

[5] B. Nizar, H. A. Ben and M. Ali. “Automatic inpainting scheme for video text detection and removal. *IEEE Transactions on Image Processing*, vol. 22, pp. 4460-4472, 2013.

[6] J. K. Chhabra and V. Birchha. “An enhanced technique for exemplar based image inpainting”. *International Journal of Computer Applications*, vol. 115, pp. 20-25, 2015.

[7] R. H. Park and Y. Seunghwan. Red-eye detection and correction using inpainting in digital photographs”. *IEEE Transactions on Consumer Electronics*, vol. 55, pp. 1006-1014, 2009.

[8] M. S. Kankanhalli and W. Q. Yan. “*Erasing Video Logos Based on Image Inpainting*”. Vol. 2. IEEE, Lausanne, Switzerland, pp. 521-524, 2002.

[9] Wu, Y., K. Zhonglin and Z. Hongying. "*An Efficient Scratches Detection and Inpainting Algorithm for old Film Restoration*". Vol. 1. IEEE, Kiev, Ukraine, pp. 75-78, 2009.

[10] Y. Mecky, G. Sergios, Y. Bin and A. Karim. “*Adversarial Inpainting of Medical Image Modalities*”. IEEE, Brighton, United Kingdom, pp. 3267-3271, 2019.

[11] M. B. Vaidya and K. Mahajan. “Image in painting techniques: A survey”. *IOSR Journal of Computer Engineering*, vol. 5, no. 4, pp. 45-49, 2012.

[12] Jain, L., A. G. Patel and K. R. Pate. "Image inpainting-a review of the underlying different algorithms and comparative study of the inpainting techniques". *International Journal of Computer Applications*, vol. 118, no. 10, 2015. Available from: http://www.citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.695.9341&rep=rep1&type=pdf.

[13] B. Limbasiya and N. Pandya. “A survey on image inpainting techniques”. *International Journal of Current Engineering and Technology*, vol. 3, no. 5, pp. 1828-1831, 2013.

[14] P. Perez, K. Toyama and A. Criminisi. “*Object Removal by Exemplar-based Inpainting*”. IEEE, Madison, WI, USA, 2003.

[15] P. Perez, K. Toyama and A. Criminisi. “Region filling and object removal by exemplar-based image inpainting”. *IEEE Transactions of Image Processing*, vol. 13, no. 9, pp. 1200-12012, 2004.

[16] C. W. Hsieh, S. K. Lin, C. W. Wang and J. L. Wu, W. H. Cheng. “Robust Algorithm for Exemplar-based Image Inpainting”. *Proceedings of International Conference Computer Graphics, Imaging and Vision*, pp. 64-69, 2005.

[17] A. Wong and J. Orchard. “*A Nonlocal-means Approach to Exemplar-based Inpainting*”. IEEE, San Diego, CA, USA, pp. 2600-2603, 2008.

[18] T. Huang, X. Zhao and L. Deng. “Exemplar-based image inpainting using a modified priority definition”. *Neurocomputing*, vol. 10, no. 10, pp. 1-18, 2015.

[19] H. Liu, G. Lu, X. Wang, J. Wei, Y. Chao and X. Bi. “*Exemplar-based Inpainting under Boundary Contraction Constraints*”. IEEE, Shenyang, China, pp. 295-300, 2018.

[20] A. Awati, B. Pandurngi, M. R. Patil and H. C. Rao. “*Image Inpainting using Exemplar Based Technique with Improvised Data Term*”. IEEE, Belgaum, India, pp. 162-166, 2018.

[21] S. Wang and Y. Xu. “*Image Inpainting Based on Color Differences and Structure Differences*”. IEEE, Dalian, China, pp. 364-368, 2013.

[22] Available from: http://www.escience.cn/people/dengliangjian/Data.html. [Last accessed on 2019 Dec 12].

ORIGINAL ARTICLE
UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Thresholding-based White Blood Cells Segmentation from Microscopic Blood Images

Zhana Fidakar Mohammed[1], **Alan Anwer Abdulla**[2,3]

[1]Department of Information Technology, Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, Iraq, [2]Department of Information Technology, College of Commerce, University of Sulaimani, Sulaimani, Iraq, [3]Department of Information Technology, University College of Goizha, Sulaimani, Iraq

ABSTRACT

Digital image processing has a significant role in different research areas, including medical image processing, object detection, biometrics, information hiding, and image compression. Image segmentation, which is one of the most important steps in processing medical image, makes the objects inside images more meaningful. For example, from microscopic images, blood cancer can be identified which is known as leukemia; for this purpose at first, the white blood cells (WBCs) need to be segmented. This paper focuses on developing a segmentation technique for segmenting WBCs from microscopic blood images based on thresholding segmentation technique and it compares with the most commonly used segmentation technique which is known as color-k-means clustering. The comparison is done based on three well-known measurements, used for evaluating segmentation techniques which are probability random index, variance of information, and global consistency error. Experimental results demonstrate that the proposed thresholding-based segmentation technique provides better results compared to color-k-means clustering technique for segmenting WBCs as well as the time consumption of the proposed technique is less than the color-k-means which are 70.8144 ms and 204.7188 ms, respectively.

Index Terms: Medical image processing, Segmentation techniques, Thresholding, White blood cells

1. INTRODUCTION

The image processing techniques have been developed in the various of areas such as pattern recognition, biometrics, image inpainting, medical image processing, image compression, information hiding [1], and multimedia security [2]. Medical image processing is a collection of techniques that help the clinician in the diagnosis of different diseases from medical images such as X-ray, magnetic resonance imaging, computed tomography, and ultrasound and microscopic images. Thus, various types of cancer can be detected based on medical image processing from medical images such as breast cancer, brain tumor, lung cancer, skin cancer, and blood cancer (leukemia). The advantage of creating a system based on medical image processing techniques is extracting the targeted diseases in higher accuracy, with reducing time consumption as well as decreasing cost, otherwise, the manual processing is taken a lot of time and it also needs a professional staff for detecting of diseases [3]. In general, processing medical images include four main steps which are: Pre-processing, segmentation, feature extraction, and classification. This paper is mainly focused on the segmentation step in processing microscopic blood images which help the clinician in identifying various diseases in human's blood such as blood cancer (leukemia), and anemia. Segmentation of white blood cells (WBCs) is the most important step in identifying leukemia. Leukemia is a type of cancer that affects blood, lymphocyte system, and bone marrow. Thus, the correct segmentation of WBCs has an impact on further steps, such as feature extraction and classification, to

Access this article online

DOI: 10.21928/uhdjst.v4n1y2020.pp9-17	E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Alan Anwer Abdulla, Department of Information Technology, College of Commerce, University of Sulaimani and University college of Goizha, Sulaimani, Iraq. E-mail: Alan.abdulla@univsul.edu.iq

Received: 12-01-2020 Accepted: 09-02-2020 Published: 13-02-2020

obtain results more accurately [3]-[5]. In general, the main components of blood are four types as follows:

- Red blood cells (RBCs): Are responsible for delivering oxygen from the lungs to all the parts of the human's body [6]
- WBCs: Are responsible for body's immune system. The WBCs are larger in size and fewer in number compared to RBCs [6]
- Platelets: Are responsible for blood clotting [6]
- Plasma: Is responsible for carrying nutrients that are required by the cells [3]. It makes up about half of the blood volume.

Most blood cells are produced in the bone marrow, and there are cells known as stem cells (lymphoid and myeloid) that are responsible for producing different blood cells [7] in general, there are different types of WBCs which are Basophil, eosinophil, neutrophil, lymphocyte, and monocyte, as illustrated in Fig. 1.

In this paper, a segmentation technique based on thresholding is proposed to segment WBCs from the microscopic blood images using public and well-known dataset acute lymphoblastic leukemia-image database (ALL-IDB). The proposed technique includes the following steps: Pre-processing step to enhance the image quality, segmentation step to separate WBCs from other blood components, and image cleaning step to remove the unwanted objects inside the segmented image. Consequently, the proposed segmentation technique is compared with other technique, which is known as color-k-means clustering, in terms of time consumption as well as performance of the segmentation technique based on probability random index (PRI), variance of information (VOI), and global consistency error (GCE) which are measurements that using for evaluating a segmentation techniques.

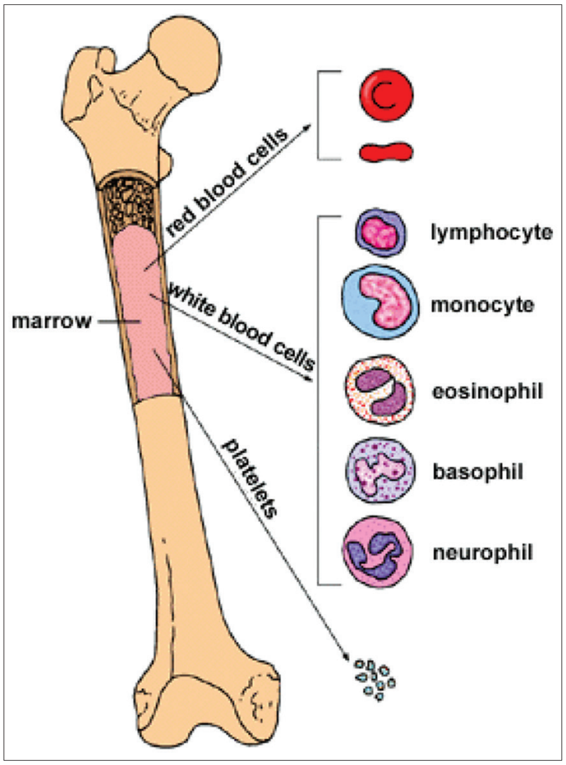


Fig. 1. Bone marrow and blood components.

1.1. Problem Statement

The manual processing of medical images is time consuming and it requires a professional staff which mainly depends on personal skills, and sometimes it may produce inaccurate results.

1.2. Objective of the Research

The main objective of the research is to segment WBCs from microscopic blood images accurately and quickly because accurate segmentation of WBCs makes other processes (such as feature extraction) easer consequently producing an accurate result of classification. Thus, the overall process could help the clinician in identifying different diseases accurately in a faster way, which further helps them to provide treatment for the patients sooner.

The rest of the paper is organized as follows: In Section 2, the literature review is presented. Section 3 presents the proposed approach. The experimental results are illustrated in Section 4. Finally, our conclusions are given in Section 5.

2. LITERATURE REVIEW

Nowadays, processing medical images have a crucial role for early identification of diseases. Thus, segmenting WBCs from microscopic blood images are the most important step in identifying leukemia. This section focuses on reviewing the most important existing works on segmenting WBCs.

In 2009, Sadeghian *et al.* [8] proposed a segmentation technique to segment WBCs as well as their nuclei and a segmentation technique to separate the cytoplasm of the cell. The red, green, and blue (RGB) image is converted into gray image; then canny edge detection is applied followed by gradient vector flow to connect the boundary of the nucleus. Consequently, the hole filling technique is applied to get the nucleus. Furthermore, Zack algorithm is applied into the gray image to get the binary image to extract the cytoplasm of the cell by subtracting the binary image from the gray image. Finally, their proposed work is obtained 92% accuracy for nucleus segmentation and 70% accuracy for cytoplasm segmentation.

In 2015, Marzukia *et al.* [9] proposed a system to segment the nuclei of WBCs based on the active contour technique. The RGB image is converted to a gray image and then the active contour is applied to the resulted image to get the segmented nucleus. Continuously, the obtained image is converted to a binary image to find the roundness value to determine the grouped and individual WBCs. As they claimed, their proposed system can accurately extract the boundary of WBC nuclei.

Continuously, in 2015 Madhloom *et al.* [10] proposed an approach to segment WBCs and their nucleus. First, it segments the WBCs based on the color transformation as well as mathematical morphology. Moreover, the marker-control watershed technique is applied to separate overlapped cells. Furthermore, the seeded region growing technique is used to segment the nucleus of the cells. Finally, the performance of their approach is evaluated using relative ultimate measurement accuracy and misclassification error to measure the accuracy and it achieves 96% for WBCs segmentation and 94% for nucleus segmentation.

In 2016, Sobhy *et al.* [11] proposed two segmentation techniques for segmenting WBCs. First, color correction is applied to extract the mean intensity from the histogram of each RGB channel of the original image. Moreover, the image is converted to hue saturation intensity color space, and then the two tested segmentation techniques are applied to the S component of the image. The first technique was Otsu's thresholding and the second technique was marker-control watershed segmentation. Furthermore, the exoskeleton algorithm is used to separate the adjusted WBCs. Finally, they compared their work with another study which is manually counted the WBCs based on 30 images. As they claimed, their proposed is achieved an accuracy of 93.3% for Otsu's segmentation and 99.3% for marker-control watershed segmentation.

In 2017, Gowda and Kumar [12] proposed a system to segment WBCs based on k-means clustering and Gram-Schmidt orthogonalization. The proposed system is started by pre-processing step which includes: (1) Converting image from RGB to gray image, (2) median filter is implemented to remove noise, and (3) image normalization is applied for contrast stretching. Consequently, k-means clustering segmentation technique is applied to segment WBC and its subtypes such as: Basophil, eosinophil, neutrophil, lymphocyte, and monocyte. Furthermore, Gram-Schmidt orthogonalization scheme is applied to segment the nucleus from the cell.

Finally, in 2017, Nain *et al.* [13] proposed a system to differentiate individual and overlapped WBCs based on the watershed segmentation scheme to segment the cells. The edge map extraction technique, which is a circle fitted on each cell, is used to identify both individual and overlapped cells. To evaluate the performance of the proposed work, two evaluation measurements which are detection rate (DR) and false alarm rate (FAR), are used. As they claimed, the proposed system is achieved 97.10% and 7.80%, respectively, for DR and FAR, the lower value of FAR and higher value of DR means better segmentation of WBCs.

3. PROPOSED APPROACH

A microscopic image of the blood sample contains three components, as discussed before, which are RBCs, WBCs, and background (platelets and plasma). The proposed approach focuses on segmenting WBCs from other components (which are RBCs and background). The segmentation is done based on the thresholding technique. This section concentrates to explain the steps of the proposed work, as illustrated in Fig. 2.

3.1. Pre-processing

The first step of the proposed work is pre-processing, which is an important step to provide a better quality of the input image and make the next step (segmentation step) easier. In this step, the three channels of the input image, which are RGB are separated, and then the median filter (to remove noise) and histogram equalization (to contrast enhancement) are applied on the only green channel of the input image to enhance the image quality, Fig. 3.

The median filter is a non-liner filter, which is used for removing noise. While, the histogram equalization is a technique that uses image's histogram to improve low or high contrast of the image which makes better distribution of intensities of the image's histogram [14].

3.2. Segmentation

To segment the WBCs from the other components of the microscopic image, i.e., RBCs and background, the thresholding segmentation technique is applied to the image.

The thresholding-based segmentation technique is the simplest and useful segmentation technique that segments an image based on the intensity of the pixels value [15]. This technique is

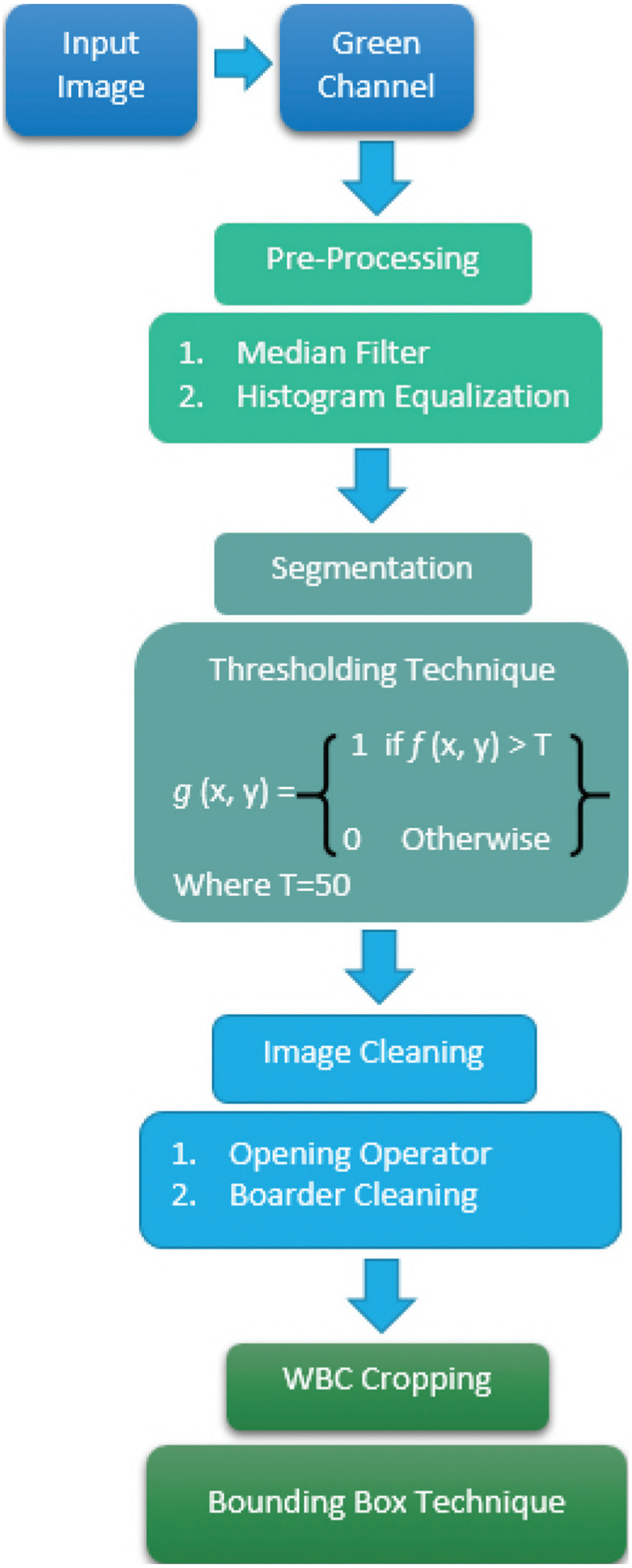


Fig. 2. Steps of the proposed approach.

produce a binary image from a gray image (the value of 0, which represents background and it is black and the value of one which represents object inside the image and it is white color). This process has the advantage of reducing complex information and simplifies the classification and recognition processes [16]. The threshold value of this technique can be selected manually or automatically based on the information from the features of the image [17]. There are types of thresholding which are:

3.2.1. Global thresholding (single thresholding)

This technique use only one threshold (T) to segment the whole image into objects and background; this technique is more appropriate for those images that have bimodal histogram. The global thresholding can be defined as equation (S1) [16], [17]: Suppose a sample image of $f(x, y)$ has the following diagram (Fig. 4):

Then, the binary image $g(x, y)$ of the $f(x, y)$ can be defined as following equation:

$$g(x, y) = 1 \text{ if } f(x, y) > T \text{ otherwise } 0 \text{ if } f(x, y) \leq T \quad (1)$$

Where T is the threshold value.

3.2.2. Local thresholding (multiple thresholding)

In this thresholding technique, the image segments based on multiple threshold values and the image partitions into multiple region of interests (ROIs) and backgrounds [15]. Moreover, it segments an image by partitioning an image into (n × n) sub-images and then selects a threshold *Tij* for every sub-image [16], as illustrated in Fig. 5.

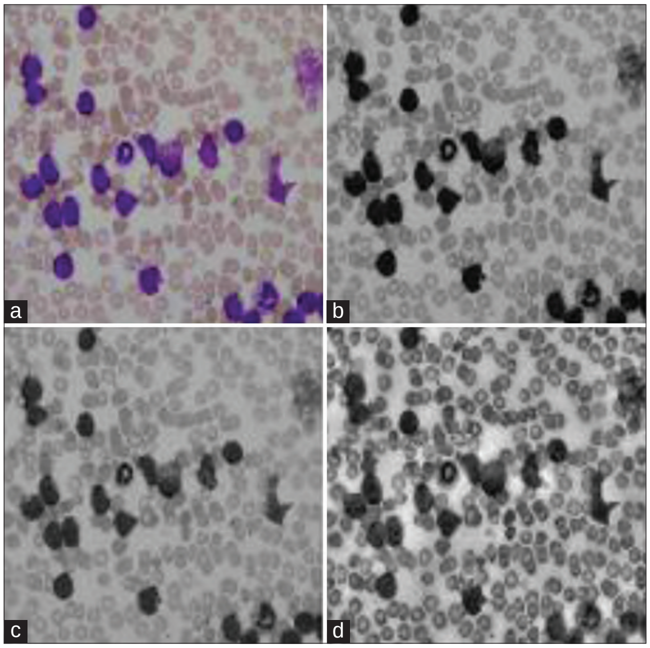


Fig. 3. Pre-processing: (a) Input image, (b) green channel, (c) resulted in image after the median filter is applied, (d) resulted image after histogram equalization is applied.

This technique is more appropriate for images that are contained disparate illuminations, this technique can be defined as follows [16]:

$$g(x, y) = 0 \text{ if } f(x, y) < T(x, y) \text{ otherwise}$$
$$1 \text{ if } f(x, y) \geq T(x, y) \quad (2)$$

Where $f(x, y)$ is an input image and $g(x, y)$ is a binary image produces depending on multiple threshold value T (x, y).

In our proposed approach, we segment the microscopic images into ROI (which are WBCs) and background based on global threshold value with (T = 50) as clarified in the diagram of the work (Fig. 2.). We apply this technique in the resulted image obtained after the pre-processing step is done (i.e., image d from Fig. 3). Moreover, we obtain a binary image that contains WBCs which are larger in size in the image and some remaining as unwanted objects (RBCs and platelets), as illustrated in Fig. 6.

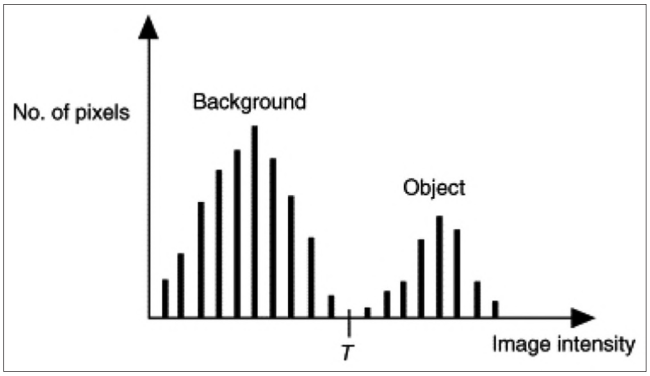


Fig. 4. Histograms of a sample image.

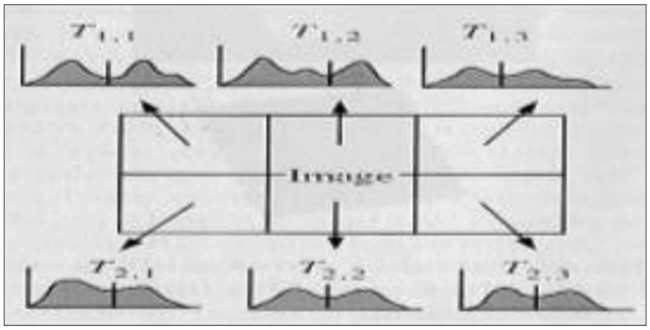


Fig. 5. Segmentation based on local thresholding.

3.3. Image Cleaning

The resulted image obtained in the previous step, as illustrated in Fig. 6, includes some unwanted objects such as RBCs or platelets. To overcome this drawback, the morphological opening operator is applied to the resulted image of the segmentation step.

The opening operator is a morphological operation that uses to eliminating imperfections in the images that impact on texture and shape of images. Thus, morphological operations have a crucial role in image processing especially image segmentation process because they treat with shape extraction within the image and they describe the structure of the image [18]. The most significant morphological operations are dilation and erosion. There are also two other operations which are opening and closing which work depending on the combination of dilation and erosion [18], [19]. The opening operator is a combination of both erosion and dilation. It first erodes an image depends on the structuring element and then dilated it by the same structural element. Opening smoothens the boundary of an object and deletes small unwanted objects inside the image. It can be defined as the following equation [18], [19]:

$$A \circ B = (A - B) + B \tag{3}$$

Where A: Is an image

B: Is a structuring element.

Moreover, some of the WBCs are located in the edge of the image; thus, the border cleaning scheme is used to remove those cells because they have a negative impact on the accuracy rate in further steps such as feature extraction and classification. Thus, to make only complete WBCs remains in the image, the border cleaning is applied to the resulted image of the opening operator, as illustrated in Fig. 7. The border cleaning is a morphological technique that used to remove an object that touches the edge of the image [11].

3.4. WBCs Cropping

To crop each WBCs as an individual image, we used the bounding box technique. This technique can be defined as the smallest rectangle that soured an object of interest, and it extracts the minimum area of the box and it can be represented as follows [20]:

$$\text{Area} = \text{Major axis length} \times \text{Minor axis length} \tag{4}$$

Where the major axis represents the longest line that can be drawn between two points in the object, and the minor axis is the longest line between two points of the object that can be drawn perpendicularly to the major axis, as illustrated in Fig. 8. [20].

We applied the bounding box technique on the original image (input image) based on the location of WBCs in segmented and cleaned images obtained in Fig. 7, which further can be used for more processing such as feature extraction and classification.

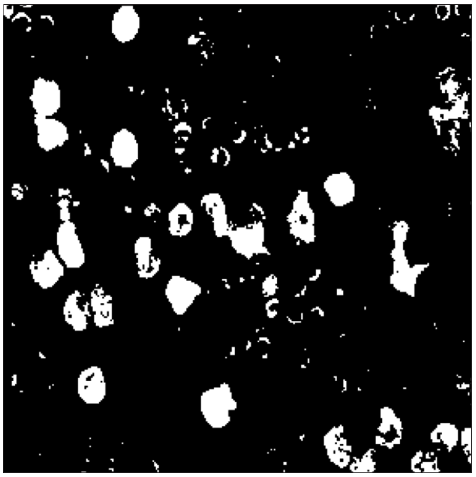


Fig. 6. The resulted image after thresholding segmentation is applied.

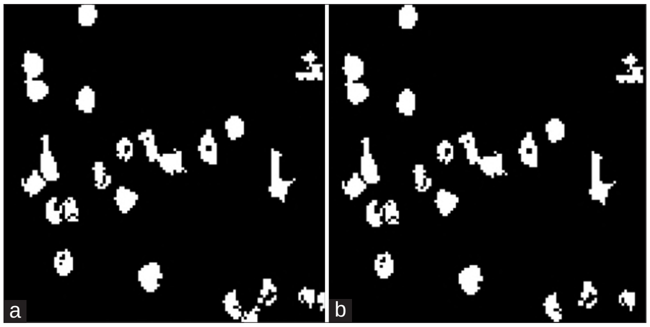


Fig. 7. Image cleaning: (a) Resulted image after opening operator is applied, (b) resulted image after border cleaning is applied.

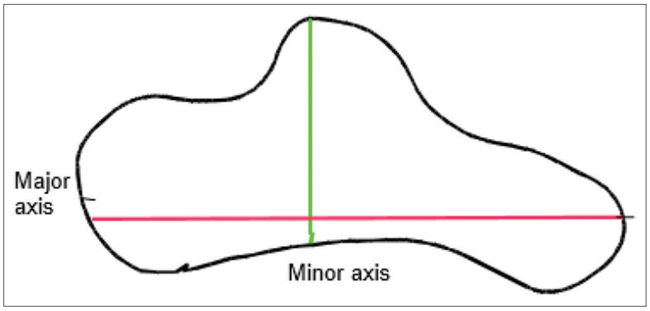


Fig. 8. The major axis and the minor axis.

The result of the bounding box is represented in Fig. 9. And Fig. 10 illustrate cropped WBCs as individual images.

4. EXPERIMENTAL RESULTS

The main purpose of the proposed approach is to segment the WBCs from microscopic blood images. In this section, to evaluate the performance of the proposed approach, experiments are conducted on the dataset (explained in 4.1).

4.1. Dataset

The input images are taken from a public and commonly used dataset known as ALL-IDB, which consists of two groups of images:

- The first one is ALL-IDB1 which was designed for testing the segmentation techniques, and it contains 108 microscopic blood images, including (49 abnormal images of leukemia patients and 59 normal images) [21]
- The second one is called ALL-IDB2 which was designed for testing the classifier techniques and it consists of 260 cropped WBC images (50% abnormal cells and 50% normal cells) which are taken from ALL-IDB1 [21].

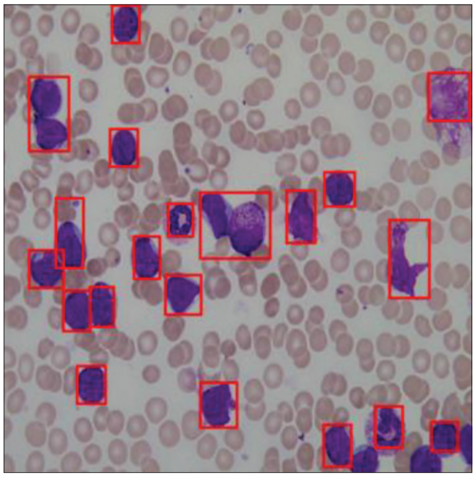


Fig. 9. Applying bounding box technique on the original image based on the segmented image, as presented in Fig. 5.

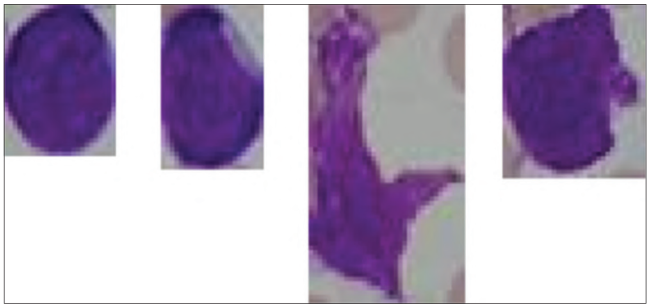


Fig. 10. Example of cropped white blood cells.

As this study is mainly focused on the segmentation of WBCs, thus the first group of the dataset is used (i.e., ALL-IDB1). Because as mentioned above, this group of the dataset is dedicated for testing segmentation techniques.

4.2. Results

As mentioned in Section 3, in the thresholding segmentation, only the green channel of the input microscopic image was used for the segmentation process. Because the green channel gives better results in terms of segmentation compared to the other two channels (red and blue), since it holds more contrast information regarding to WBC [22]. In the segmentation step, the thresholding technique used for segmenting WBCs with T = 50, in our experiments first different threshold values are tested on all three channels of normal and abnormal images and the experimental results demonstrate that the best result is given by the green channel with T = 50. As illustrated in Fig. 11. the WBCs have higher contrast in the green channel of the image compared to two other channels; thus, it makes the segmentation process more accurate.

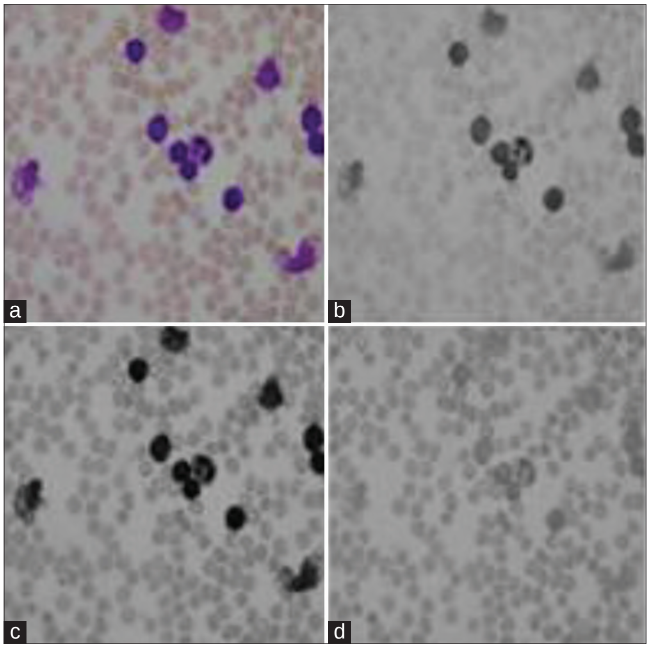


Fig. 11. A blood microscopic image of leukemia patient: (a) Original image, (b) red channel, (c) green channel, (d) blue channel.

To evaluate the performance of the proposed approach, we compare our proposed approach with the commonly used segmentation technique which is known as color-k-means clustering, which was used by many existing works such as Mishra *et al.*, Kumar and Vasuki, Ferdosi *et al.*, Sarrafzadeh and Dehnavi [23]-[26]. We also applied the color-k-means clustering on the same dataset (ALL-IDB1) to evaluate both segmentation technique, the comparison was done in terms of time consumption as well as the performance of the segmentation techniques based on the PRI, GCE, and VOI. The PRI is a nonparametric technique used to measure the performance of the segmentation technique and it can be calculated as follows [22], [27]:

$$\mathrm{PRI}\left(S_{test}, G_k\right) = \frac{1}{N/2} \Sigma_{\forall i,j\ \&\ i<j} \left[C_{i,j}P_{i,j} + \left(1 - C_{i,j}\right)\left(1 - P_{i,j}\right)\right] \tag{5}$$

Moreover, the GCE is a region-based segmentation consistency, which measures to quantify the consistency between image segmentation of differing granularities, and it can be calculated as follows [28]:

$$\mathrm{GCE}\left(S_1,\ S_2\right) = \frac{1}{n}\min\{\Sigma_i\ X_i(S_1,S_2), \Sigma_i\ X_i(S_1,S_2)\} \tag{6}$$

While the VOI is calculated the distance between two segmented areas which provide information about the participation of pixels in different clusters and measures the knowledge lost or gained in two clusters it can be calculated as follows [28]:

$$\mathrm{VOI} = H(X) = H(Y) – 2I(X,Y) \tag{7}$$

The experimental results are as follows:

4.2.1. Execution time

The time consumption has been calculated for both segmentation techniques which implemented on the ALL-IDB1, as illustrated in Table 1. The systems were implemented using MATLAB on Lenovo computer with 8 GB of RAM and 64-bit Operating System\Windows 10.

Table 1 demonstrates that the proposed thresholding-based technique needs less time compared to the color-k-means technique for segmenting the WBCs from microscopic images which applies to 49 abnormal images and 59 normal images.

TABLE 1: Time consumptions.

Segmentation techniques	Execution time\ms
Proposed	70.8144
Color-k-means clustering	204.7188

TABLE 2: Segmentation evaluation.

Measurements	Proposed technique	Color-k-means clustering technique
PRI	0.126347	0.515437
GCE	0.04317	0.159923
VOI	5.419551	5.164454

PRI: Probability random index, GCE: Global consistency error, VOI: Variance of information

4.2.2. Performance of the segmentation techniques

Three well-known segmentation measurements such as PRI, GCE, and VOI are calculated for both proposed technique and color-k-means clustering and Table 2 illustrates the performance of each technique. The lower value of PRI, GCE, and VOI means the better/higher performance of the segmentation technique [27].

Table 2 demonstrates that the proposed technique has a lower value of PRI, GCE, and VOI. Thus, it is better for segmenting WBCs from microscopic images.

5. CONCLUSION

The progression of using techniques of image processing in medical fields provides better accuracy for identifying different diseases from medical images. This paper focuses on the segmentation step, which is the most important step in medical image processing for segmenting WBCs from microscopic blood images, depending on the threshold-based technique. Experimental results prove that the proposed segmentation technique can extract the WBCs from images better than color-k-means clustering in terms of segmentation evaluation as well as time consuming.

REFERENCES

[1] H. Sellahewa, S.A. Jassim and A.A. Abdulla. Stego Quality Enhancement by Message Size Reduction and Fibonacci Bitplane Mapping. United Kingdom, London, 2014, pp. 151-166.

[2] A.A. Abdulla. Exploiting Similarities between Secret and Cover Images for Improved Embedding Efficiency and Security in Digital Steganography, Department of Applied Computing, University of Buckingham, PhD Thesis, 2015. Available from: http://www.bear.buckingham.ac.uk/149.

[3] H. B. Kekre, B. Archana and H. R. Galiyal. “Segmentation of blast

using vector quantization technique". *International Journal of Computer Applications*, vol. 72, pp. 20-23, 2013.

[4] M. A. Bennet, G. Diana, U. Pooja and N. Ramya. "Texture metric driven acute lymphoid leukemia classification using artificial neural networks". *International Journal of Recent Technology and Engineering*, vol. 7, no. 6S3, pp. 152-156, 2019.

[5] K. A. ElDahshan, M. I. Youssef, E. H. Masameer and M. A. Mustafa. "An efficient implementation of acute lymphoblastic leukemia images segmwntation on FPGA". *Advances in Image and Vedio Prpcessing*, vol. 3, no. 3, pp. 8-17, 2015.

[6] V. Venmathi, K. N. Shobana, A. Kumar and D. G. Waran. "Leukemia detection using image processing". *International Journal for Scientific Research and Development*, vol. 5, no. 1, pp. 804-808, 2017.

[7] S. C. Neoh, W. Srisukkham, L. Zhang, S. Todryk, B. Greystoke, C. P. Lim, M. A. Hossain and N. Aslam. "An intelligent decision support system for leukaemia diagnosis using microscopic blood images". *Scientific Reports*, vol. 5, p. 14938, 2015.

[8] F. Sadeghian, Z. Seman, A. R. Ramli, B. H. A. Kahar and M. Saripan. "A framework for white blood cell segmentation in microscopic blood images using digital image processing". *Biological Procedures Online*, vol. 11, pp. 196-206, 2009.

[9] N. I. C. Marzukia, N. H. Mahmoodb and M. A. A. Razakb. "Segmentation of white blood cell nucleus using active contour". *Jurnal Teknologi*, vol. 74, pp. 115-118, 2015.

[10] H. T. Madhloom, S. A. Kareem and H. Ariffin. "Computer-aided acute leukemia blast cells segmentation in peripheral blood images". *Journal of Vibroengineering*, vol. 17, pp. 4517-4532, 2015.

[11] N. M. Sobhy, N. M. Salem and M. El Dosoky. "A comparative study of white blood cells segmentation using otsu threshold and watershed transformation". *Journal of Biomedical Engineering and Medical Imaging*, vol. 3, no. 3, pp. 15-24, 2016.

[12] J. P. Gowda and S. C. P. Kumar. "Segmentation of white blood cell using K-means and gram-schmidt orthogonalization". *Indian Journal of Science and Technology*, vol. 10, pp. 1-6, 2017.

[13] K. N. Sukhia, M. M. Riaz, A. Ghafoor and N. Iltaf. "Overlapping white blood cells detection based on watershed transform and circle fitting". *Radioengineering*, vol. 24, pp. 1177-1181, 2017.

[14] S. Shafique and S. Tehsin. "Computer-aided diagnosis of acute lymphoblastic leukaemia". *Computational and Mathematical Methodsin Medicine*, vol. 2018, p. 6125289, 2018.

[15] S. Yuheng and Y. Hao. "Image segmentation algorithms overview". *arXiv Preprint*, vol. 2017, pp. 1-7, 2017.

[16] K. Bhargavi and S. Jyothi. "A survey on threshold based segmentation technique in image processing". *International Journal of Innovative Research and Development*, vol. 3, pp. 234-239, 2014.

[17] D. Kaur and Y. Kaur. "Various image segmentation techniques: A review". *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 5, pp. 809-814, 2014.

[18] S. Ravi and A. M. Khan. "Morphological Operations for Image Processing: Understanding and its Applications". In: *NCVSComs-13 Conference Proceedings*, 2013.

[19] S. Singh and S. K. Grewal. "Role of mathematical morphology in digital image processing: A review". *International Journal of Scientific Engineering and Research*, vol. 2, no. 4, 2014.

[20] A. E. Huque. "*Shape Analysis and Measurement for the HeLa Cell Classification of Cultured Cells in High Throughput Screening*". University of Skövde, Skövde, Sweden, 2006.

[21] R. D. Labati, V. Piuri and F. Scotti. "ALL-IDB: The Acute Lymphoblastic Leukemia Image Database for Image Processing". In: *18th IEEE International Conference on Image Processing*, 2011.

[22] S. Kumar, S. Mishra, P. Asthana and Pragya. "Automated Detection of Acute Leukemia Using k-mean Clustering Algorithm". In: *Advances in Computer and Computational Sciences*, Proceedings of ICCCCS, pp. 655-671, 2018.

[23] S. Mishra, L. Sharma, B. Majhi and P. Kumar Sa. "Microscopic Image Classification Using DCT for the Detection of Acute Lymphoblastic Leukemia (ALL)". *Proceedings of International Conference on Computer Vision and Image Processing*, pp. 171-180, 2017.

[24] P. S. Kumar and S. Vasuki. "Automated diagnosis of acute lymphocytic leukemia and acute myeloid leukemia using multi-SV". *Journal of Biomedical Imaging and Bioengineering*, vol. 1, pp. 20-24, 2017.

[25] B. J. Ferdosi, S. Nowshin, F. A. Sabera and Habiba. "White Blood Cell Detection and Segmentation from Fluorescent Images with an Improved Algorithm using K-means Clustering and Morphological Operators". In: *4th International Conference on Electrical Engineering and Information and Communication Technology* (*iCEEiCT*), 2018.

[26] O. Sarrafzadeh and A. M. Dehnavi. "Nucleus and cytoplasm segmentation in microscopic images using K-means clustering and region growing". *Advanced Biomedical Research*, vol. 4, p. 174. 2015.

[27] R. Kumar and A. M. Arthanariee. "Performance evaluation and comparative analysis of proposed image segmentation algorithm". *Indian Journal of Science and Technology*, vol. 7, pp. 39-47, 2014.

[28] R. Sardana. "Comparitive analysis of image segmentation techniques". *International Journal of Advanced Research in Computer Engineering and Technology*, vol. 2, no. 9, pp. 2615-2619, 2013.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Big Data Sentimental Analysis Using Document to Vector and Optimized Support Vector Machine

Sozan Abdulla Mahmood, Qani Qabil Qasim

Department of Computer Science, University of Sulaimani, Sulaymaniyah, Iraq

ABSTRACT

With the rapid evolution of the internet, using social media networks such as Twitter, Facebook, and Tumblr, is becoming so common that they have made a great impact on every aspect of human life. Twitter is one of the most popular micro-blogging social media that allow people to share their emotions in short text about variety of topics such as company's products, people, politics, and services. Analyzing sentiment could be possible as emotions and reviews on different topics are shared every second, which makes social media to become a useful source of information in different fields such as business, politics, applications, and services. Twitter Application Programming Interface (Twitter-API), which is an interface between developers and Twitter, allows them to search for tweets based on the desired keyword using some secret keys and tokens. In this work, Twitter-API used to download the most recent tweets about four keywords, namely, (Trump, Bitcoin, IoT, and Toyota) with a different number of tweets. "Vader" that is a lexicon rule-based method used to categorize downloaded tweets into "Positive" and "Negative" based on their polarity, then the tweets were protected in Mongo database for the next processes. After pre-processing, the hold-out technique was used to split each dataset to 80% as "training-set" and rest 20% "testing-set." After that, a deep learning-based Document to Vector model was used for feature extraction. To perform the classification task, Radial Bias Function kernel-based support vector machine (SVM) has been used. The accuracy of (RBF-SVM) mainly depends on the value of hyperplane "Soft Margin" penalty "C" and γ "gamma" parameters. The main goal of this work is to select best values for those parameters in order to improve the accuracy of RBF-SVM classifier. The objective of this study is to show the impacts of using four meta-heuristic optimizer algorithms, namely, particle swarm optimizer (PSO), modified PSO (MPSO), grey wolf optimizer (GWO), and hybrid of PSO-GWO in improving SVM classification accuracy by selecting the best values for those parameters. To the best of our knowledge, hybrid PSO-GWO has never been used in SVM optimization. The results show that these optimizers have a significant impact on increasing SVM accuracy. The best accuracy of the model with traditional SVM was 87.885%. After optimization, the highest accuracy obtained with GWO is 91.053% while PSO, hybrid PSO-GWO, and MPSO best accuracies are 90.736%, 90.657%, and 90.557%, respectively.

Index Terms: Document to Vector, Grey Wolf Optimizer, Particle Swarm Optimizer, Hybrid Particle Swarm Optimizer_Grey Wolf Optimizer, Opinion Mining, Radial Bias Function Kernel-based Support Vector Machine, Sentiment Analysis, Support Vector Machine Optimization, Twitter Application Programming Interface

1. INTRODUCTION

Nowadays, the use of the internet has become inseparable from our daily routines. Social media networks such as Facebook and Twitter have also been developed to give a right to people to easily share their viewpoints about any

Access this article online

DOI:10.21928/uhdjst.v4n1y2020.pp18-28	E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Qani Qabil Qasim, Department of Computer Science, University of Sulaimani, Sulaymaniyah, Iraq. E-mail: qani.qabil@gmail.com

Received: 23-10-2019 Accepted: 06-01-2020 Publishing: 13-02-2020

product or service in the form of short text. This makes them to be rich sources of data that can be valuable for various organizations and companies to find their fans' or customers' opinions about their products and services. In spite of companies, well-known people such as politicians and athletes may need to exploit those opinions and attitudes as well as to help them for making better decision-making in the future. However, data diversity and sparsity make it impossible for human to be able to analyze it. Here, the role of machine learning and automation can take a part to solve the problem of big data. Sentiment analysis (SA) or opinion mining techniques could be used [1].

SA refers to the task of finding the opinions of authors about specific entities that expressed in a written text [2].

In recent years, Twitter has become one of the most popular social media and microblogging platform where it is a convenient way for users to write and share their thoughts about anything within 280-characters length (called tweets). Twitter is used extensively as a microblogging service worldwide. Tweets consist of misspellings, slangs, and symbolic forms of words, which poses a major challenge for the conventional natural language processing or machine learning systems to be used on tweets [3].

Sentiment analyzer model can be built in three main approaches – lexicon-based approach, machine learning-based approach, and hybrid of both lexicon-based and machine learning approach. The machine learning approach is one of the most popular techniques that are widely used to build an automated classification model with the help of algorithms such as support vector machine (SVM), Naïve Bayes (NB), and so on. This is due to their ability to handle a large amount of data [4].

In this study, we propose a technique to promote SVM performance for SA by implementing four different meta-heuristic optimizers, namely, particle swarm optimizer (PSO), modified PSO (MPSO), grey wolf optimizer (GWO), and hybrid of PSO-GWO. The sentiment classification goes through four phases: Data collection, data pre-processing, feature extraction, and classification. In the first phase, Twitter Application Programming Interface (Twitter-API) enables developers to collect tweets about any keyword they desire and then followed by preprocessing phase to remove least informative data such as URL, hashtags, numbers, and so on. In the third phase, Document to Vector (Doc2Vec) approaches were used for vectorizing cleaned text, which is the numerical representation of text. PSO, GWO, and hybrid PSO-GWO are used to select the best parameters for the classifier (SVM) to classify generated features from the previous step.

The rest of the paper is structured as follows: In section 2, some previous related works in this field that has been conducted before being discussed, section 3 describes the material and methods used in this work, section 4 describes the problem statement, section 5 illustrates the proposed system model and methodology of analyzing the datasets, section 6 shows the results obtained from the model and discussed in detail, and finally, the conclusion and future work are stated in section 7.

2. RELATED WORK

Many researches and works have been developed in the field of SA. Researchers have proposed different solutions to different issues of SA in terms of improving performance of classification models, enhancing topic specific corpus, reducing feature-set size to shrink execution time of algorithms and space usage using different techniques.

Das *et al.* [5] review basic stages to be considered in SA, such as pre-processing, feature extraction/selection, and representation along with some data-driven techniques in this field such as SVM and NB as well as to demonstrate how they work and the measuring metrics such as (Precision, Recall, F1-Score, and Accuracy) to evaluate the model efficiency. They concluded that all the SA tasks are challenging and need different techniques to deal with each stage.

Naz *et al.* [6] illustrate the impact of different weighting feature schemes such as term frequency (TF), TF-inverse document frequency (TF-IDF), and binary occurrence (BO) to extract features from tweets along with different n-gram ranges such as unigram, bigram, trigram, and their combination, followed by feeding extracted feature from SemEval2016 dataset to train SVM. The best result they achieved is 79.6% for TF-IDF with Unigram range. They also used the sentiment score vector package to calculate the score of tweets into positive and negative forms to improve the performance of SVM, along with different weighting schemes and n-gram range, the highest accuracy achieved with SVC is 81.0% for BO with unigram range.

Seth *et al.* [7] proposed a hybrid technique for improving the efficiency and reliability of their model by merging SVM with the decision tree. The model performs a classification of tweets on the basis of SVM and adaboost decision tree individually. Then, a hybrid technique will be applied by feeding the outputs obtained from the two above mentioned algorithms as the input to the decision tree. Finally, they compared traditional techniques to the proposed model and obtained the accuracy of 84%, while prior accuracies were 82% and 67%.

Sharma and Kumari [8] applied SVM to find the polarity of four smartphone product review texts, whether positive or negative. Before applying SVM, they used part of speech (POS) tagging with tokens, then used clustering for TF-IDF features to find more appropriate centroids. The accuracy of the model was evaluated based on (Precision, Recall, F-score, and Accuracy) metrics, compared to previous studies on the same datasets where no POS and no clustering were performed. They obtained the accuracy of 90.99% while the best previous study accuracy was 88.5%.

Rajput and Dubey [9] made a comparative study between two supervised classification algorithms, namely, NB and SVM for making binary classification of customers review about six Indian stock market. The results show that SVM provides better accuracy, which was 81.647%, while NB accuracy was 78.469%.

Rane and Kumar [10] worked on a six major US Airline datasets for performing a multi-class (Positive, Negative, and Neutral) SA. Doc-2Vec deep learning approach has been used for representing these tweets as vectors to do a phrase-level analysis – along with seven (7) supervised machine learning algorithms (Decision Tree, Random Forest, SVM, K-Nearest Neighbors, Logistic Regression, Gaussian NB and AdaBoost). Each classifier was trained with 80% of the data and tested using the remaining 20% data. Accuracy of all classifiers was calculated based on (Precision, Recall, F1-Score) metrics. They concluded that the classification techniques used include ensemble approaches, such as AdaBoost, which combines several other classifiers to form one strong classifier which performs much better. The maximum achieved accuracy was 84.5%.

Shuai *et al.* [11], these authors carry out a binary SA on Chinese hotel reviews by using Doc2vec feature extraction technique and SVM, logistic regression and NB as a classifier. After making a performance comparison between classification algorithms based on the precision, recall rate, and F-measure metrics, SVM achieved the best accuracy in their experiment as follows: 79.5%, 87.92%, and 81.16% for all three metrics.

Bindal and Chatterjee [3] described two-step method (lexicon-based sentiment scoring in conjunction with SVM, point-wise mutual information utilized to calculate sentiment of tweets. They also discussed the efficacy of several linguistic features, such as POS tags and higher-order n-grams (Uni + Bi Gram, Uni + Bi + Tri Gram) in sentiment mining. Their proposed scheme had better "F-Score" average than commonly used one-step methods such as Lexicon, NB, Maximum Entropy, and SVM classifier, i.e., for Unigram range lexicon-SVM outperforms other classification methods with F-score of 84.39% while other methods F-score is 82.44%, 81.85%, 80.18%, and 83.56%, respectively.

Mukwazvure and Supreethi [12] used a hybrid technique which involves lexicon-based approach for detecting "news comments" polarity in (Technology, Politics, and Business) domains. Then, the outcome of lexicon-based is then fed to train two supervised machine learning algorithms: SVM and K-nearest neighbor (kNN) classifiers. Investigational results revealed that SVM performed better than kNN which were 73.6, 61.38, and 58.00 while kNN results were 74.24%, 56.27%, and 55.58%.

Flores *et al.* [13] made a comparative analysis of SVM algorithm-sequential minimal optimization with synthetic minority over-sampling technique (SMOTE) and Naive Bayes multinomial (NBM) algorithm with SMOTE for classification of two SA datasets gathered by students of University of San Carlos. The outcomes have shown that with 10-folds cross-validation SA for their datasets could perform better compared to 70:30 split. Performance of NBM with SMOTE was 72.33% and 78.02% and SVM with SMOTE were 83.16% and 82.22% in the term of accuracy.

3. MATERIALS AND METHODS

3.1. VADER

VADER stands for Valence Aware Dictionary and sEntiment Reasoner. It is a lexicon and rule-based SA tool that was developed by Hutto and Gilbert [14] in 2014. It is specifically attuned to do calculate the sentiment scores of texts expressed on social media. VADER uses a combination of a sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative. VADER not only tells about the positivity and negativity score but also tells us about how much positive or negative a sentiment is? VADER produces four sentiment metrics from these word ratings, the first three, positive, neutral, and negative, represents the proportion of the text that falls into those categories and the final metric is compound score which is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). According to their experiment, it is more effective than other existing lexicon-based approaches, for example, **SentiWordNet.**

3.2. Word Embedding and DOC2VEC

Word embedding, also known as (Word2Vec), is a technique for unique vector representation of each word with its semantic meaning of the word taken into consideration. Unlike bag of words, which is one of the most common techniques used for numerical representation of words that convert word to a fixed-length feature vector, it has some shortcomings. First, it does not consider the ordering of the words, ignores semantics of the words. For example, “powerful,” “strong,” and “Paris” are equally evaluated and generate a high dimensional feature set, so, it needs a lot of memory space [15].

In Word2Vec approach, each word is mapped to a vector in a predefined vector space. These vectors are learned using neural networks. The learning process can be done with a neural network model or by using an unsupervised process involving document statistics. Word2Vec can be implemented in two different architectures, first is continuous bag of word (CBoW), as shown in Fig. 1 which is designed to predict current words at an input of future words and history words and the second is skip-gram (SG) which is used to maximize the probability of surrounding words given the current word being used in word embedding [15], [16].

Doc2Vec, also called paragraph vector (PV), is a (Word2Vec) based learning approach that converts entire paragraph to a unique vector which is represented by a column in matrix D and every word is mapped to unique vector mapped in matrix W. The word and PVs are then concatenated to predict the next word. CBoW and SG methods have been tuned for Doc2Vec and converted into two methods, namely, distributed bag of words version of PVs (PV-DBOW) and distributed memory of PVs (PV-DM) [10], as shown in Figs. 2 and 3.

The DBOW model ignores the context words in the input, but force the model to predict words randomly sampled from the paragraph in the output [15].

In DM model, to predict the next word in a context, the paragraph and word vectors either being averaged (mean) which is called DM mean (DMM), or concatenated which is called DM concatenation (DMC) [15].

3.3. PSO Algorithm

PSO is a type of meta-heuristic algorithm developed by Dr. Kennedy and Dr. Eberhart in 1995 to optimize numeric problems iteratively. PSO simulates the behaviors of the animals’ groups searching for food, especially bird flocking or fish schooling. PSO starts through a randomly distributed group of agents called particles in a search space; every particle has self-own velocity [17].

Each particle has two “best” achieved positions; the first one is its best position or (local best position) referred to as “pbest.” And the second one is (global best position) referred to as “gbest.”

At each time the particles will move toward “pbest” and “gbest” based on a new “velocity” and some constant

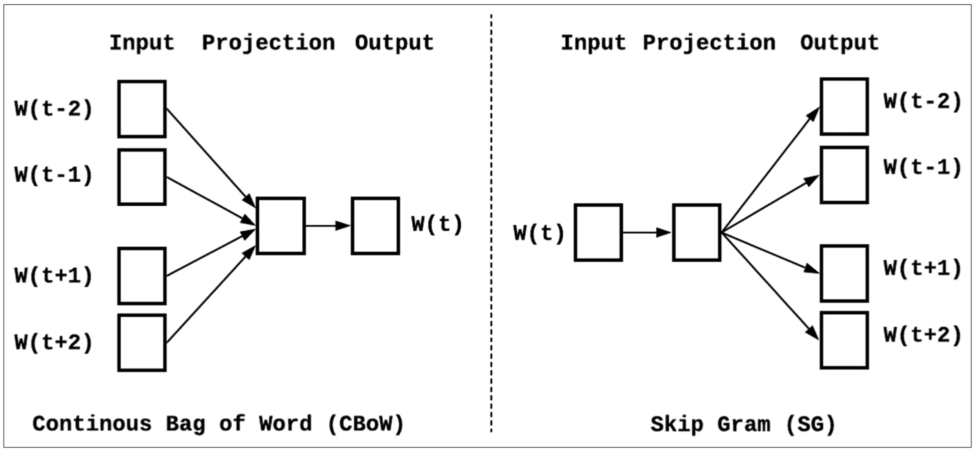


Fig. 1. Continuous bag of word and skip-gram.

coefficient parameters such as c_1, c_2, and w (inertia weight) and two random numbers.

In D-dimensional space, PSO algorithm can be described as follows:

$X_i = (X_{i1}, X_{i2}, X_{i3}, \ldots, X_{iD})$ represents the current position of the "particle," $V_i = (V_{i1}, V_{i2}, V_{i3} \ldots . V_{iD})$ and it refers to its velocity, the local best location is denoted as $Pbest,i = (P_{i1}, P_{i2}, P_{i3} \ldots . P_{iD})$, and global best position of all particles refers to $Pgbest,i = (P_{g1}, P_{g2}, P_{g3} \ldots . P_{gD})$.

At every iteration, each particle changes its position according to the new velocity.

$$v_i^{t+1} = w^* v_i^t + c_1 r_1 + \left(pBest_i^t - x_i^t\right) + c_2 r_2 + \left(gBest_i^t - x_i^t\right) \quad (1)$$

In this study, instead of multiplying w to only current velocity, after changing the current particle best position and group best position, we multiplied them all to "w." The formulated equation is:

$$v_i^{t+1} = w^* \left(\begin{array}{l} v_i^t + c_1 r_1 + \left(pBest_i^t - x_i^t\right) \\ + c_2 r_2 + \left(gBest_i^t - x_i^t\right) \end{array} \right) \quad (2)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (3)$$

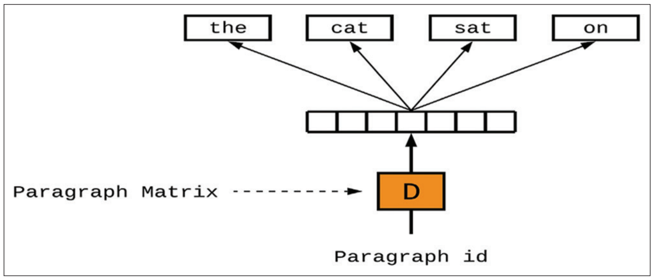


Fig. 2. Distributed bag of word of paragraph vector.

Where "i" refers to a particle, pBest, and gBest as the best particle position, best group position, and the parameters w, c_1 and c_2 are called inertia weighs. r_1 and r_2 are two random numbers in the range of (0, 1), v_i^t is a current velocity, v_i^{t+1} indicates new velocity in the next time or iteration. Furthermore, x_i^t is current particle position, x_i^{t+1} indicates the new particle position.

The pseudocode of PSO is:

```
Initialized number of particles (n_particle), D, n_iterations, c1, c2, and w.
For each particle i ∈ (n_particle)
Initialize $X_i$, $V_i$
End for
For each particle i in n_particle do
If $f(X_i) < f(Pi)$
$Pbest_i = X_i$
End if
If $f(Pbest_i) < f Gbest_i$
$Gbest = Pbest_i$
End if
End for
For each particle i in n_particle do
For each dimension d in D
Update velocity according to equation (1) for PSO and equation (2) for MPSO
Update position according to equation (3)
End for
End for
Iteration = Iteration +1
Until iteration > n_iterations.
```

3.4. GWO Algorithm

GWO algorithm is another type of swarm intelligence algorithm, proposed by Mirjalili *et al.* in 2014 [18], that mimics the leadership hierarchy and hunting mechanism of

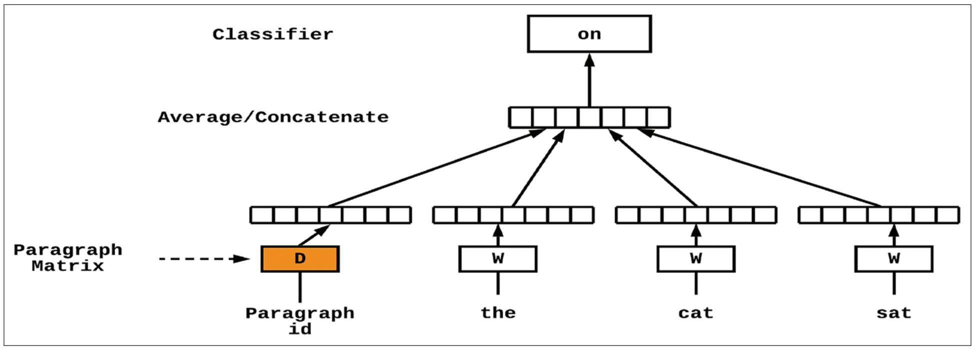


Fig. 3. Distributed memory of paragraph vector.

grey wolves in nature. Four types of grey wolves, such as alpha, beta, delta, and omega, are employed for simulating the leadership hierarchy. Furthermore, the three main steps of hunting, searching for prey, encircling prey, and attacking prey, are implemented.

3.4.1. Social hierarchy

The social hierarchy in this algorithm consists of four groups of wolves, namely, alpha (α), beat (β), and delta (δ), and the other is called omega (ω). In the GWO algorithm, the hunting (optimization) process is guided by α, β, and δ. The ω wolves follow these three wolves [18].

3.4.2. Encircling prey

Encircling prey means that grey wolves surround prey during the hunt, the following mathematical equations form the encircling behavior [18]:

$$\vec{D} = \left|\vec{C}.\vec{X}p(t) - \vec{X}(t)\right| \tag{4}$$

$$\vec{X}(t+1) = \left|\vec{X}p(t) - \vec{X}.\vec{D}\right| \tag{5}$$

Where t indicates the current iteration, $\vec{A}$ and $\vec{C}$ are coefficient vectors, $\vec{X_P}$ is the position vector of the prey, and $\vec{X}$ indicates the position vector of a grey wolf.

The vectors A and C are calculated as:

$$\vec{A} = 2\,\vec{a}\,.\,\vec{r_1} - \vec{a} \tag{6}$$

$$\vec{C} = 2\,\vec{r_2} \tag{7}$$

Where $\vec{a}$ linearly decreased from 2 to 0 throughout iterations and r_1, r_2 are two random vectors in the range of 0, 1.

3.4.3. Hunting

Grey wolves can identify the location of prey and encircle them. The hunt is usually guided by the alpha. Sometimes beta and delta might also get involved in hunting, alpha (best candidate solution), beta, and delta have better knowledge about the potential location of prey. Thus, the first three best solutions are selected to update their positions according to the position of the best search agents based on the following mathematical formulas [18]:

$$\vec{D}_\alpha = \left|\vec{C}_1 . \vec{X}_\alpha - \vec{X}\right|$$

$$\vec{D}_\beta = \left|\vec{C}_2 . \vec{X}_\beta - \vec{X}\right| \tag{8}$$

$$\vec{D}_\delta = \left|\vec{C}_3 . \vec{X}_\delta - \vec{X}\right|$$

$$\begin{aligned}\vec{X}_1 &= \vec{X}_\alpha - \vec{A}_1 .(\vec{D}_\alpha)\\ \vec{X}_2 &= \vec{X}_\beta - \vec{A}_2 .(\vec{D}_\beta)\\ \vec{X}_3 &= \vec{X}_\delta - \vec{A}_3 .(\vec{D}_\delta)\end{aligned} \tag{9}$$

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \tag{10}$$

The pseudocode of GWO as follows:

```
Initialize the grey wolf population Xi (i = 1, 2..., n)
Initialize a, A, and C
Calculate the fitness of each search agent using equations (8) and (9)
$X_\alpha$ = The first best search agent
$X_\beta$ = The second-best search agent
$X_\delta$ = The third best search agent
While (t < Max number of iterations)
For each search agent
Update the position of the current search agent according to equation (9)
End for
a=2−t*(2/(Max_iteration))
Calculate A, C using equations (6) and (7)
Calculate the fitness of each search agent using equations (8) and (9)
Update position of the current search agents according to equation (10)
t=t + 1
End while
Return $X_\alpha$.
```

3.5. Hybrid PSO-GWO

In hybrid PSO-GWO, the first three agents' position is updated in the search space by a mathematical equation 8. Instead of using common mathematical formulas, the exploration and exploitation of the grey wolf in the search space have been controlled by inertia constant [19]. The modified set of dominant equations is as follows:

$$\vec{D}_\alpha = \left|\vec{C}_1 . \vec{X}_\alpha - w^* \vec{X}\right|$$

$$\vec{D}_{\beta} = \left|\vec{C}_2 . \vec{X}_{\beta} - w^* \vec{X}\right| \tag{11}$$

$$\vec{D}_{\delta} = \left|\vec{C}_3 . \vec{X}_{\delta} - w^* \vec{X}\right|$$

Where $c_{1,}$ c_2, c_3, and w are constants,

To combine PSO and GWO variants, the velocity and updated equation are calculated as follows:

$$v_i^{t+1} = w^* \begin{pmatrix} v_i^t + c_1 r_1 \left(x_1 - x_i^t\right) \\ + c_2 r_2 \left(x_2 - x_i^t\right) + c_3 r_3 \left(x_3 - x_i^t\right) \end{pmatrix} \tag{12}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{13}$$

The pseudocode of hybrid PSO-GWO as follows:

```
Initialize $c_{1,}$ $c_2$, $c_{3,}$ t = 0,
w = 0.5 + r/2, velocity=random (search Agents No. dim),
Postion=dot (random (search Agents No, dim), (ub−lb)) + lb
While (t <Max_iteration)
For each search agent
a=2−t*(2/Max_iteration)
Calculate $A_1$, $A_2$, and $A_3$ according to equation (6)
Calculate the fitness of each search agent using equations (9) and (11)
Update velocity and position of the current search agent according to equations (12) and (13)
End for
t=t + 1
End while.
```

4. PROBLEM STATEMENT

As described in Section 1, machine learning techniques are popular ways of sentiment classification. In this work, to perform sentiment classification, Radial Bias Function kernel-based SVM (RBF-SVM) has been used. The accuracy and performance of this type of SVM mainly depend on the value of two parameters, namely, penalty **"C"** and **"gamma"** which known as hyperplane **"Soft Margin"** parameters. Hence, selecting optimal value for those parameters is a challenge to boost the classification model accuracy. To solve this problem, four meta-heuristic optimizer algorithms: PSO, MPSO, GWO, and hybrid of PSO-GWO have been implemented to select the best values for those parameters.

5. PROPOSED SYSTEM MODEL

In this study, four meta-heuristic optimizer algorithms have been implemented for selecting the best value to ***"Soft Margin" penalty "C" and "gamma"*** parameters to improve the accuracy of the RBF-SVM classifier. The work implemented on Dell Latitude E6540, Intel(R) Core(TM) i7-4610M CPU at 3.00GHz, 8-GB RAM, 64-Bit Windows-7 operating System. Fig. 4 is a flow diagram that displays basic architecture and steps of the proposed sentiment classification model.

5.1. Tweet Collection

To access Twitter and reading tweets from it, you have to make a Twitter developer account that known as Twitter-API. Twitter-API is an interface between the developers and Twitter that enables them to search for tweets based on their desired keyword through some secret key and tokens. In this work a Twitter-API is created called "Twitter-Sentiment-Analysis-20," to collect the most recent tweets according to some keyword such as Trump, Bitcoin, IoT, and Toyota using python code and categorizing to "Positive" and "Negative" using "VADER" [14] lexicon rule-based method then persist in mongo database collection or table. Table 1 shows the details about each keyword dataset size, and Fig. 5 shows a sample of data.

5.2. Pre-processing

Pre-processing means cleaning the text from the least important data. The datasets will go through the following steps for pre-processing task:

Removing duplicate tweets, convert the words to the lowercase, and replace emoticons symbols with a positive or negative opinion, according to Table 2.

The next step is removing URLs, slang correction (omg → oh my god), expand contraction (can't → cannot), stripping punctuation marks, special character and numbers, as well as multiple spaces, clearing from stop words, tokenizing, and

TABLE 1: Dataset size description

Keyword	Positive	Negative	Total
Trump	1339	1626	2965
Bitcoin	4923	2341	7264
IoT	10,700	1929	12,629
Toyota	14,332	6594	20,926
Total	31,294	12,490	43,784

TABLE 2: Emoticons and their meaning

:-), :-D, :-j, =p, :], :3	positive
:(, :[, ^o), :^), :@, =/	negative

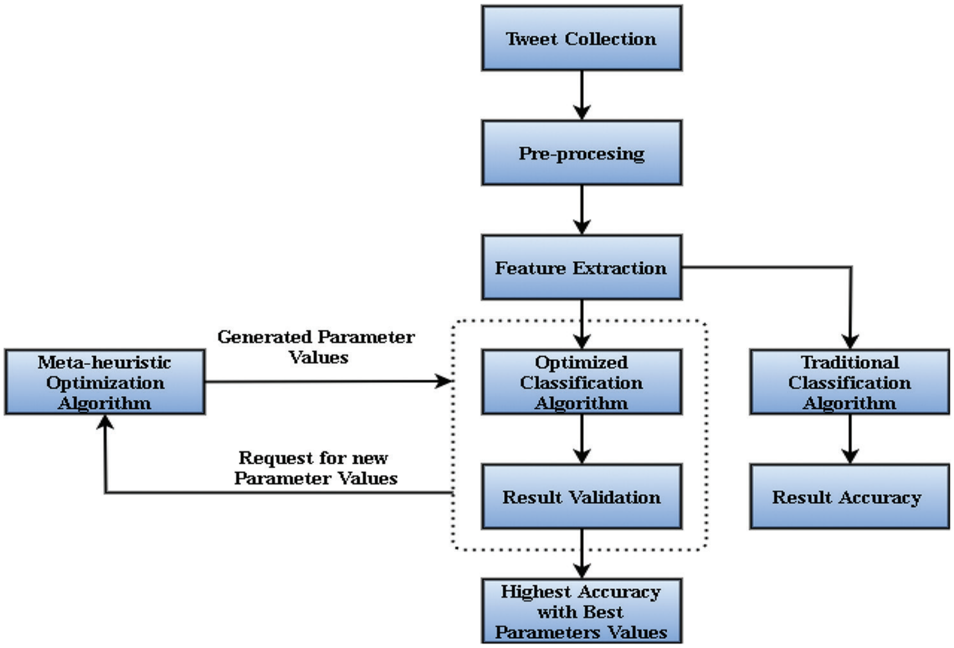
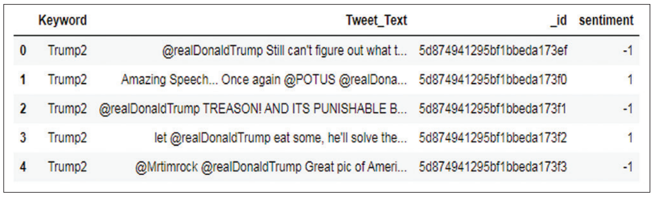

Tweet Collection
Pre-procesing
Feature Extraction
Generated Parameter Values
Meta-heuristic Optimization Algorithm
Optimized Classification Algorithm
Traditional Classification Algorithm
Request for new Parameter Values
Result Validation
Result Accuracy
Highest Accuracy with Best Parameters Values

Fig. 4. Flow diagram of the proposed model.



	Keyword	Tweet_Text	_id	sentiment
0	Trump2	@realDonaldTrump Still can't figure out what t...	5d874941295bf1bbeda173ef	-1
1	Trump2	Amazing Speech... Once again @POTUS @realDona...	5d874941295bf1bbeda173f0	1
2	Trump2	@realDonaldTrump TREASON! AND ITS PUNISHABLE B...	5d874941295bf1bbeda173f1	-1
3	Trump2	let @realDonaldTrump eat some, he'll solve the...	5d874941295bf1bbeda173f2	1
4	Trump2	@Mrtimrock @realDonaldTrump Great pic of Ameri...	5d874941295bf1bbeda173f3	-1

Fig. 5. Sample of collected tweets.

lemmatizing and finally, dropping duplicate tweets after pre-processing and protecting them in another mongo database collection.

5.3. Feature Extraction

Feature extraction is the most important phase. The purpose of this phase is to normalize the data by converting the words into vectors for the classification process. Gensim's deep learning library has been utilized for the numerical representation of each document. Doc2vec is a way of document embedding where each document is mapped to a vector in space. Doc2vec is Gensim's extended library of word2vec, which is used to find vector representations for each word [15]. Doc2Vec was proposed in two models, namely, DBoW and DM. DM is divided into two sub-model, namely, DMC and DMM. After preprocessing, the cleaned tweets will be split into two parts, which are training-set, composed of 80% of tweets, and test-set, composed of 20% of tweets, after that Doc2Vec models has been used to extract features from train-set and test-set. Doc2Vec models and their combination DBoW + DMC and DBoW + DMM are used to extract features from pre-processed tweets.

5.4. Classification

To perform the classification task, the RBF-SVM has been used. SVM one of the well-known supervised machine learning that broadly use in classification and regression tasks due to the ability to work with large amounts of data.

In the first approach, the traditional SVM with default value "1", and "scale" for **C** and **gamma** parameters, used to classify tweets. In the second approach, at each iteration, the RBF-SVM's **"C"** and **"gamma"** parameters took the position value of each agent. After finishing the last iteration, the best accuracy with respect to the best **C** and **gamma** values was presented. Finally, the accuracy of both classification approaches has been compared.

6. RESULTS AND DISCUSSION

Figs. 6-9 show an accuracy comparison between traditional SVM and optimized SVM with different Doc2Vec feature extraction models.

As it is shown in Fig. 6, all optimizers provide a better result for all Doc2Vec feature extraction methods. The hybrid of PSO-GWO provides better results in DBoW and DMC

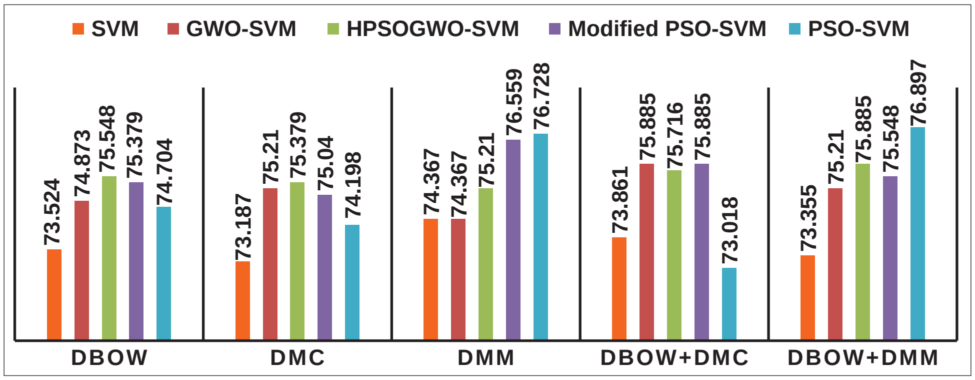


Fig. 6. Results of Trump dataset.

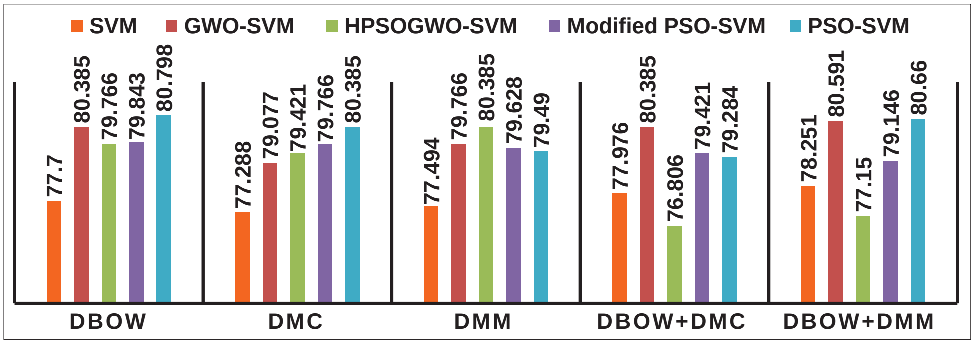


Fig. 7. Results of Bitcoin dataset.

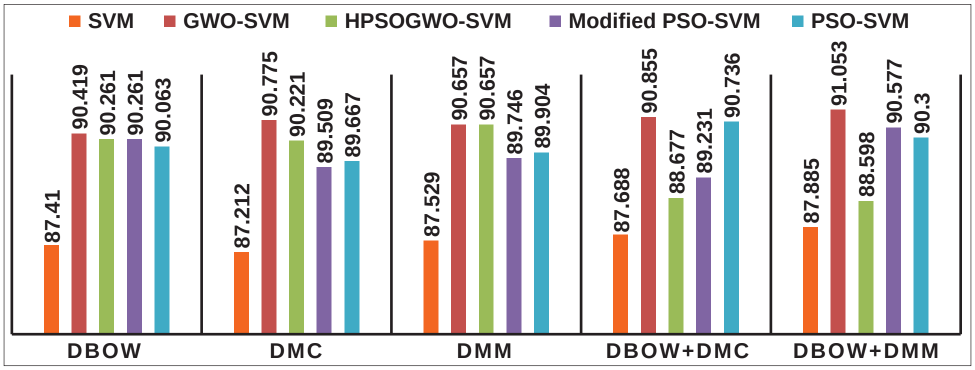


Fig. 8. Results of IoT dataset.

models. Furthermore, MPSO-SVM outperforms original PSO-SVM for DBoW, DMC, and DBoW + DMC models, respectively.

By looking at the "Bitcoin" dataset results, for DBoW, DMC, and DMM models, all optimizers provide a remarkable accuracy compared to traditional SVM, except for hybrid PSO-GWO that could not get expectable result for DBoW + DMC and DBoW + DMM models. MPSO-SVM provides better results than original PSO-SVM for both Doc2Vec DMM and DBoW + DMC models.

The results show that the model accuracy remarkably increased for all optimizers with different Doc2Vec models and their combinations, especially GWO that achieves the highest accuracy result that is 91.093% in DBoW + DMM, followed by MPSO and PSO. In DBoW, DMC, and DMM models hybrid of PSO-GWO provides a better result than PSO and MPSO, but in DBoW + DMC and DBoW + DMM combinations, it increased the model accuracy by <1%.

Finally, Fig. 9 illustrates that all optimizers outperform SVM when used alone, like "IoT" dataset, GWO-SVM

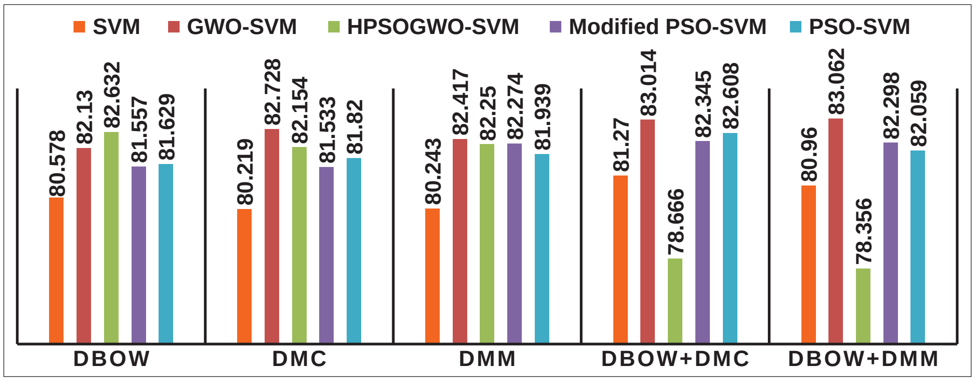


Fig. 9. Results of Toyota dataset.

outperforms other optimizers in all Doc2Vec models. Except for PSO-GWO with SVM that could not grant the expected result for DBoW + DMC and DBoW + DMM the same as the "Bitcoin" dataset.

7. CONCLUSION AND FUTURE WORK

In this work, we have carried out a comparative analysis between classification with traditional RBF-SVM and optimized RBF-SVM using four meta-heuristic optimizers, namely, PSO, MPSO, GWO, and hybrid of PSO and GWO. These optimizers are implemented for selecting the best values for hyperplane "***Soft Margin***" penalty "C" and ***gamma*** parameters of the RBF-SVM classifier. After testing our model on each dataset and with different Doc2Vec feature extraction methods. We came to the point that these optimizers have an important role in enhancing the accuracy of the classifier.

The results show that with a small dataset, MPSO provides a better result than the original PSO. In contrast, with increasing the dataset size, SVM with GWO achieves better accuracy compared to the rest optimizers.

Hybrid of PSO-GWO is effective in improving SVM accuracy in Doc2Vec DBoW, DMC, and DMM models, but it is not work well for combinations of DBoW + DMC and DBoW + DMM because of feature set nature was generated by merging these two models.

In future works, we will try to use these optimizers for parameter optimizing of some deep learning algorithms, i.e., rectified neural network weights to examine whether it performs better results than existing RBF-SVM model or not.

REFERENCES

[1] A. Go, R. Bhayani and L. Huang. "Twitter Sentiment Classification using Distant Supervision". Technical Report, Stanford University. p. 6, 2009.

[2] R. Feldman. "Techniques and applications for sentiment analysis: The main applications and challenges of one of the hottest research areas in computer science". *Communication of the ACM*, vol. 56, no. 4, pp. 82-89, 2013.

[3] N. Bindal and N. Chatterjee. "A two-step method for sentiment analysis of tweets." In: *15th International Conference Information Technology 2016*, Bhubaneswar, pp. 218-224, 2017.

[4] S. K. Jain and P. Singh. "Systematic Survey on Sentiment Analysis". In: *2018-1st International Conference on Secure Cyber Computing and Communication*, Jalandhar, pp. 561-565, 2019.

[5] M. K. Das, B. Padhy and B. K. Mishra. "Opinion mining and sentiment classification: A review". In: *Proceedings of the International Conference on Inventive Systems and Control 2017*, Malaysia, pp. 4-6, 2017.

[6] S. Naz, A. Sharan and N. Malik. "Sentiment Classification on Twitter Data Using Support Vector Machine". *2018 IEEE/WIC/ACM International Conference on Web Intelligence*, Santiago, pp. 676-679, 2019.

[7] P. Seth, A. Sharma and R. Vidhya. "Sentiment analysis of tweets using machine learning approach". *International Journal of Engineering and Technology*, vol. 7, no. 3.12, p. 434, 2018.

[8] A. K. Sharma. and D. S. U. Kumari. "Sentiment Analysis of Smart Phone Product Review using SVM Classification Technique". *International Conference on Energy, Communication, Data Analytics and Soft Computing*, Chennai, India, pp.1469-1474, 2017.

[9] V. S. Rajput and S. M. Dubey. "Stock market sentiment analysis based on machine learning". In: *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, Dehradun, pp. 506-510, 2017.

[10] A. Rane and A. Kumar. "Sentiment classification system of twitter data for us airline service analysis". *International Computer Software and Applications Conference*, vol. 1, pp. 769-773, 2018.

[11] Q. Shuai, Y. Huang, L. Jin and L. Pang. "Sentiment Analysis on Chinese Hotel Reviews with Doc2Vec and Classifiers". In: *018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 1171-1174, 2018.

[12] A. Mukwazvure and K. P. Supreethi. "A Hybrid Approach to

Sentiment Analysis of News Comments". In: *2015 4th International Conference on Reliability, Infocom Technologies and Optimization*, Noida, 2015.

[13] A. C. Flores, R. I. Icoy, C. F. Pena and K. D. Gorro. "An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set". In: *2018-4th International Conference on Engineering, Applied Sciences, and Technology, Explor Innovative Smart Solutions Social*, Phuket, pp. 1-4, 2018.

[14] J. Hutto and E. E. Gilbert. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text". In: *8th International Conference on Weblogs and Social Media*, Michigan, 2014.

[15] Q. Le and T. Mikolov. "Distributed Representations of Sentences and Documents". *31st International Conference on Machine Learning*, vol. 4, pp. 2931-2939, 2014.

[16] M. Bilgin and İ. F. Şentürk. "Sentiment Analysis on Twitter Data with Semi-supervised Doc2Vec". In: *2nd International Conference on Computer Science and Engineering UBMK 2017*, Turkish, pp. 661-666, 2017.

[17] R. Eberhart and J. Kennedy. "New Optimizer Using Particle Swarm Theory". In: *Proceedings International Symposium on Micro Machine and Human Science*, New York, pp. 39-43, 1995.

[18] S. Mirjalili, S. M. Mirjalili and A. Lewis. "Grey wolf optimizer". *Advances Engineering Software,* vol. 69, pp. 46-61, 2014.

[19] N. Singh and S. B. Singh. "Hybrid algorithm of particle swarm optimization and grey wolf optimizer for improving convergence performance". *Journal of Applied Mathematics*, vol. 2017, pp. 15, 2017.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Sentiment Analysis Using Hybrid Feature Selection Techniques

Sasan Sarbast Abdulkhaliq[1], Aso Darwesh[2]

[1]Department of Computer Science, University of Sulaimani, Sulaymaniyah, Iraq, [2]Department of Information Technology, University of Human Development, Sulaymaniyah, Iraq

ABSTRACT

Nowadays, people from every part of the world use social media and social networks to express their feelings toward different topics and aspects. One of the trendiest social media is Twitter, which is a microblogging website that provides a platform for its users to share their views and feelings about products, services, events, etc., in public. Which makes Twitter one of the most valuable sources for collecting and analyzing data by researchers and developers to reveal people sentiment about different topics and services, such as products of commercial companies, services, well-known people such as politicians and athletes, through classifying those sentiments into positive and negative. Classification of people sentiment could be automated through using machine learning algorithms and could be enhanced through using appropriate feature selection methods. We collected most recent tweets about (Amazon, Trump, Chelsea FC, CR7) using Twitter-Application Programming Interface and assigned sentiment score using lexicon rule-based approach, then proposed a machine learning model to improve classification accuracy through using hybrid feature selection method, namely, filter-based feature selection method Chi-square (Chi-2) plus wrapper-based binary coordinate ascent (Chi-2 + BCA) to select optimal subset of features from term frequency-inverse document frequency (TF-IDF) generated features for classification through support vector machine (SVM), and Bag of words generated features for logistic regression (LR) classifiers using different n-gram ranges. After comparing the hybrid (Chi-2 + BCA) method with (Chi-2) selected features, and also with the classifiers without feature subset selection, results show that the hybrid feature selection method increases classification accuracy in all cases. The maximum attained accuracy with LR is 86.55% using (1 + 2 + 3-g) range, with SVM is 85.575% using the unigram range, both in the CR7 dataset.

Index Terms: Binary Coordinate Ascent, Bag of Words, Chi-square, Logistic Regression, n-grams, Opinion Mining, Sentiment Analysis, Support Vector Machine, Twitter-Application Programming Interface, Term Frequency-Inverse Document Frequency

1. INTRODUCTION

In the past two decades, the internet and more specifically social media have become the main huge source of opinionated data. People broadly use social media such as Twitter, Facebook, Instagram to express their attitude and opinion toward things such as products of commercial companies, services, social issues, and political views in the form of short text. This steadily growing subjective data makes social media a tremendously rich source of information that could be exploited for the decision-making process [1], [2].

As mentioned before, one of the most popular and widespread social media is Twitter. It is a microblogging platform that allows people to express their feelings toward vital aspects in the form of a 280-character length text called

Access this article online

DOI:10.21928/uhdjst.v4n1y2020.pp29-40	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: Sasan Sarbast Abdulkhaliq, Department of Computer Science, University of Sulaimani, Sulaymaniyah, Iraq. E-mail: Sasan.abdulkhaliq@uhd.edu.iq

Received: 19-12-2019 Accepted: 07-01-2020 Published: 13-02-2020

tweet. Moreover, Twitter is used by almost all famous people and reputable companies around the world [3]. This has made Twitter become a very rich source for data, as it has approximately 947 million users and 500 million generated tweets each day. Hence, companies and organizations trying to benefit from this huge and useful data to find their customer's satisfaction with their products and service levels they offer, politicians wish to envisage their fans' sentiments. However, it is impractical for a human to analyze this massive data, to avoid this, sentiment analysis, or opinion mining techniques can be used to automatically discover knowledge and recognize predefined patterns within large sets of data [4].

Sentiment analysis is a natural language processing (NLP) technique for detecting or calculating the mood of people about a particular product or topic that has been expressed in the form of short text using machine learning algorithms. The main goal of sentiment analysis is to build a model to collect and analyze views of people about a particular topic and classify them into two main classes positive or negative sentiment [5].

One major step in sentiment analysis is feature extraction, which is the numerical representation of tokens in a given document. But actually, features can be noisy due to the data collection step as a consequence of data collecting technologies imperfection or the data itself which contains redundant and irrelevant information for a specific problem. This degrades the performance of the learning process, reduces the accuracy of classification models, increases computational complexity of a model, and leads to overfitting. Thus, high dimensionality problem should be handled when applying machine learning and data mining algorithms with data that have high dimensional nature. To handle this problem, feature selection techniques could be used to select the best features from available feature space for classification, regression, and clustering tasks.

Besides, one of the most important aspects of classification is accuracy. Feature selection plays a key role in improving accuracy by identifying and removing redundant and irrelevant features. Feature selection techniques are broadly classified into three categories, which are filter-based, wrapper-based, and embedded or hybrid methods [6].

In this work, Twitter-application programming interface (Twitter-API) being used to extract most recent tweet according to specific keywords such as (Amazon, Trump, Chelsea FC, CR7) and label them using lexicon-based approach into two categories which are positive and negative, followed by pre-processing step to remove irrelevant terms. Bag of word (BoW) technique and term frequency-inverse document frequency (TF-IDF) weighting scheme along with different n-gram ranges such as (Unigram, Bigram, Trigram, and Uni + Bi + Tri-gram) are used to extract features from tweets. The next step is selecting the best features to feed to our model which consists of a filter-based method Chi-square (Chi-2) to select the most relevant attributes within generated features, followed by selecting the best feature-subset within features using wrapper-based method binary coordinate ascent (BCA) to improve classification accuracy. Two supervised machine learning algorithm has been chosen for their performance and simplicity, namely, logistic regression (LR) and linear support vector machine (SVM) to perform binary sentiment classification on selected feature-subsets.

1.1. Related Works

In Zhai *et al.* [7], the authors utilize Chi-2 for feature selection with single and double words, together with SVM and Naïve Bayesian as classifiers. Obtained is that accuracy gradually increases as the number of features increase. With 1300 features accuracy hits 96%, then it remains slightly stable until 2000 features. Meanwhile, the accuracy of information (information gain [IG]) remains below Chi-2 for the whole features. Besides, we can see that the feature selection applied as combination features could also affect the performance of the classification. It extracts context-related multi-features with little redundancy which can help to reduce the internal redundancy, consequently improve the classification performance. Shortcomings of Chi-2 is also pointed, as it only considers the frequency of words within the document, regardless of the effectiveness of the word, and as result, it can cause the removal of some effective but low-frequency words during feature selection step.

In contrast, the researchers in Kurniawati and Pardede [8] proposed a hybrid feature selection technique on a balanced dataset, which composed of particle swarm optimization (PSO) plus IG together, followed by classification step using SVM, achieving a better result than using each one separately. The results are as follows: Compared to using SVM alone, the proposed method achieves 1.15% absolute improvements. Compared to IG + SVM and PSO + SVM, the method achieves 1.97% and 0.6% improvements, respectively. Overall, the system achieved 98% accuracy using area under the curve accuracy measure.

In the research done by Kaur *et al.* [9], the proposed system uses k-nearest neighbors (KNN) as a classifier for classifying sentiments of text on e-commerce sites into positive, negative, and neutral sentiments on tweeter dataset. Features generated using n-gram before the KNN classifier took place. The performance of the proposed model was analyzed using precision, recall, and accuracy followed by comparing them to results obtained from the SVM classifier. The outcome was that the proposed system could outperform SVM classifier by 7%.

On the other hand, the work done by Zhang and Zheng [10] incorporates part of speech tagging to specify adverbs, adjectives, and verbs in the text first, then applied term frequency-inverse document frequency (TF-IDF) for generating features as a result of their corresponding word weights. Then, features were adopted for classification and fed to both SVM and extreme learning machine with kernels to classify sentiments of Hotel reviews in Chinese. They attained that essential medicines list accuracy is slightly better than SVM when introduced with the kernel and takes an effectively shorter time of training and testing than SVM.

In the work Joshi and Tekchandani [11], researchers have made a comparative study among supervised-learning algorithms of machine learning such as SVM, maximum entropy (MaxEnt), and Naïve Bayes (NB) to classify Twitter movie review sentiments using unigram, bigram, and unigram-bigram combination features.

Their study result shows that SVM reaches maximum accuracy of 84% using hybrid feature (unigram and bigram), leaving other algorithms behind. Furthermore, they observed that MaxEnt Excels NB algorithm when used with bigram feature.

In Luo and Luo [12] researchers proposed a new odds ratio (OR) + SVM-recursive feature elimination (RFE) algorithm that combines OR with a recursive SVM (SVM-RFE), which is an elimination based function. OR is used first as a filtering method to easily select a subset of features in a very fast manner, followed by applying SVM-RFE to precisely select a smaller subset of features. Their observation result emphasizes that OR + SVM-RFE attains better classification performance with a smaller subset of features.

In Maipradit *et al.* [13], a group of researchers suggests a method for classifying sentiments with a general framework for machine learning.

n-gram IDF has been used in feature generation and selection stage. As the classification stage, an automated machine learning tool has been used which makes use of auto-sklearn for choosing the best classifier for their datasets and also choosing the best parameter for those classifiers automatically. Classification is applied on different publicly available datasets (Stack Overflow, App reviews, Jira issues).

However, their study might not be feasible to be generalized for every other dataset; their datasets were specifically chosen for comments, reviews, and questions and answers. Their classification result achieved the best average model evaluation metrics F1, precision, and recall score values for all datasets in predicting class labels for positive, negative, and neutral classes for abovementioned datasets. Moreover, the highest F1-score value achieved was 0.893 in positive comments, 0.956 in negative comments of Jira issues dataset, and 0.904 F1-score value for in neutral comments of stack overflow dataset.

In work done by researchers in Rai *et al.* [14], tweets have been gathered from Twitter's API first. Later on weights for each word within review tweets have been calculated. Followed by selecting the best features using the NB algorithm, and consequently classifying the sentiment of reviews using three different machine learning classifiers, namely, NB classifier, SVM, and Random Forest Algorithm. After measuring they realized that all three algorithms are performing the same for 50 tweets, but increasing the number of tweets and adding more features changes the accuracy and other measures dramatically. As a part of their observation, they noticed that increasing the number of tweets from 50 to 250 will increase the accuracy of NB and SVM up to 83% approximately while adding more features to each algorithm gives slightly better classification accuracy up to 84% for 250 tweets.

Another group of researchers in Naz *et al.* [15] has employed another method to classify sentiments of Twitter data. The method composed of a model that employs a machine learning algorithm utilizing different feature combinations (unigram, bigram, trigram, and the combination of unigram + bigram + trigram) + SVM to improve classification accuracy. Furthermore, three different weighting approaches (TF and TF-IDF and binary) have been tried with the classifier using different feature combinations to see the effect of changing weights on classification accuracy. The best accuracy achieved by this approach was 79.6% using unigram with TF-IDF. Furthermore, sentiment score vector is created to save overall scores tweets and then associated with the feature vector of tweets, then classified them using SVM with different n-grams of features from different feature selection methods as mentioned before. The result shows that using a sentiment score vector with unigram + SVM gives the best accuracy result compared to other n-grams which were 81%.

Another research has been carried out by Wagh and Punde [16], a comparative study among different machine learning approaches have been applied by other researchers. The focus of their work was to discuss the sentiment analysis of Twitter tweets, considering what people like or dislike. They perceived that applying machine learning algorithms such as SVM, NB, and Max-Entropy on results of semantic analysis WordNet to form hybrid approach can improve accuracy of sentiment analysis classification by 4–5% approximately.

Another research has been performed by Iqbal *et al.* [17], in which multiple feature combinations are fed to (NB, SVM, and MaxEnt) classifiers for classifying movie reviews from the IMDb dataset and tweets from Stanford Twitter sentiment 140 dataset, in the term of people's opinion about them. The experiment incorporates four different sets of features, each of which are a combination of different single features as following: Combined single word features with stopword filtered word features as (set 1), unigram with bigram features as (set 2), bigram with stopword filtered word features as(set 3), and most informative unigram with most informative bigram features as(set 4). Chi-2 has been used as a supervised feature selection technique to obtain more enhanced performance by selecting the most informative features. And also, Chi-2 helps to decline the size of training data. Their result shows that combining both unigram and bigram features and subsequently feeding it to MaxEnt algorithm gives the best result in term of F1-score, precision, and recall compared to two other algorithms, and also compared to using single feature and baseline model which is SentiWordnet (SWN) method by 2–5%.

Another research was done by Rane and Kumar [18] on a dataset containing tweets of 6 US Airlines and carried out sentiment analysis to extract sentiments as (positive, negative, and neutral). The motivation of the research was to provide airline companies a general view of their customer's opinions about airline services to provide them a good level of service to them. As the first step preprocessing has been performed, followed by a deep learning concept (Doc2vec) to represent tweets as vectors which makes use of distributed BoW and distributed memory model, which preserves ordering of words throughout a paragraph, to do phrase-level sentiment analysis. The classification task has been done using seven different supervised and unsupervised learning, namely, decision tree, random forest, SVM, KNN, LR, Gaussian NB, and AdaBoost. After classification, they attained acceptable accuracy that can be used by airline companies with most of the classifiers as follows: Random forest (85.6%), SVM (81.2%), AdaBoost (84.5%), and LR (81%) are among the best classifiers as result, they concluded that the accuracy of the classifiers are high enough, that makes them reliable to be used by airline industry to explore customer satisfaction.

Another work done by Jovi *et al.* [19] to review available feature selection approaches for classification, clustering and regression tasks, along with focusing on their application aspects. Among which IG (precision) and normal separation (accuracy, F-measure, and recall) have the best performance for text classification tasks, whereas iterative feature selection (Entropy, precision) attains the best performance for text clustering. Results show that using hybrid approaches for feature selection, consisting of a combination of the best properties from filter, and wrapper methods giving out the best result by applying first, a filter method to reduce feature dimensions to obtain some candidate subsets. Then applying a wrapper method was based on a greedy approach to find the best candidate subset.

In Rana and Singh [20] authors have proposed a model for classifying movie reviews using NB classifier and Linear SVM classifiers. They realized that applying the classifiers after omitting synthetic words gives a more accurate result. Their result shows that SVM achieves better accuracy than the NB classifier. Furthermore, both algorithms distinctly performed better for genre drama, reaching 87% with SVM and 80% with the NB algorithm.

In Kumar *et al.* [21] authors have developed a classification model to classify reviews from websites such as amazon.com. After extracting reviews of three different products, namely, APPLE IPHONE 5S, SAMSUNG J7, and REDMI NOTE 3 from the website automatically, they applied NB, LR, and SWN algorithms for classifying reviews in the term of positive and negative. After using quality measure metrics (F1 score, recall, and precision), NB has achieved the best result among three classifiers with F1-scores: 0.760, 0.857, and 0.802 for three above-mentioned datasets, respectively.

In Iqbal *et al.* [22] researchers proposed a hybrid framework to solve scalability problems that appear when feature set grows in sentiment analysis. Using genetic algorithm (GA) based technique to reduce feature set size up to 42% without effecting accuracy. Comparing their proposed (GA) based feature reduction technique against two other well-known techniques: Principal Component Analysis (PCA) and Latent Symantec Analysis, they affirmed that GA based technique had 15.4% increased accuracy over PCA and up to 40.2% increased accuracy over latent semantic analysis. Furthermore, they employed three different methods of sentiment analysis, which are SWN, Machine Learning, and

Machine Learning with GA optimized feature selection. In all cases, the SWN approach has lower accuracy than two other mentioned approaches achieving its best accuracy of 56%, which impractical for real-time analysis. Their developed model which incorporates GA results in reducing feature size by 36–43% in addition to 5% increased efficiency when compared to the ML approach due to reduced feature size. They have tested their proposed model using six different classifiers on different datasets, the classifiers are, namely, J48, NB, PART, sequential minimal optimization, IB-k, and JRIP. Among all classifiers, NB classifier has shown the highest accuracy (about 80%) while using GA based feature selection on Twitter and reviews dataset, on the other hand, IB-k outperformed other classifiers with accuracy 95% while applying on the geopolitical dataset. Another evaluation is done for the scalability and usability of their proposed technique using execution time comparison. They found that the system showed a linear speedup with the increased dataset size. However, the technique consumed 60–70% of the aggregate execution time on customer reviews dataset, but it results in a speedup of modeling the classifiers up to 55% and remains linear, confirming that proposed algorithm is fast, accurate, and scales well as the dataset size grows.

1.2. Problem Statement

Thus, twitter is one of the richest sources of opinionated data; there is a big demand on analyzing twitters' data nowadays for the process of decision making. However, these data are unstructured and contain a lot of irrelevant and redundant information that leads to high-dimensionality of feature space, consequently analyzing them properly and accurately by machine learning, and data mining techniques is a big challenge. High dimensionality degrades the performance of the learning process, reduces the accuracy of classification models, increases computational complexity of a model, and leads to model overfitting. To overcome this problem, a hybrid of two feature selection methods proposed to remove the redundant and irrelevant features to select the best feature subset for classification task automatically. In this work, we use Chi-2 to calculate the correlation between attributes and the class labels. Low correlation of a particular feature means that the feature is irrelevant to the class label and needs to be removed prior to classification. In this way features are reduced. but there is still the problem of redundant features. The redundant features were removed by applying BCA, which uses an objective function to selects optimal feature subset from features were selected by Chi-2. As result irrelevant and redundant features were removed, which lead to solve high dimensionality problem in model building, and eventually classification accuracy is improved.

2. METHODOLOGY

The main objective of this study is to select optimal or sub-optimal feature-subset to perform Twitter sentiment analysis throughout utilizing filter-based method (Chi-2) and hybrid filter + wrapper method, namely, Chi-2 + BCA to improve the accuracy of the classification model. The work was implemented on Acer vn7-591g series laptop, Intel(R) Core(TM) i7-4710HQ CPU at 2.50 GHz (8 CPUs), 16-GB RAM, Windows 10 Home 64-bit. Fig. 1 depicts the flow diagram of our proposed system model and the following subsections explain each step of developing the proposed model in detail:

2.1. Data Collection

Twitter-API with python code is used to automatically download most recent tweets about (Amazon, Trump, Chelsea FC, and CR7) keywords, respectively, and lexicon rule-based method being utilized to assign positive and negative scores for each tweet, then protecting it in Comma Separated Value (.csv) file. Table 1 illustrates the details of each keyword dataset and Fig. 2 shows a sample of collected tweets.

2.2. Pre-processing

Text pre-processing is the first step in Twitter sentiment analysis, the tweets should go through some pre-processing step such as removing duplicate tweets, converting to lowercase, replacing emoji's with their meaning, removing URLs, usernames, and expand contractions such as (can't → cannot), replacing slang word like (omg → oh my god), reducing repeated character to only two character, removing numbers, special characters, punctuation marks, multiple space, tokenizing, removing stop-words, and lemmatizing, followed by removing duplicate tweets after pre-processing. Finally persisting the cleaned in another (.csv) file. Table 2 shows the emoji's and their meaning.

TABLE 1: Dataset size description

Keyword	Positive	Negative	Total # of tweets
Amazon	432	791	1223
Trump	1054	1200	2254
Chelsea FC	2239	761	3000
CR7	2816	1184	4000

TABLE 2: Emoji's and their meaning

Emoji	Meaning
:), :-D, :-j, =p, :3	Positive
:(, :|, :^), +o(, :=&	Negative

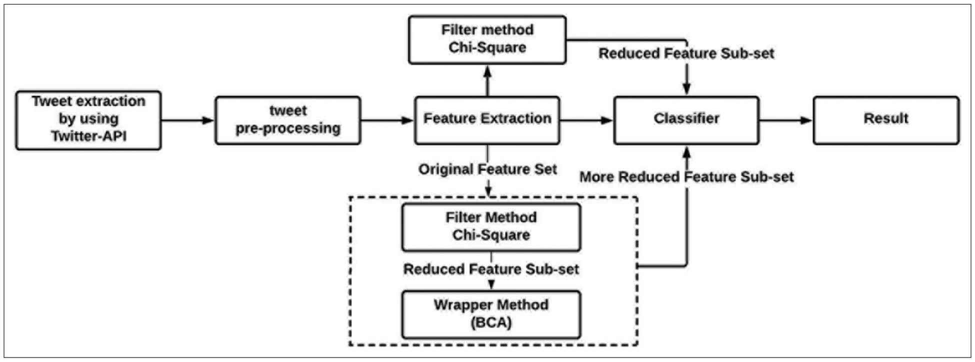
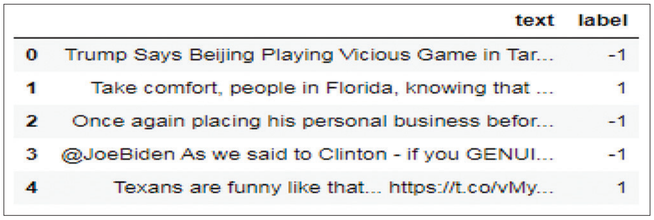


Fig. 1. Flow diagram of proposed system model.



	text	label
0	Trump Says Beijing Playing Vicious Game in Tar...	-1
1	Take comfort, people in Florida, knowing that ...	1
2	Once again placing his personal business befor...	-1
3	@JoeBiden As we said to Clinton - if you GENUI...	-1
4	Texans are funny like that... https://t.co/vMy...	1

Fig. 2. Sample of collected tweets.

2.3. Feature Extraction

Feature extraction is the process of converting the text data into a set of features or numerical representations of words or phrases. The performance of the machine learning process depends heavily on its features, so it is crucial to choose appropriate features for your classification model. On the other hand, applying different n-grams which are a different combination of words within the document gives out different accuracy results. We used (unigram, bigram, trigram, and combination of all) as the most commonly used ranges to see the impact of each of them on classification results. The followings are two Feature Extraction Methods used by the proposed model

1) TF-IDF: TF-IDF stands for Term Frequency – Inverse Document Frequency, it is a simple and effective metric that represents how "important" a word is to a document in the document set. It has many uses; one of its common uses is for automated text analysis. It is very useful for scoring words in machine learning algorithms in NLP.
   TF-IDF for a word in a document is calculated by multiplying two different metrics:
   - TF: Calculating how many times a term occurs in a document. The reason behind using it is that words that frequently occur in a document are probably more important than words that rarely occur. The result is then normalized by dividing it by the number of words in the whole document. This normalization is done to prevent a bias toward longer documents.
     TF(t) = (Number of times term t appears in a document)/(Total number of terms in the document).
     And its mathematical formula is:

$$tf_{td} = \frac{n_{td}}{\sum_k n_{kd}} \tag{1}$$

     Where n_{td} is the number of times that term t occurs in document d, and n_{kd} is the number of occurrences of every term in document d.
   - IDF: Measures how important term is by taking the total number of documents in the corpus and divide it by the number of documents where the term appears. It is calculated by:
     IDF(t) = log(Total number of documents/Number of documents with term t in it).
     And its mathematical formula is:

$$idf_t = \log \frac{|D|}{|D_t|} \tag{2}$$

     Where $|D_t|$ is the total number of documents in the corpus, $|D_t|$ is the number of documents where the term t appears in it.
     Hence, the TF-IDF is the multiplication of eq. 1 and eq. 2. Which is:

$$tf - idf_t = \frac{n_{td}}{\sum_k n_{kd}} \log \frac{|D|}{|D_t|} \tag{3}$$

2) BoW: One of the simplest types of feature extraction models is called BoW. The name BoW refers to the fact that it does not take the order of the words into account. Instead one can imagine that every word is put into a bag. It simply counts the number of occurrences of each word within a document and keeps the result in a vector which is known as count-vector.

2.4. Feature Selection

The main goal of feature selection is to select an optimal group of features for learning algorithms from the original feature set to retain the most useful features as possible and removes useless features that do not affect classification result. In this way, feature selection reduces the high dimensionality of data by eliminating irrelevant and redundant features. Thus, improves the model accuracy, reduces computation, and training time. It also reduces storage requirement, and avoids overfitting. Feature selection methods mainly divided into three categories which are Filter Methods, Wrapper Methods, and Hybrid or embedded methods. In general, feature selection methods are composed of four main steps, namely, feature subset generation, subset evaluation, stopping criterion, and result validation. Fig. 3 illustrates the basic steps of the feature selection process.

In this work, we are using Chi-2 filter-based method in conjunction with BCA wrapper-based method to form a hybrid feature subset selection technique to select the best subset for our classification models. First, we employed Chi-2 to remove irrelevant features, leading to produce reduced feature set. Then applied BCA for selecting more optimal subset of features that are more reduced feature subset. The operation details of both methods are described below:

2.5. Chi-2

Chi-2 is a type of filter-based feature selection method; it is used to select informative features and ranking them to remove irrelevant features with low ranks. In statistics, the Chi-2 test is used to examine the independence of two events. The events, and are assumed to be independent if:

$$p\left(XY\right) = p\left(X\right) p(Y) \qquad (4)$$

In text feature selection, these two events correspond to the occurrence of a particular term and a class, respectively. Chi-2 can be computed using the following formula:

$$Chi - 2\left(t,C\right) = \sum_{t\in\{0,1\}} \sum_{C\in\{0,1\}} \frac{(N_t,C - E_t,\, C)^2}{E_t,C} \qquad (5)$$

N is the observed frequency, and E is expected frequency for both of term t and Class C. CHI2 is a measure of how much expected count E and observed count N deviate from each other. A high value of Chi-2 indicates that the hypothesis of independence is not correct. The occurrence of the term makes the occurrence of the class more likely if the two events are dependent. Consequently, the regarding term is relevant as a feature. The Chi-2 score of a term is calculated for individual classes. This score can be globalized over all classes in two ways. The first way is to compute the weighted average score for all classes and the second way is to choose the maximum score among all classes. In this paper, the former approach is preferred to globalize the Chi-2 value for all classes in the corpus.

$$\sum P\left(C_i\right).\mathrm{Chi} - 2(t,C_i) \qquad (6)$$

Where $P(C_i)$ is the class probability and Chi-2 (t, C_i) is the class-specific Chi-2 score of term t.

2.6. BCA

BCA is a wrapper based feature selection method, was introduced by Zarshenas and Suzuki [23] in 2016. The goal of the BCA algorithm is to choose optimal or sub-optimal sub-set from available features from feature space that makes machine-learning algorithms the highest possible performance for a specific task, such as classification. The BCA algorithm iteratively adds and removes features to and from the selected subset of features based on the objective function values starting from an empty sub-set. At each iteration, the BCA checks whether the existence of a particular feature, in a given subset of features, improves or degrades the classification performance. If a feature was included in or removed accidentally from the feature sub-set, the BCA algorithm will be capable of correcting the wrongly taken decisions in the proceeding scans to approximate the optimal solution as much as possible. Compared to Sequential Feature Selection (SFS) and Sequential Forward Floating Selection (SFFS) are two of the most popular wrapper-based FSS techniques and filter-wrapper Incremental Wrapper Subset Selection (IWSSr), the BCA is more efficient than

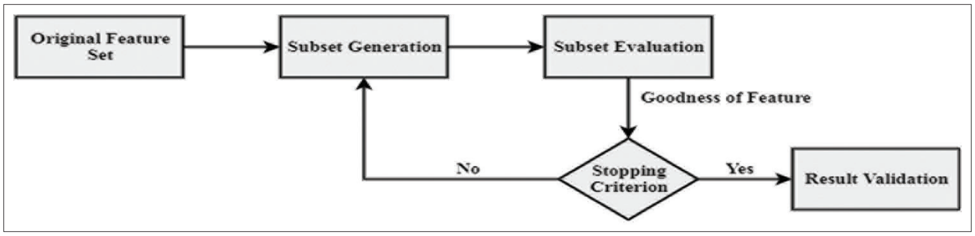


Fig. 3. A general framework of feature selection.

both of them in terms of processing time and classification accuracy. We came to the point that this algorithm is an effective feature selection method for the classification of datasets with a high number of initial attributes. Fig. 4 shows the work of the algorithm.

2.7. Hybrid Feature Selection Method

Filter methods are fast because they use mathematics and statistics for selecting features. The proposed model uses Chi-2 as filter-based method to remove irrelevant features and producing a reduced feature subset from the original feature set. In addition, wrapper-based methods are more accurate because they work as a part of the classification algorithms to evaluate usefulness of a particular feature, but they are computationally slow when applied on original feature set. Hence, the proposed model takes advantage of characteristics of both feature selection methods by first removing irrelevant features from the original feature set using Chi-2, and then applying BCA to the features those are selected by Chi-2 to select more optimal feature subset to enhance classification accuracy.

2.8. Classification Algorithms

In sentiment analysis classification essentially means categorizing data into different classes based on some calculation to determines the sentiment of the text. In our study, we applied two machine learning algorithms, namely, Linear SVM (LSVM) and LR for binary classification (positive and negative) of Twitter data.

SVM is a non-probabilistic machine learning algorithm. It is primarily used for classification in machine learning and could be fine-tuned for using with regression. The aim of SVM is to find the optimal decision boundary between classes

By transforming our data with the help of mathematical functions called Kernels. The best decision boundary created is called a hyperplane.

With a linearly separable data linear kernel is used. Since our class labels are linear (only positive and negative), we will perform classification with "linear SVM."

LR is a statistical machine learning algorithm for predicting classes which have dichotomous nature. Dichotomous mean having just two possible classes, binary by another mean. The term logistic mean logit function (a probabilistic function which returns values just in [0,1]).

3. RESULTS AND DISCUSSION

Based on the results attained from the two classifiers: TF-IDF with SVM and BoW with LR along with different n-grams, five-fold cross-validation, using Chi-2 and Hybrid Chi-2 + BCA feature selection methods, we achieved accuracy levels illustrated in the following Figs. 5-12:

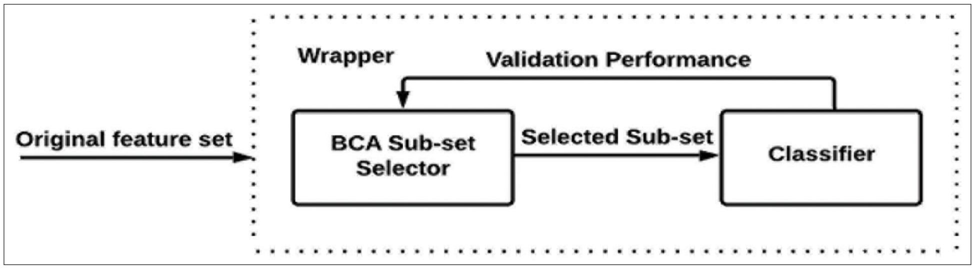


Fig. 4. Binary coordinate ascent.

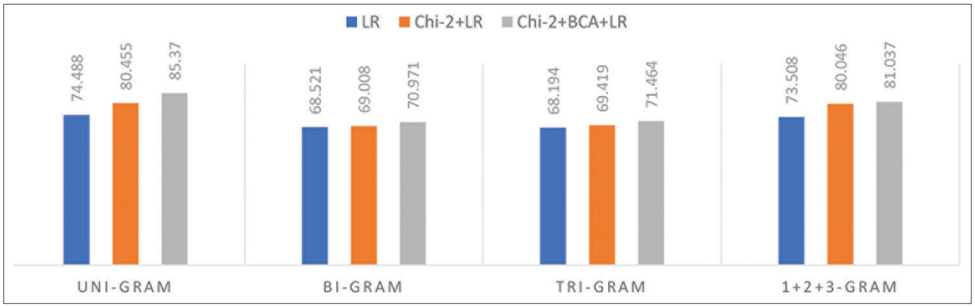


Fig. 5. Accuracies of Amazon dataset.

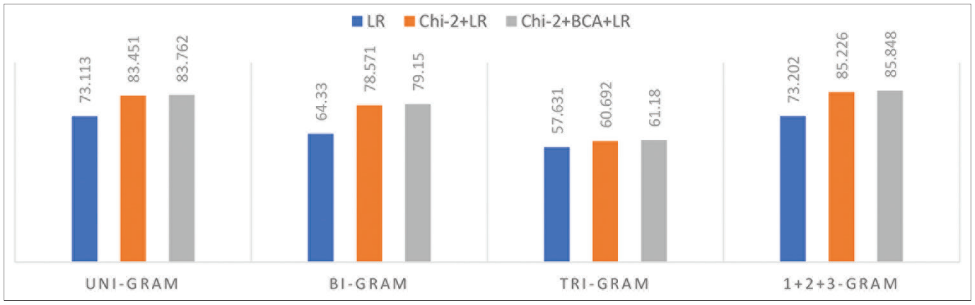


Fig. 6. Accuracies of Trump data set.

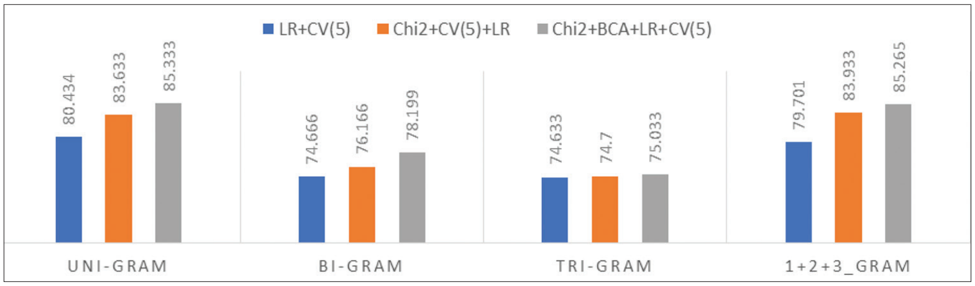


Fig. 7. Accuracies of Chelsea FC data set.

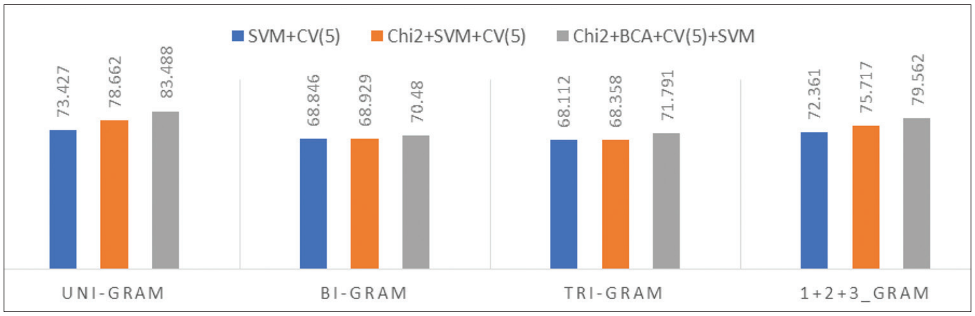


Fig. 8. Accuracies of CR7 data set.

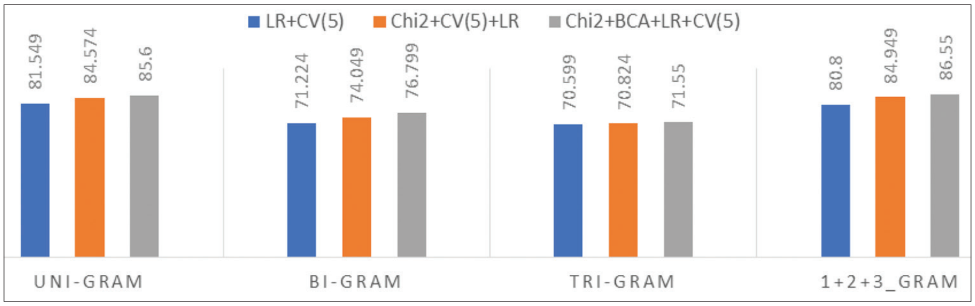


Fig. 9. Accuracies of Amazon dataset.

The graph shows that LR classifier attains best result in unigram range followed by 1 + 2 + 3-g, bigram, and trigram, respectively. However, after applying Chi-2, unigram, and 1 + 2 + 3-g accuracy dramatically increased with more than 5%, followed by bigram and trigram with a slight increase. Finally, applying BCA achieves a dramatical accuracy rate increase by approximately 5% with unigram, followed by the rest three approaches bigram and trigram and 1 + 2 + 3-g with a slight increase.

The graph shows that LR classifier attains best result in unigram range followed by 1 + 2 + 3-g, bigram, and trigram, respectively, which is in all cases less than Chi-2 result, that achieve big raise in accuracy with unigram by 10% bigram

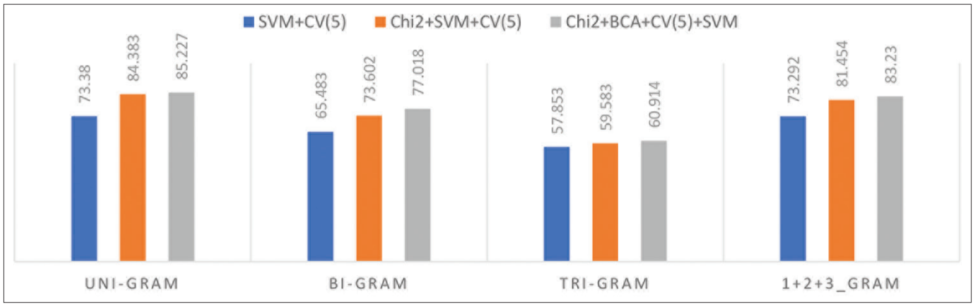


Fig. 10. Accuracies of Trump dataset.

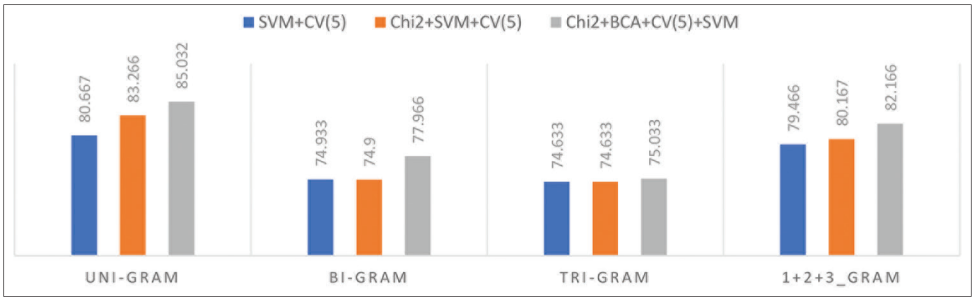


Fig. 11. Accuracies of Chelsea FC dataset.

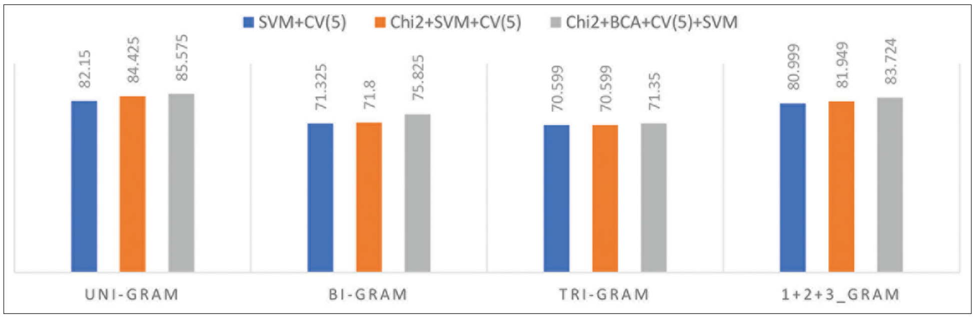


Fig. 12. Accuracies of CR7 dataset.

14% and 1 + 2 + 3-g with 10% followed by trigram with 3% increase. Finally, applying BCA accuracy has a slight raise in all cases with less than 1%.

The graph shows that LR classifier attains best accuracy with unigram range followed by 1 + 2 + 3-g, bigram, and trigram, respectively. After applying Chi-2, unigram and 1 + 2 + 3-g accuracy increased with more than 3%, followed by bigram and trigram with a slight increase. Finally, applying BCA achieves a better accuracy rate increase in all cases. In unigram bigram, and 1 + 2 + 3-g accuracy increased by more than 1.5%, followed by trigram with a small increase.

The graph shows that LR classifier attains best result with 81.549% in unigram range followed by 1 + 2 + 3-g, bigram, and trigram respectively. Consequently, after applying Chi-2, unigram, 1 + 2 + 3-g, and bigram accuracy increased by more than 3%, followed by and trigram with a small increase. Finally, applying BCA achieves even more accuracy rate increased by approximately 1.5% with unigram, bigram, and 1 + 2 + 3-g, and a small change can be observed with trigram.

All bar charts illustrate that the accuracy attained from Hybrid Chi-2 +BCA outperforms the accuracy of LR, LR when applied only with features selected by Chi-2, in all n-gram ranges. Moreover, unigram and (1 + 2 + 3-g) achieve higher results than bigram and trigram in (Chi-2+BCA) feature selection. Results also show that with the growth of datasets, the accuracy of the classifier increases.

The graph shows that SVM classifier attains the best result in unigram range followed by 1 + 2 + 3-g, bigram, and trigram, respectively. Then, after applying Chi-2, unigram, and 1 + 2 + 3-g accuracy dramatically increased with more than 3% and 5%, respectively, followed by bigram and trigram with a slight increase. Finally, applying BCA achieves a dramatical accuracy rate increase by approximately 5% with unigram and (1 + 2 + 3), followed by bigram and trigram with a slight increase.

The graph shows that SVM classifier attains the best result in unigram and 1 + 2 + 3-g, followed by bigram and trigram, respectively. Then, Applying Chi-2, unigram, bigram, and 1 + 2 + 3-g accuracy dramatically increased with more than 10%, 8%, and 7%, respectively, followed by trigram with a slight increase. Finally, applying BCA to features selected by Chi-2 achieves a dramatical accuracy rate increase by approximately 3%, and more than 1% with unigram and (1 + 2 + 3), followed by and trigram with a slight increase.

The graph shows that SVM classifier achieves the best result in unigram range, followed by 1 + 2 + 3-g, bi-gram, and tri-gram, respectively. However, after applying Chi-2, unigram accuracy dramatically increased with more than 3%, followed by 1 + 2 + 3-g, bigram, and trigram with a slight increase. Finally, applying BCA to features selected by Chi-2 achieves a dramatical accuracy increase by approximately 3% with unigram, followed by 1 + 2 + 3, bigram, and trigram with a slight increase.

The graph shows that SVM classifier attains the best result in unigram range, followed by 1 + 2 + 3-g, bigram, and trigram, respectively. However, after applying Chi-2, unigram, and 1 + 2 + 3-g accuracy increased with more than 2% and 1%, respectively, followed by bigram with a slight change, while trigram increase remained same. Finally, applying BCA achieves a dramatical accuracy rate increase by approximately 4% with bigram, 1% with unigram, and more than 2% with 1 + 2 + 3, followed by trigram with a slight increase.

All Bar charts illustrate that the accuracy achieved from Hybrid Chi-2+BCA outperforms the accuracy of SVM, SVM when applied only with features selected by Chi-2, in all n-gram ranges. Moreover, unigram and 1 + 2 + 3-g achieve higher results than bigram and trigram in most cases of (Chi-2+BCA) feature selection. Results also show that with the growth of datasets, the accuracy of the classifier increases, same as with LR.

4. CONCLUSION

In the context of our work, we developed a sentiment classification model for classifying tweets into positive and negative based on the sentiment of the author. As the amount of data becomes huge, the task of classifying them becomes more challenge and the need for reducing the number of features arises to improve classification accuracy. We proposed a hybrid feature selection method by incorporating a filter-based method Chi-2, followed by wrapper-based method BCA for reducing the number of irrelevant features and selecting optimal or sub-optimal features respectively, from features generated by BoW and TF-IDF, each of which used with a different classifier. After training our model with different n-gram ranges and five-fold cross-validation, we conclude that applying our proposed hybrid feature selection method (Chi-2+BCA) reduces features and improves classification performance in the term of accuracy up to 11.847% compared to using original feature set with linear SVM and 10.882% with LR classifiers, both with unigram range. Moreover, the maximum improvement of Chi-2+BCA over using only Chi-2 was 4.915% and 4.826% for LR and SVM, respectively.

4.1. Future Work

Using the same system with a greater number of tweets to inspect the effectiveness of BCA with the growth of the dataset. Using BCA as a feature subset selection algorithm with deep learning algorithms such as LSTM and RNN. Applying BCA to other feature generation techniques such as word2vec or doc2vec. Hybridizing BCA with other filter methods.

REFERENCES

[1] H. P. Patil and M. Atique. "Sentiment Analysis for Social Media: A Survey". 2015 *IEEE 2nd International Conference Information Science Secur*, 2016.

[2] M. K. Das, B. Padhy and B. K. Mishra. "Opinion Mining and Sentiment Classification: A Review". *Proceeding International Conference Inventory System Control*, pp. 4-6.

[3] A. S. Al Shammari. "Real-time Twitter Sentiment Analysis using 3-way classifier". *21st Saudi Computer Society National Computer Conference's*, pp. 1-3, 2018.

[4] R. D. Desai. "Sentiment Analysis of Twitter Data". *Proceeding 2nd International Conference Intelligence Computing Control System* no. Iciccs, pp. 114-117, 2019.

[5] P. M. Mathapati, A. S. Shahapurkar and K. D. Hanabaratti. "Sentiment Analysis using Naïve Bayes Algorithm". *International Journal of Computational Science and Engineering*, vol. 5, no. 7, pp. 75-77, 2017.

[6] N. Krishnaveni and V. Radha. "Feature Selection Algorithms for Data Mining Classification: A Survey". *Indian Journal of Science and Technology*, vol. 12, no. 6, pp. 1-11, 2019.

[7] Y. Zhai, W. Song, X. Liu, L. Liu and X. Zhao. "A Chi-square Statistics Based Feature Selection". *2018 IEEE 9th Internatinal Conference Software Engineering Services Science*, pp. 160-163, 2018.

[8] I. Kurniawati and H. F. Pardede. "Hybrid Method of Information Gain and Particle Swarm Optimization for Selection of Features of SVM-Based Sentiment Analysis". *2018 Internatinal Conference Information Technology System innovation*, pp. 1-5, 2019.

[9] S. Kaur, G. Sikka and L. K. Awasthi. "Sentiment Analysis Approach Based on N-gram and KNN Classifier". *ICSCCC 2018 1st International Conference Security Cyber Computer communication*, pp. 13-16, 2019.

[10] X. Zhang and X. Zheng. "Comparison of Text Sentiment Analysis Based on Machine Learning". *Proceeding 15th Internatioanl Symposium Parallel Distributed Computing ISPDC 2016*, pp. 230-233, 2017.

[11] R. Joshi and R. Tekchandani. "Comparative Analysis of Twitter Data Using Supervised Classifiers". *Proceeding International Conference Invention Computer Technology ICICT 2016*, vol. 2016, 2016.

[12] M. Luo and L. Luo. "Feature Selection for Text Classification Using OR+SVM-RFE". *2010 Chinese Control Decision Conference CCDC 2010*, pp. 1648-1652, 2010.

[13] R. Maipradit, H. Hata and K. Matsumoto. "Sentiment classification using N-gram IDF and automated machine learning". *IEEE Software*, vol. 7459, pp. 10-13, 2019.

[14] S. Rai, S. M. Shetty and P. Rai. "Sentiment Analysis of Movie Reviews using Machine Learning Classifiers". *International Journal of Computer Applications*, vol. 182, no. 50, pp. 25-28, 2019.

[15] S. Naz, A. Sharan and N. Malik. "Sentiment Classification on Twitter Data Using Support Vector Machine". *Proceeding 2018 IEEE/WIC/ACM International Confernce Web Intell. WI 2018*, pp. 676-679, 2019.

[16] R. Wagh and P. Punde. "Survey on Sentiment Analysis Using Twitter Dataset". *Proceeding 2nd International Conference electronic communications Aerospace Technology ICECA 2018*, No. Iceca, pp. 208-211, 2018.

[17] N. Iqbal, A. M. Chowdhury and T. Ahsan. "Enhancing the Performance of Sentiment Analysis by Using Different Feature Combinations". *International Conference Compututing Communication IC4ME2 2018*, pp. 1-4, 2018.

[18] A. Rane and A. Kumar. "Sentiment Classification System of Twitter Data for US Airline Service Analysis". *Proceeding International Computing Software APPL Conference*, vol. 1, pp. 769-773, 2018.

[19] A. Jovi, K. Brki and N. Bogunovi. "A Review of Feature Selection Methods with Applications". *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics*, pp. 25-29, 2015.

[20] S. Rana and A. Singh. "Comparative Analysis of Sentiment Orientation Using SVM and Naive Bayes Techniques". *Proceeding 2016 2nd Interenational Confernce Next General Computer Technologies*, *2016*, pp. 106-111, 2017.

[21] K. L. S. Kumar, J. Desai and J. Majumdar. "Opinion Mining and Sentiment Analysis on Online Customer Review". *2016 IEEE Interenatioanl Conference Computing Intelligence computing Research ICCIC 2016*, 2017.

[22] F. Iqbal, J. Maqbool, B. C. M. Fung, R. Batool, A. M. Khaytak, S. Aleem and P. C. K. Hung. "A hybrid framework for sentiment analysis using genetic algorithm based feature reduction". *IEEE Access*, vol. 7, pp. 14637-14652, 2019.

[23] A. Zarshenas and K. Suzuki. "Binary coordinate ascent: An efficient optimization technique for feature subset selection for machine learning". *Knowledge-Based Systems*, vol. 110, pp. 191-201, 2016.

ORIGINAL ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Effect of Benzalkonium Chloride on Properties of Zinc Oxide Nanoparticles Synthesized through Sol-Gel Technique

Hwda Ghafur Rauf[1], Madzlan Aziz[2], Sattar Ibrahim Kareem[1]

[1]Department of Medical Laboratory Sciences, University of Human Development, Sulaymaniyah, Kurdistan Region, Iraq,
[2]Department of Chemistry, University Technology Malaysia, Johor Bahru, Malaysia

ABSTRACT

In the present study, to synthesize controllable sized metal oxide particles, benzalkonium chloride (BAK) as cationic surfactant was added to zinc oxide (ZnO) nanostructures synthesis at room temperature using sol–gel method. The effect of cationic surfactant BAK concentrations, on the optical properties, size, and morphology of ZnO nanoparticles synthesized through sol–gel method was studied. The characterization of ZnO nanostructures was occurred using transmission electron microscopy (TEM), X-ray diffraction (XRD), ultraviolet–visible near infrared (UV-Vis) spectrophotometer, Fourier transform infrared spectroscopy (FTIR), and scanning electron microscopy (SEM). ZnO nanostructures shape and size were revealed by SEM and TEM. The hexagonal (wurtzite structure) of ZnO was confirmed by an X-ray diffractogram. The bandgap energy of the prepared ZnO samples was determined by UV-Vis spectrophotometer. FTIR analyzed the presence of functional groups.

Index Terms: Benzalkonium Chloride, Cationic Surfactants, Sol–Gel Method, Zinc Oxide, Nanostructures

1. INTRODUCTION

In an experimentally accessible size range, one of the few oxides that show quantum confinement effects is zinc oxide (ZnO), which is a semiconductor with a wide bandgap of 3.4 eV and a large exciton binding energy of 60 MeV at room temperature. What makes ZnO favorable is its qualities such as low-temperature process, transparency, and high carrier mobility in addition to cost savings. As well as ZnO properties, size-dependent optical absorption is also a valuable tool. Therefore, to quantum size ZnO particles, sol–gel preparation method can be used [1]. In many particle synthetic processes, especially surfactant-free chemical reactions, as particles start to generate, aggregation immediately occurs. This agglomeration for nanostructured materials does cause a problem in the many chemical and pharmaceutical products. The direct mutual attraction between particles through chemical bonding or Van der Waals forces, simply leads to aggregation as represented in conventional studies [2]. Although this agglomeration can be counted onto prevent toxicity of nanoparticles, especially in cosmetic products as ZnO is widely used such as sunscreen or industrial factories like paint factories [3]. Nanoparticles recently considered to be dangerous, both medically and environmentally [4]. In fact, it has the ability to make the particles very reactive or catalytic because it has a high surface [5]. They are also capable of harming cell membranes in organisms and biological systems [6]. However, if the particles are larger in size, they are unlikely to harm the cell nucleus and other internal cellular components due to the particle agglomeration [7].

Access this article online

DOI: 10.21928/uhdjst.v4n1y2020pp41-50	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: Hwda Ghafur Rauf, Department of Medical Laboratory Sciences , University of Human Development, Sulaymaniyah, Kurdistan Region of Iraq. E-mail: hwda.rauf@uhd.edu.iq

Received: 01-09-2019 Accepted: 01-02-2020 Published: 15-02-2020

Lung inflammatory and systemic activity can be noticed in experimental animals when they inhaled ultrafine particles unlike an equal mass of larger particles, which may also affect adult human subjects [8].

Due to its bulky bi-tailed structure, which is insoluble, benzalkonium chloride (BAK) particles tend to agglomerate, which caused the increase in the particle size [9].

Several physical and chemical methods could be taken under consideration to synthesize ZnO nanostructures. Chemical vapor deposition (CVD), metal-organic CVD and molecular beam epitaxy, pyrolysis, vapor-liquid-solid growth, and vapor-solid processes such as thermal reduction, are considered physical methods. At the same time, to have a very large scale production, there are chemical methods that are very simple yet effective such as precipitation, sol–gel, and solvothermal processes. Unlike general methods, these methods do not require high temperatures and sophisticated instruments [10]. In this work, there must be a controllable grain size of particles as a result, to approach that goal, it's better to use the sol–gel method, especially when we adjust the experimental conditions such as concentration, temperature, pH, and reaction time. In addition, this method is simple yet reproducible and does not cost much, not to mention its reliability of stoichiometry control, which makes it a totally suitable process for industrial production of ZnO [11].

2. EXPERIMENTAL

2.1. Preparation of ZnO Nanoparticles

ZnO nanocrystals were prepared by adding 100.0 mL of 0.5 mol/l NaOH solution dropwise slowly into 250.0 mL of 0.1 mol/l $Zn(Ac)_2$ distilled water using distilled water under vigorous stirring and closed vessel to produce the $Zn(OH)_2$ precipitate, then an appropriate amount of NH_4HCO_3 (0.8 g) powder was added to adjust ph value of the solution. After stirring for 30 min, a semitransparent zinc carbonate hydroxide colloid was obtained. After 30 min, the colloid was centrifuged and dried at 80°C. Thus, the precursor of a small crystallite of $Zn_5(CO_3)_2(OH)_6$ with white color was formed then calcined at 450°C for 2 h. The sample was washed by deionized water, and finally, the product was dried at 70°C to obtain the white colored sample of ZnO nanoparticle.

2.2. Preparations of ZnO Nanoparticles with BAK

The procedure was repeated as in 2.1 with addition BAK (95% purity Sigma-Aldrich) (1×10^{-5} M below, 5×10^{-5} M at, 15×10^{-5} M above) critical micelle concentration (CMC), respectively. The CMC value of BAK was taken from literature data [12]. Although this time, the precursor was calcined at 550°C for 5 h to obtain the sample, as shown in Fig. 1. Finally, the product was dried at 70°C to obtain the white-colored sample ofZnO.

3. RESULTS AND DISCUSSION

The CMC value of BAK was taken from literature data and it was 0.04 mM [12].

3.1. Analysis of Fourier Transformed Infrared (FTIR)

Fig. 2a shows the FTIR spectrum of prepared ZnO nanoparticles without BAK. Furthermore, Fig. 2b-d shows

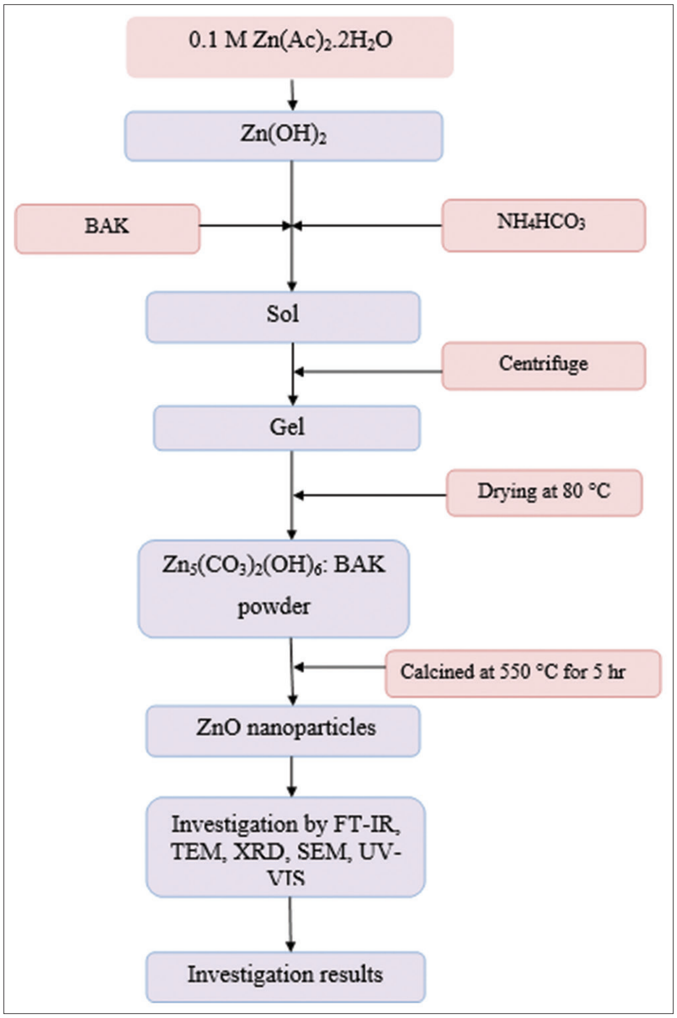


Fig. 1. Preparation of benzalkonium chloride with benzalkonium chloride surfactant.

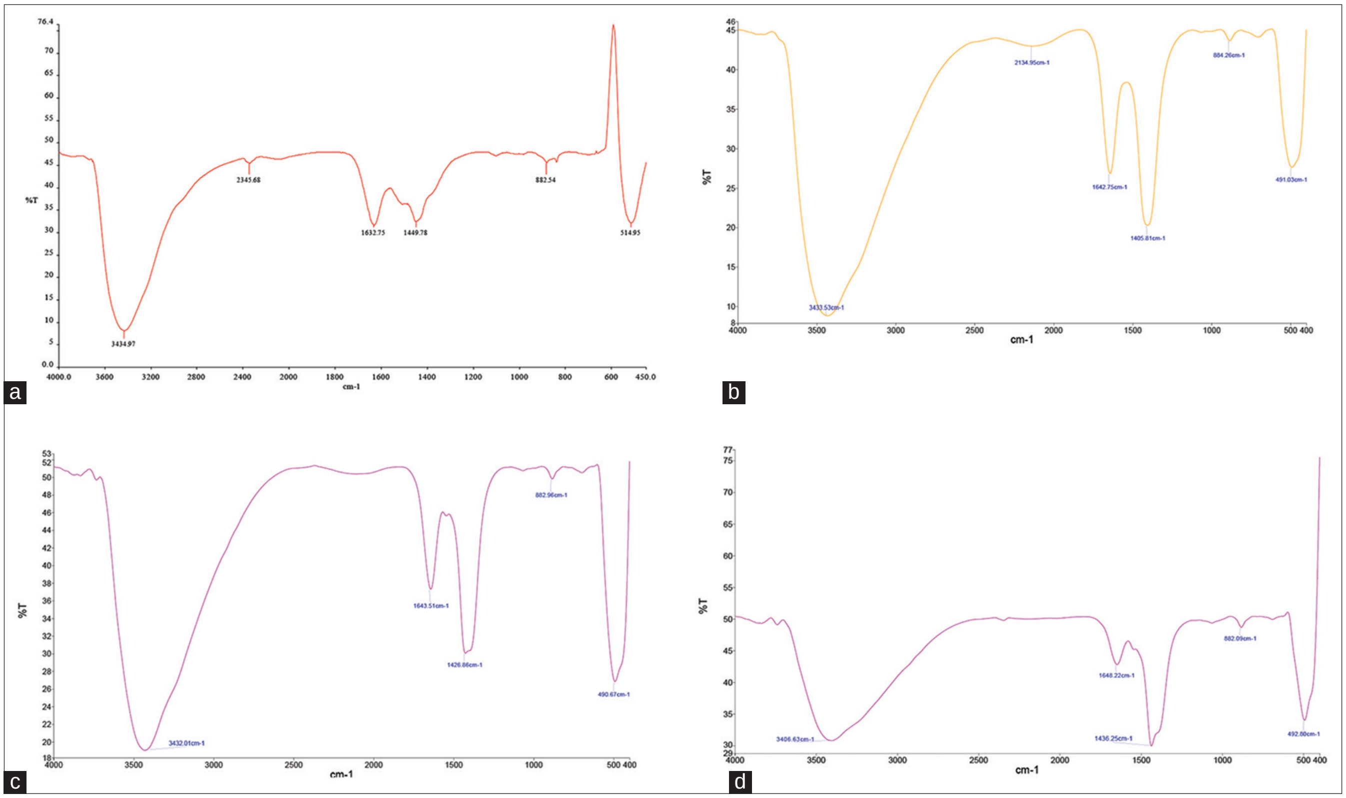


Fig. 2. Combined Fourier transformed infrared spectra of zinc oxide nanoparticles without and with benzalkonium chloride (BAK) surfactant (a) without surfactant, (b) with BAK below critical micelle concentration (CMC), (c) with benzalkonium chloride at CMC, (d) with BAK above CMC.

the FT-IR spectra of prepared ZnO particles in the presence of BAK (below CMC, at CMC, and after CMC), respectively. The broad absorption band centered above 3000 cm^{-1} is attributable to the band O–H stretching vibrations of water molecules on ZnO, and the band near 1630 cm−1 is assigned to the covalent bond in H–O–H bending vibrations mode which were also presented due to the adsorption of humidity in the air when FT-IR sample disks were prepared in an open air.

The spectrums near 1450 cm^{-1} indicate the existence of C-O. However, for ZnO samples with BAK surfactant, the calcination temperature was raised at 550°C for 5 h and is not blue shifted. The difference in the calcination temperatures and the time could explain the difference among the figures.

3.2. Analysis of Scanning Electron Microscope (SEM)

The morphology of the prepared ZnO samples was determined by SEM analysis, as shown in Figs. 3-6. ZnO samples with BAK that is calcined at 550°C for 5 h, do not reveal uniform morphology. The structure of BAK is bulky and bi-tailed which is insoluble; the particles tend to agglomerate, which caused the increase in the particle size [9]. The particle size of ZnO sample was taken by line intersecting method; the result revealed that the prepared ZnO sample without BAK in Fig. 3 was in the range of 80–85 nm which smaller than the prepared samples in the presence of BAK Figs. 4-6, respectively.

Although SEM is not an accurate technique to determine the particle size, according to particle size distributions of ZnO nanoparticles based on SEM images, ZnO particles without surfactant reveal an average size of 86 nm, as shown in Fig. 7. However, the results for ZnO samples with BAK revealed a bigger average size of 125 nm for ZnO with BAK below CMC point, 145 nm for ZnO with BAK at CMC point, and 212 nm for ZnO with BAK above CMC point. These samples are not nanosized, as shown in Figs. 8-10. The actual reason is due to BAK's bulky structure which contains 25 carbon atoms in the tail that cause the agglomeration; as a result of this led to an increase in the size of the particles [9].

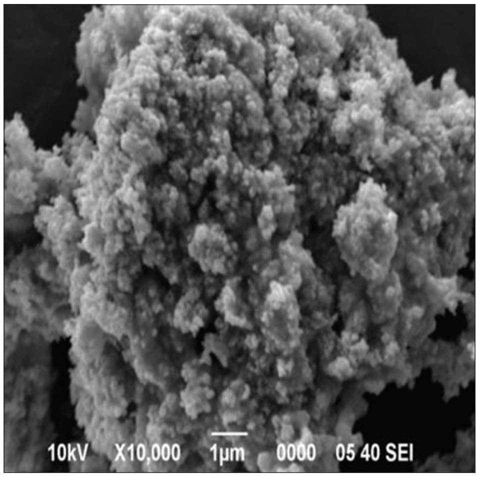


Fig. 3. SEM analysis of ZnO without surfactant.

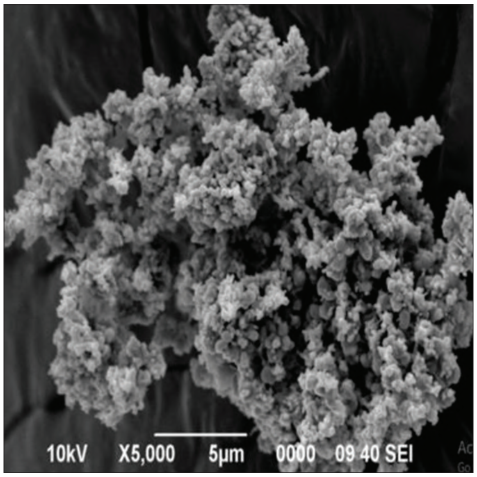


Fig. 4. SEM analysis of ZnO with benzalkonium chloride surfactant below critical micelle concentration.

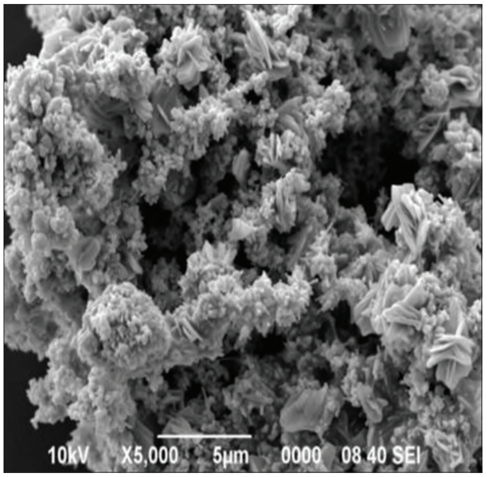


Fig. 5. SEM analysis of ZnO with benzalkonium chloride surfactant at critical micelle concentration.

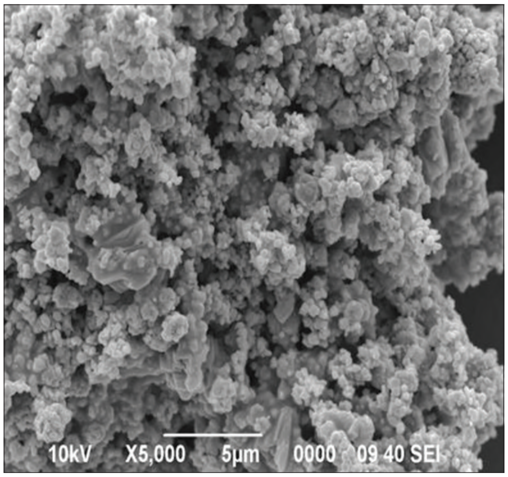


Fig. 6. SEM analysis of ZnO with benzalkonium chloride surfactant above critical micelle concentration.

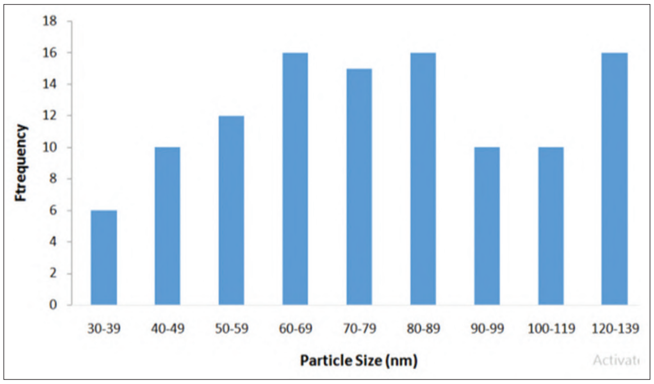


Fig. 7. The particle size distribution of ZnO without surfactant.

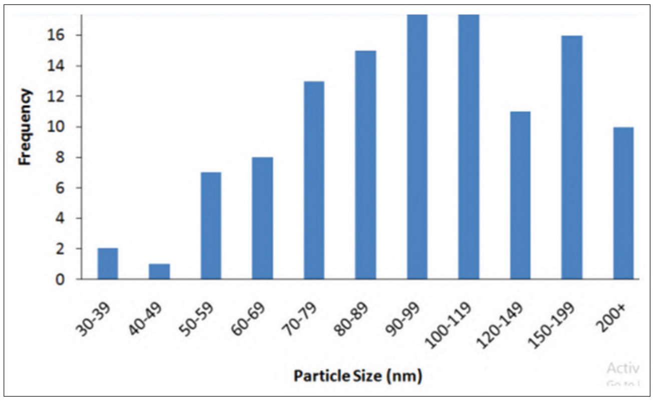


Fig. 8. The particle size distribution of ZnO with benzalkonium chloride surfactant below critical micelle concentration.

3.3. Analysis of Ultraviolet–Visible Near Infrared (UV-Vis-NIR) Spectroscopy

Figs. 11-14 illustrate the $\alpha h\upsilon^2$ versus $h\upsilon$ plot used for the estimation of the bandgap of ZnO nanoparticles calcined

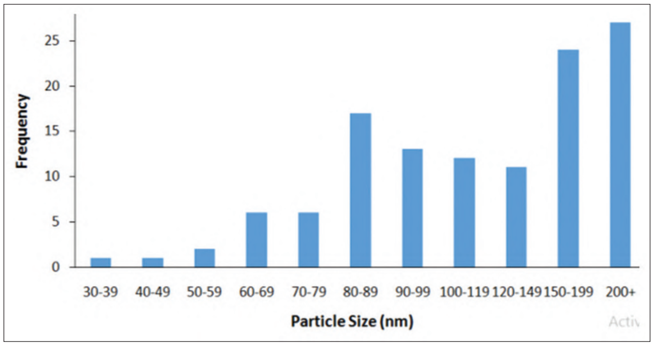


Fig. 9. The particle size distribution of ZnO with benzalkonium chloride surfactant at critical micelle concentration.

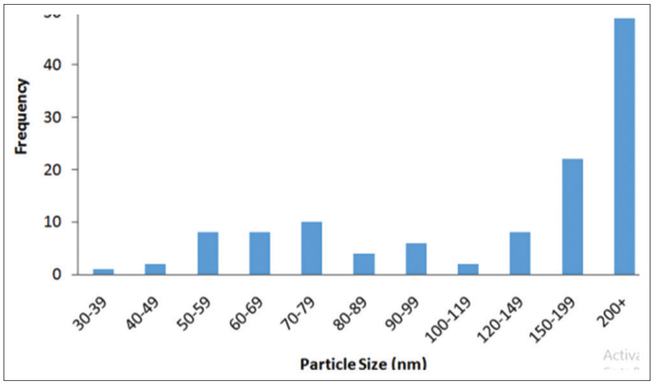


Fig. 10. The particle size distribution of ZnO with benzalkonium chloride surfactant above critical micelle concentration.

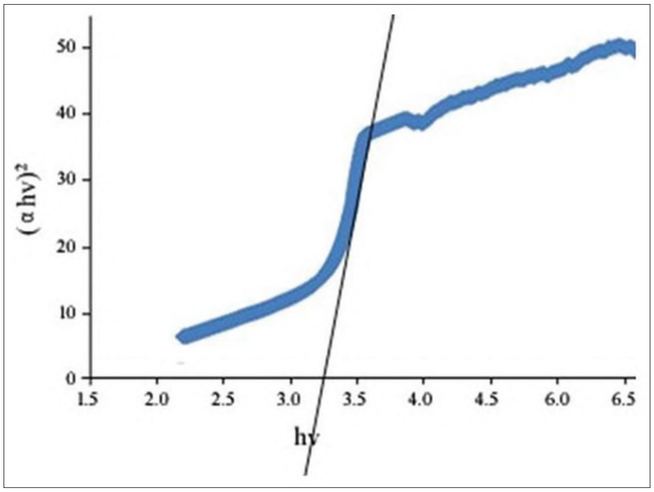


Fig. 11. Analysis of (UVVis-NIR) Spectroscopy of ZnO without surfactant.

at 550°C by extrapolating the graph to X-axis so as to calculate the bandgap of the samples. The bandgap is found to be 3.26 eV, 3.27 eV, 3.24 eV, and 3.21 eV for the samples prepared without surfactant and BAK individually at different concentrations (below, at, and after), respectively.

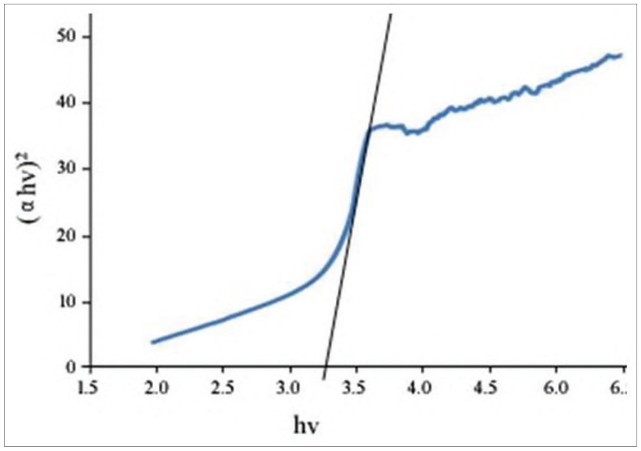


Fig. 12. Analysis of (UVVis-NIR) Spectroscopy of ZnO with benzalkonium chloride below critical micelle concentration.

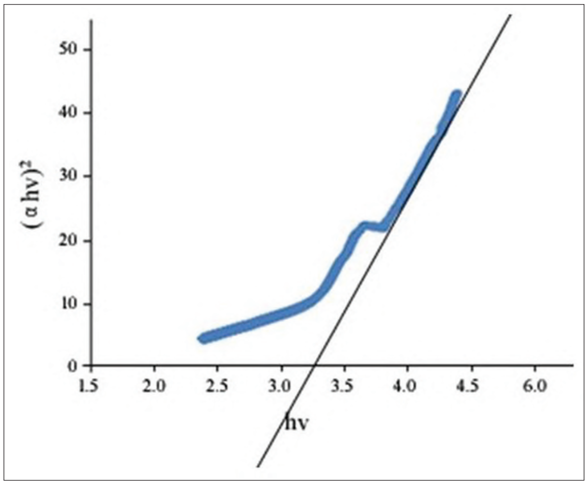


Fig. 13. Analysis of (UVVis-NIR) Spectroscopy of ZnO with benzalkonium chloride at critical micelle concentration.

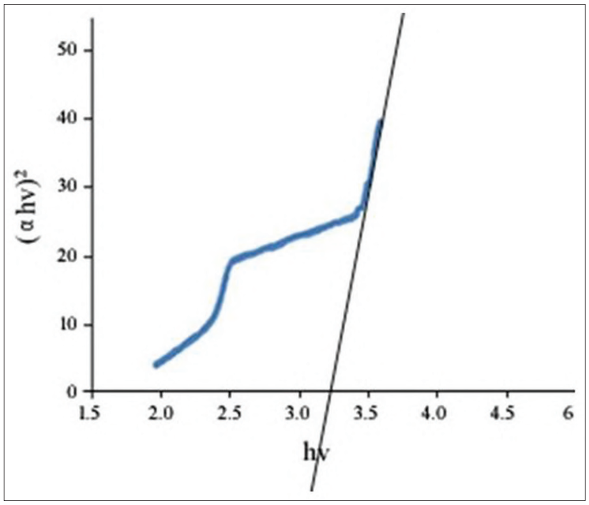


Fig. 14. Analysis of (UVVis-NIR) Spectroscopy of ZnO with benzalkonium chloride above critical micelle concentration.

The bandgap decreases in the presence of BAK surfactant because of its bulky structure which contains 25 carbon atoms in the tails that cause agglomeration [9]. Although the results show two curves, which apparently means impurities presence.

Table 1 reveals a comparison of bandgap energy for each sample:

3.4. Analysis of X-ray Diffractometer (XRD)

Fig. 15 shows the XRD patterns of the ZnO nanoparticles calcined at 450°C for 2 h. A study of standard data JCPDS 76-0704 confirms that all the synthesized materials are hexagonal ZnO phase (wurtzite structure). The XRD pattern for the ZnO nanoparticle prepared without surfactant calcined at 450°C for 2 h is shown in Fig. 15. The three highest spectrums were assigned as 100, 002, and 101 matched planes for 2ϴ values of 31.769, 34.422, and 36.254°, respectively. The calculated average size of the most intense three diffraction spectrums was 76.9 nm.

The XRD pattern of the ZnO nanoparticle prepared with BAK surfactant below CMC calcined at 550°C for 5 h is shown in Fig. 16. The three highest spectrums were assigned as (100), (002), and (101) matched planes for 2ϴ values of 31.769, 34.422, and 36.254°, respectively. The structure of synthesized material is hexagonal ZnO (wurtzite structure), according to standard data JCPDS 76-0704. The calculated average size of the most intense three diffraction spectrums was 116.4 nm. Moreover, the structure was hexagonal as the others. On the other hand, Fig. 17 shows the XRD pattern of the ZnO nanoparticle prepared with BAK surfactant at CMC calcined at 550°C for 5 h. The three highest spectrums were assigned as (100), (002), and (101) matched planes for 2ϴ values of 31.769, 34.422, and 36.254°, respectively. The calculated average size of the most intense three diffraction spectrums was 130.0 nm.

TABLE 1: Bandgap energy for each sample.

Status	Bandgap for ZnO: Benzalkonium chloride
Below CMC	3.27 eV
At CMC	3.24 eV
Above CMC	3.21 eV

CMC: Critical micelle concentration, ZnO: Zinc oxide

The XRD pattern for the ZnO nanoparticle prepared with BAK surfactant above CMC calcined at 550°C for 5 h is shown in Fig. 18. The three highest spectrums were assigned as (100), (002), and (101) matched planes for 2ϴ values of 31.769, 34.422, and 36.254°, respectively. The calculated average size of the most intense three diffraction spectrums was 240.0 nm. As its obvious, the particle size of ZnO samples in the presence of BAK increased due to the bulky and hydrophobic bi-tailed structure of BAK. The tail of BAK contains 25 carbon atoms, which cause agglomeration and lead to increase in the particle size [9].

3.5. Analysis of Transmission Electron Microscopy (TEM)

TEM was used to determine the average particle size. The results that obtained by TEM image and TEM particle

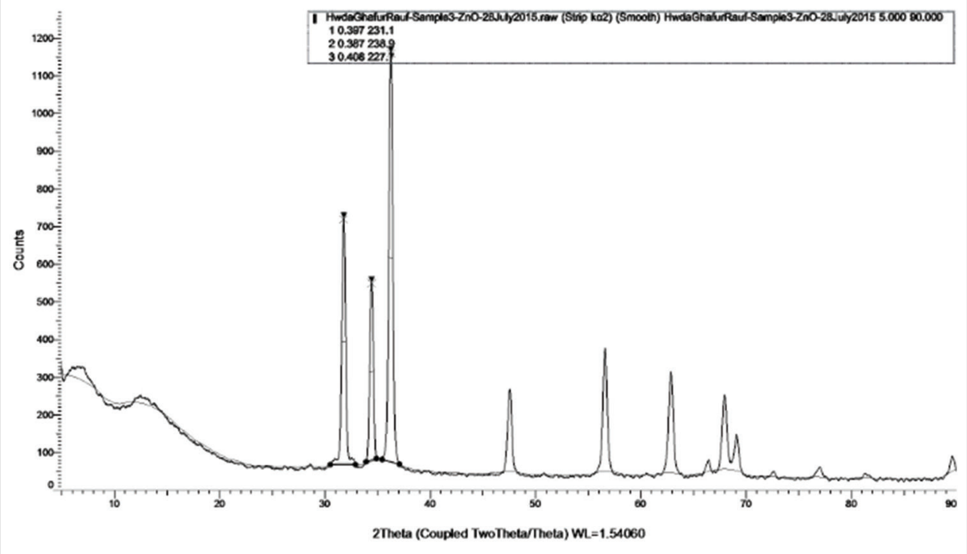


Fig. 15. XRD diffractogram of zinc oxide nanoparticles without surfactant.

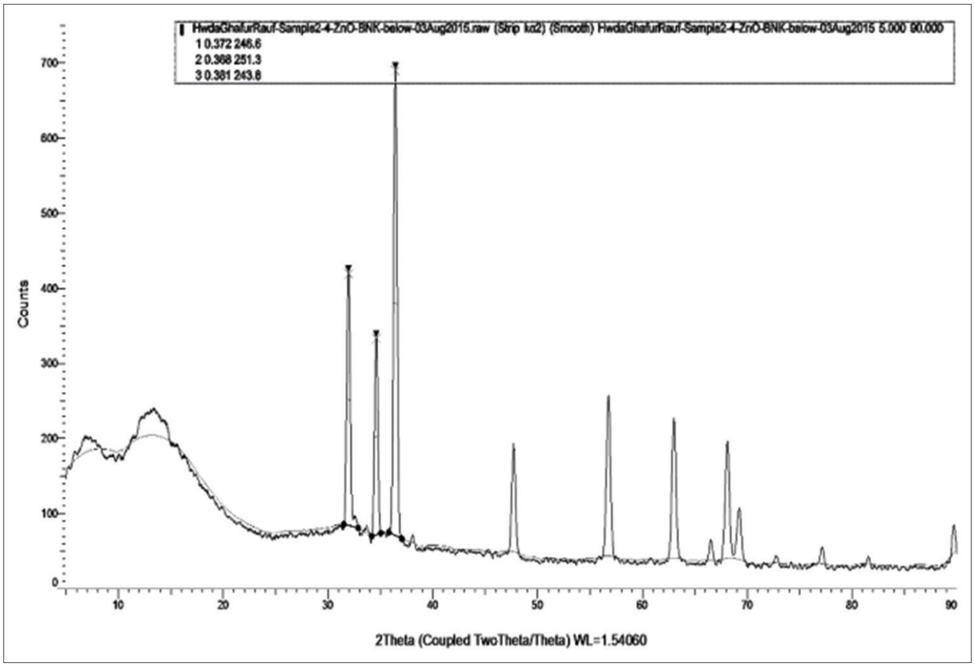


Fig. 16. XRD diffractogram of zinc oxide nanoparticles with benzalkonium chloride Surfactant below critical micelle concentration.

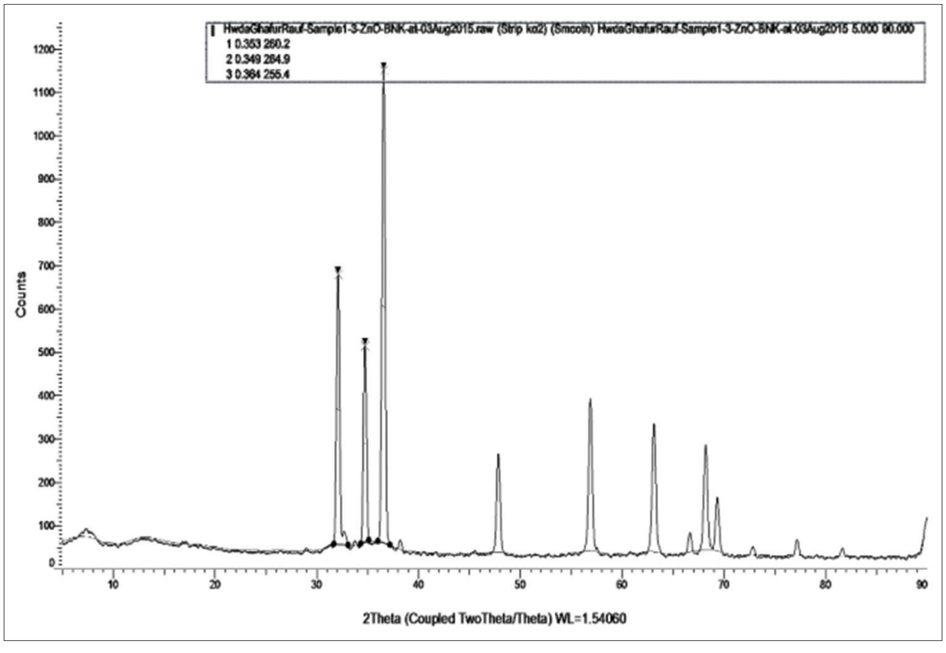


Fig. 17. XRD diffractogram of zinc oxide nanoparticles with benzalkonium chloride surfactant at critical micelle concentration.

size distribution of the ZnO sample without surfactant were 66 nm, 100 nm for ZnO with BAK below CMC, and 140 nm for ZnO with BAK at CMC, as shown in Figs. 19-24, respectively. As it is been mentioned in the previous investigations, the particle size of ZnO samples in the presence of BAK increased due to the bulky and bi-tailed structure of BAK. The tail of BAK contains 25 carbon atoms, which causes agglomeration and lead to an increase in the particle size [9].

The calculated particle sizes for each sample from XRD, TEM, and SEM are depicted in Table 2.

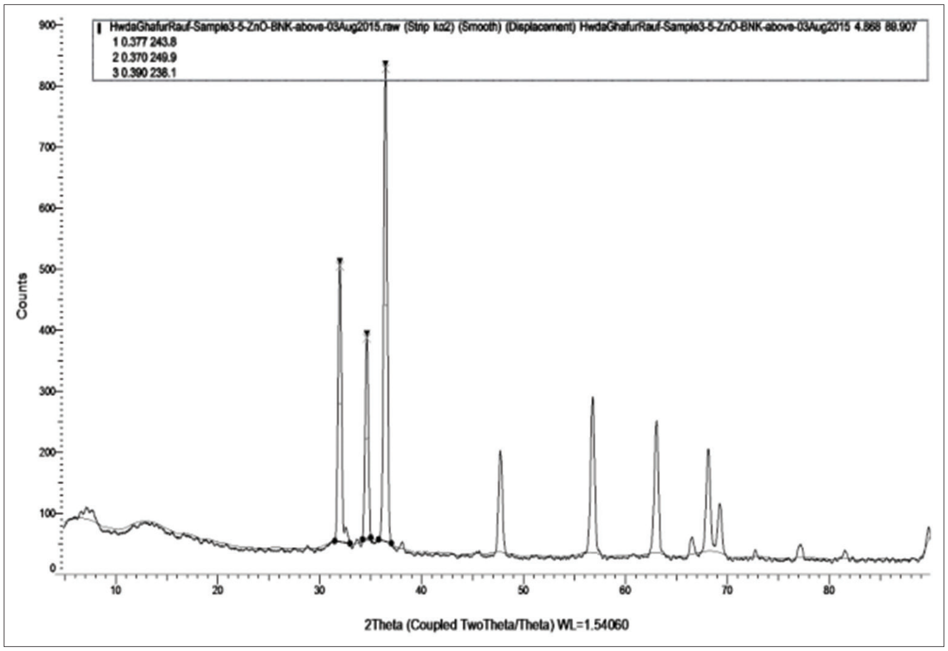


Fig. 18. XRD diffractogram of zinc oxide nanoparticles with benzalkonium chloride surfactant above critical micelle concentration.

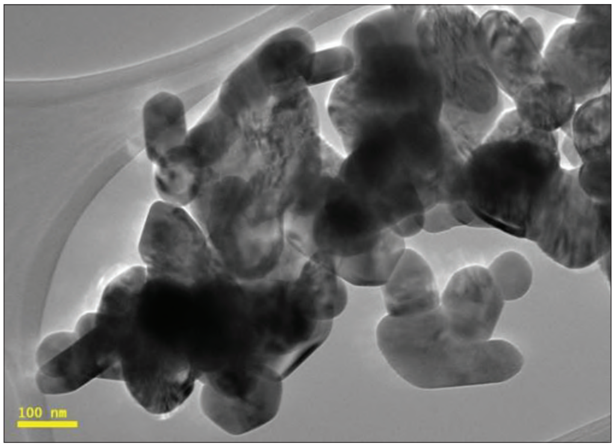


Fig. 19. TEM analysis of ZnO without surfactant.

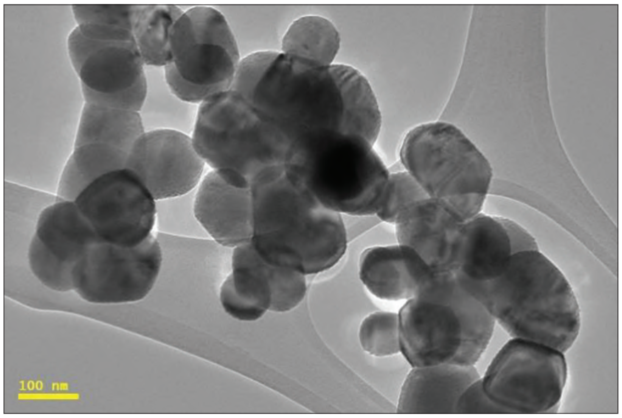


Fig. 21. TEM analysis of ZnO with benzalkonium chloride below critical micelle concentration.

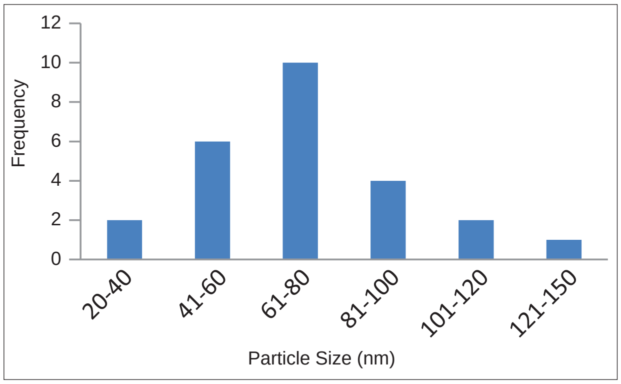


Fig. 20. Particle size distribution of ZnO based on TEM without surfactant.

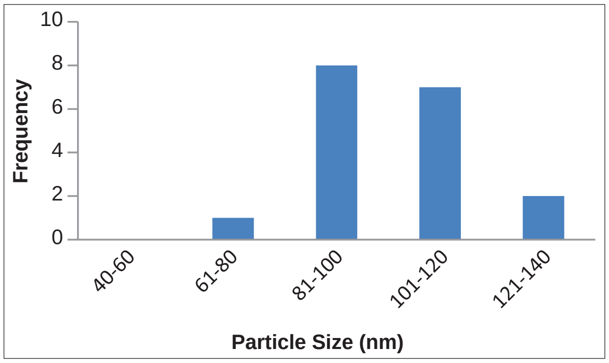


Fig. 22. Particle size distribution of ZnO based on TEM with benzalkonium chloride below critical micelle concentration.

TABLE 2: Calculated particle sizes for each sample from XRD, TEM, and SEM in different conditions.

Status	Temperature°C	Average particle size from XRD nm	Average particle size from TEM nm	Average particle size from SEM nm
ZnO without surfactant	500	76.9	66	87
ZnO with BAK below CMC	500	116.4	100	125
ZnO with BAK at CMC	500	130	140	145
ZnO with BAK above CMC BAK	500	240	-	212

BAK: Benzalkonium chloride, ZnO: Zinc oxide, CMC: Critical micelle concentration, TEM: Transmission electron microscopy, SEM: Scanning electron microscope

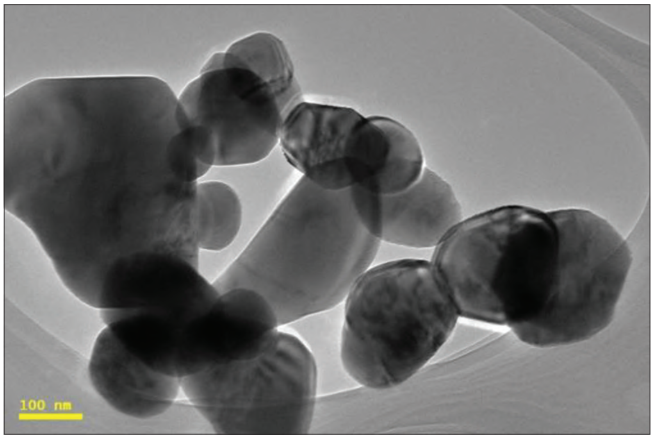


Fig. 23. TEM analysis of ZnO with benzalkonium chloride at critical micelle concentration.

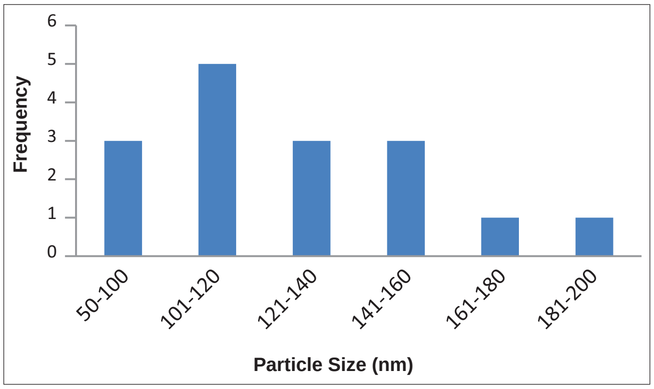


Fig. 24. Particle size distribution of ZnO based on TEM with benzalkonium chloride at critical micelle concentration.

4. CONCLUSION

In this paper, sol–gel method was used to synthesize ZnO with controllable particles. Afterward, three samples of ZnO nanoparticles were synthesized with and without surfactant, to study the influence of BAK as cationic surfactants on properties of ZnO nanoparticles such as particle size, morphology, and bandgap energy [2]. Although this agglomeration can be counted on to prevent toxicity of nanoparticles, especially in cosmetic products as zinc oxide is widely used such as sunscreen or industrial factories like paint factories [3]. Nanoparticles recently conceded to be dangerous, both medically and environmentally [4]. In fact, It has the ability to make the particles very reactive or catalytic because it has a high surface [5]. They are also capable of harming cell membranes in organisms and biological systems [6]. However, if the particles are larger in size, they are unlikely to harm the cell nucleus and other internal cellular components due to the particle agglomeration [7].

Lung inflammatory and systemic activity can be noticed in experimental animals when they inhaled ultrafine particles unlike an equal mass of larger particles, which they may also affect adult human subjects [8].

TEM, SEM, FTIR spectroscopy, X-ray diffraction (XRD), and UV-visible spectroscopy were used to characterize the structure, morphology, and size of the synthesized ZnO nanoparticles. The surfactants create their own interface and form micelles at CMC [13]. Under these conditions, there would be micelle formation in the bulk phase and surfactant-coated nanoparticles formed start to decrease by increasing surfactant concentration beyond the CMC point [14].

Moreover, with the addition of BAK surfactant, the particle size of all the samples starts to increase due to the bulky hydrophobic part and carbon number in the structure of BAK which is for both tails is 25, which causes agglomeration [9].

The XRD patterns of the ZnO nanoparticles; based on the study of standard data JCPDS 76-0704 confirm that all the samples of synthesized ZnO are hexagonal (wurtzite structure). On the other hand, UV-Vis-NIR result revealed that the bandgap value for ZnO without any surfactant was 3.26 eV while with BAK at the same point was 3.24 eV.

5. ACKNOWLEDGMENTS

The authors are thankful to Prof. Madzlan University Technology Malaysia for his review and guidance.

6. FUNDING

This work is not funded by anyone.

7. COMPETING INTEREST

The authors declare that they have no competing interest.

REFERENCES

[1] E. A. Meulenkamp. "Synthesis and growth of ZnO nanoparticles". *The Journal of Physical Chemistry B*, vol. 102, no. 29, pp. 5566-5572, 1998.

[2] D. Li and R. B. Kaner. "Shape and aggregation control of nanoparticles". *Journal of the American Chemical Society*, vol. 128, no. 3, pp. 968-975, 2006.

[3] M. J. Osmond and M. J. McCall. "Zinc oxide nanoparticles in modern sunscreens: an analysis of potential exposure and hazard". *Nanotoxicology*, vol. 4, no. 1, pp. 15-41, 2010.

[4] M. A. Zoroddu, S. Medici, A. Ledda, V. M. Nurchi, J. Lachowicz and M. Peana. "Toxicity of nanoparticles". *Current Medicinal Chemistry*, vol. 21, no. 33, pp. 3837-3853, 2014.

[5] Y. Jackie. *Nanostructured Materials*. Academic Press, New York, 2001.

[6] M. W. Frampton. "Systemic and cardiovascular effects of airway injury and inflammation: Ultrafine particle exposure in humans". *Environmental Health Perspectives*, vol. 109, pp. 529-532, 2001.

[7] C. Greulich, J. Diendorf, T. Simon, G. Eggeler, M. Epple and M. Köller. "Uptake and intracellular distribution of silver nanopahrticles in human mesenchymal stem cells". *Acta Biomaterialia*, vol. 7, no. 1, pp. 347-354, 2011.

[8] G. Oberdörster, J. Ferin, S. Soderholm, R. Gelein, C. Cox, R. Baggs and P. Morrow. "Increased pulmonary toxicity of inhaled ultrafine particles". *Annals of Occupational Hygiene*, vol. 38, pp. 295-302, 1994.

[9] R. Mandavi, R., K. S. Santosh and N. Rathore. "Critical micelle concentration of surfactant, mixed surfactant and polymer by different method at room temperature and its importance". *Oriental Journal of Chemistry*, vol. 24, no. 2. pp. 559-564, 2008.

[10] M. Gusatti, G. Barroso, J. Rosário and C. E. Campos. "Synthesis of ZnO nanostructures in low reaction temperature". *Chemical Engineering Transactions*, vol. 17, pp. 1017-1021, 2009.

[11] D. J. Sornalatha and P. Murugakoothan. "Room temperature synthesis of ZnO nanostructures using CTAB assisted sol-gel method for application in solar cells". International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 9, pp. 414-418, 2013.

[12] Y. Gargouri, R. Julien, A. Bois, R. Verger and L. Sarda. "Studies on the detergent inhibition of pancreatic lipase activity". *Journal of Lipid Research*, vol. 24, no. 10, pp. 1336-1342, 1983.

[13] M. A. Farrukh, P. Tan and R. Adnan. "Influence of reaction parameters on the synthesis of surfactant-assisted tin oxide nanoparticles". *Turkish Journal of Chemistry*, vol. 36, no. 2, pp. 303-314, 2012.

[14] E. Y. Bryleva, N. Vodolazkaya, N. Mchedlov-Petrossyan, L, Samokhina, N. Matveevskaya and A. Tolmachev. "Interfacial properties of cetyltrimethylammonium-coated SiO_2 nanoparticles in aqueous media as studied by using different indicator dyes". *Journal of Colloid and Interface Science*, vol. 316, no. 2, pp. 712-722, 2007.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Digital Medical Image Segmentation Using Fuzzy C-Means Clustering

Bakhtyar Ahmed Mohammed[1], **Muzhir Shaban Al-Ani**[2]

[1]Department of Computer Science, University of Human Development, College of Science and Technology, Sulaymaniyah, KRG, Iraq, [2]Department of Information Technology, University of Human Development, College of Science and Technology, Sulaymaniyah, KRG, Iraq

ABSTRACT

In the modern globe, digital medical image processing is a major branch to study in the fields of medical and information technology. Every medical field relies on digital medical imaging in diagnosis for most of their cases. One of the major components of medical image analysis is medical image segmentation. Medical image segmentation participates in the diagnosis process, and it aids the processes of other medical image components to increase the accuracy. In unsupervised methods, fuzzy c-means (FCM) clustering is the most accurate method for image segmentation, and it can be smooth and bear desirable outcomes. The intention of this study is to establish a strong systematic way to segment complicate medical image cases depend on the proposed method to share in the decision-making process. This study mentions medical image modalities and illustrates the steps of the FCM clustering method mathematically with example. It segments magnetic resonance imaging (MRI) of the brain to separate tumor inside the brain MRI according to four statuses.

Index Terms: Medical image, Medical image modality, Segmentation, Fuzzy C-means clustering

1. INTRODUCTION

The interest is shown by medical professionals in deepening their knowledge of internal anatomy plays an essential part in the importance of medical images that are used in both treatment and diagnosis [1]. Numerous methods of diagnostic medical imaging have been created dependent on different types of electromagnetic band imaging which includes gamma-ray and X-ray imaging, cross-sectional pictures, such as computed tomography (CT), single-photon emission CT, positron emission tomography, magnetic resonance imaging (MRI), or ultrasound [1,2]. These various applications and techniques in medical image processing rely on different ranges of electromagnetic spectrum bands [2]. Improvements in technology caused to increase size and volume of medical imaging, also these developments raise demand on automated diagnosis with computer technology developments, also it decreases the cost and time [3].

Image segmentation can be described as; trying to find homogenous limits inside an image and after that the classification of them, and can be considered as the most significant field of medical image processing, also it allows images to be divided into relevant areas according to homogeneity or heterogeneity criteria, and it is an automatic or semi-automatic process used for separating the region of interest (ROI), also there are numerous medical applications that use to differentiate in the segmentation of body organs and tissues, these include cardiology image analysis, breast tumor detection, autoclassification in hematology field, brain cognitive development, mass segmentation, mass detection,

Access this article online

DOI:10.21928/uhdjst.v4n1y2020.pp51-58	E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Bakhtyar Ahmed Mohammed, Department of Computer Science, University of Human Development, College of Science and Technology, Sulaymaniyah, KRG, Iraq. E-mail: bakhtyar.mohammed@uhd.edu.iq

Received: 30-12-2019 Accepted: 23-02-2020 Publishing: 27-02-2020

surgery simulations, plan of surgery, and detection of vessel boundary in coronary angiograms [1,4]. Techniques of image segmentation can be characterized according to these essential terminologies, such as oriented of pixel, color, region, model, and hybrid [4]. Intelligent decision support systems commonly use the prevailing image segmentation to accurately organize image pixels [5]. The procedure divides the picture into systematic and defined sectors depending on their similarities [5]. Image segmentation is one main component of analysis processes which use in these techniques; remote sensing, computer vision, medical image processing, and geographical information system [6]. Image segmentation plays a key role in automated object recognition systems in the process of computer vision and medical imaging for the analysis of details, image segmentation enables greater ease in detecting and quantifying abnormalities in anatomical structures, such as the brain and lung [5].

However, the processing of the image segmentation can be affected by improper illumination, noise disturbances, environmental factors, and blurring of images, an important phase toward the automatic segmentation of images is region segmentation since this is the step taken to determine and segment the area of interest [7]. Because of ease to apply fuzzy c-mean (FCM) and its high accuracy, it has become one of the best way for image segmentation [8]. Nevertheless, FCM has inadequacies in confusion acknowledgment; various undertakings are practiced for covering this deficiency, with using the objective work FCM and the use of neighbor pixels despite the pixel and besides pixel division have been used, FCM methodology is used for improving the accuracy in picture division, enlistment work is changed [8].

The best criterion to find the optimum solution for these issues is the method known as FCM clustering which is a clustering method whereby points of data can be designated to more than one group each based on shared correlations, and then tries to identify parallels and relationships within each set, least-squares solutions are employed to identify the ideal location for any point of data which may lie in a space of probability bounded by two or more clusters, also there should be as higher level as possible in the likeness of clusters and as lower-level as possible in differences and fuzzy boundaries are easier to develop from a computational point of view [9]. The purpose of writing this paper is to indicate the importance of the segmentation process in image processing and the mechanism of brain segmenting image using FCMs clustering and how it can find the optimum solution for segmenting images and diagnosis using medical image techniques. Furthermore, there are four major steps in the medical imaging field, which consist of capturing the image, digitalizing it, processing for segmentation and finally extracting important information [9,10].

2. LITERATURE REVIEW

In 2008, Ahmed and Mohamad explored that fuzzy clustering is very important in image segmentation, using parallel length to calculate fuzzy weights [11].

In 2010, Naz *et al.* found that the reason digital image processing developers are innovating this method is the best and most accurate way of diagnosing medical imaging can be found, improvements have been made rapidly with many methodologies being put forward supported by a wide range of literature on taking information from a picture and dividing it into defined areas. However, constraints are presented in regards to intricacy, time, and precision as a result of unclear cluster borders shown in images, fuzzy techniques, on the other hand, are largely free of such problems and provide much better results in comparison to other segmented image methods [6].

In 2010, Padmavathi *et al.* stated that quality of underwater image is different from the quality of an image which capture in air, because some factors have impact of it such as; water medium, atmosphere, pressure, and temperature which means that image segmentation is necessary for digital image processing that implies image demonstrations is the need of picture segmentation, which separates a picture into portions that have strong correlations with objects to mirror the genuine data gathered from this present reality, picture segmentation is the most down to earth approach among practically all robotized picture acknowledgment frameworks, also clustering of numerical information shapes the premise of numerous arrangement and framework displaying calculations [7].

In 2011, Quintanilla-Dominguezab *et al.* tested FCM for the early stages of breast cancer detection using mammography technique [12].

In 2013, Jiang *et al.* utilized the fuzzy science strategy, and fuzzy grouping examined isolates the differentiate things and arranges them [13].

Furthermore, In 2013, Yambal and Gupta showed that segmentation process is unsupervised classification technique and an important step in advance image analysis process used as assistance to some other processes like detection for MRI brain tumor, generally original works of clustering are detecting anomalies, identifying salient features, classifying data, and compressing data, using conventional FCMs algorithm depend on hierarchical self-organized map, the aim of image segmentation process is to effective segmentation of noisy images [14].

In 2014, Khalid *et al*. illustrated that FCMs can diagnose some special diseases, such as glaucoma which is an ailment characterized by expanded weight inside the eyeball, making extreme harm the optic nerve, it is the most astounding reason for visual deficiency and irreversible, with early revelation early and appropriate treatment which it could continue for a long time [15].

Furthermore, In 2014, Norouzi *et al*. showed that mechanism of clustering algorithms is same as the classification technique without training of data, these methods have unsupervised learning algorithm and individual authority to calculate the similar features in the image and retain some things, same as; keys to recognize other features that have same attributes, this method is compatible with most of the data mining algorithms considering unsupervised methods they do not train data because its process is not time consuming in segmenting [1].

In 2017, Kumar *et al*. tested correlative distance by adding the process of eliminating, clustering, and merging to compute fuzzy weights using large initial prototypes and Gaussian weights. In standard FCM spatial FCMs methods incorporated the spatial information and altering of every cluster membership weights after considering the cluster distribution in the neighborhood [3].

In 2018, Ali *et al*. denoted that segmentation is an essential step to the sensitive analysis of human tissue lesions with aim of improving the partition of different clusters of images rely on similar features [16].

3. MEDICAL IMAGE MODALITIES

Imperative topic in medical imaging is medical image modalities or techniques. It is used to anatomical vision of body organs. There are some modalities which use in digital medical image processing.

3.1. X-Ray

Nowadays, X-ray imaging system uses in the branch of diagnostic radiology, however, there are many types of X-ray imaging systems, but the most famous one is conventional planar radiography which is too beneficial because of its low cost and low dose with side effect of interacting with useless beams. Its problem is overlapping anatomical image details and finding a lot of clusters [17].

In telemedicine uses as an integral part to automate X-ray segmentation which use broadly in remote areas which usually accidents happen there which helps the medical staffs to analyze the emergency cases, in this situation X-ray analysis in two manners; first one is segmenting the bone region from the surrounding flesh of the bone then extraction their features, the other one is determining the situation of the bone which usually happen this case, also it automatically selecting fractions aim doctors and medical staffs to analyze the level of injury to select the suitable medical treatment [17].

3.2. CT

To avoid overlapping effect, CT has been evolved. It can reshape the picture from all sides and angles until it can recreate the image of the body organ. High radiation dose is side effect of CT, also side effect of using that big amount of radiation increased day after day until it has detected that using CT cause of cancer higher than other types [17].

3.3. Digital Tomosynthesis

This technique has been invented to solve the problems of overlapping and high dose of radiation. It has moderate performance between these two disadvantages of X-ray and CT. Its angle was <360, but in digital theater systems, resolution depth increased with lower radiation compared to CT, because of this advantage, little angles accurately reshape images. Two major reshaping methods have been evolved: Analytical reconstruction and iterative reconstruction. In different algorithms of analytical reconstruction, filtered back-projection was the most famous one because of its smallest deforming and high precision. However, this method needs high pass filtering, to prevent this distortion maximum likelihood expectation maximization (MLEM) evolved which essentially involved iteration number with edge-preserving regularization. Another algorithm invented based on MLEM known as chest digital tomosynthesis (CDT) system to project data using a real CDT system [17].

3.4. MRI

Using of MRI images more common than other modalities in brain diagnosis and segmentation process. Among unsupervised learning methods FCMs clustering for medical image segmentation methods, FCM clustering is the most accurate one for segmentation process compare to other segmentation methods more appropriate for sensitivity noisy with intensity inhomogeneity MRI which can properly decrease the noises and supply superior segmentation results [18].

MRI is a wide medical modality that takes the image of the internal body and organs without contacting the skin. The characteristics of MRI are non-linear. A lot of things have influences of MRI accuracy, such as partial volumes effects which implies a pixel consists of more than one tissues, also, the volume is dependable thing in the segmentation process so as to determine the size of organs and strange things [19].

3.5. High-resolution CT

The segmentation is an indispensable step, as in many medical image analysis applications. Accurate segmentation using high-resolution CT (HRCT) images and quantification of the lungs have an important role in the early diagnosis of lung diseases. Especially detection of nodules in the lung region, airway and vessels diameter, lung volume is the key component of diagnosing lung diseases. The definition of each lung region from the CT images is the first step for the computer-aided diagnosis algorithm of lung. To extract each lung tissue, FCMs clustering algorithm has been applied to the segmentation of lung region in two-dimensions HRCT images [5].

4. METHODOLOGY

Many methods developed during several past years to enhance medical segmentation with demanding to obtain an accurate diagnosis, but FCM was the most accurate one in unsupervised learning methods. In supervised learning, when you feed the images or specified dataset to the method, it can extract similar features and cluster them according to observe of patterns. FCM distribute one piece of data to two or more clusters. However, the con of these methods is cannot label similar groups. For instance, in the tested image, FCM can determine the tumor without labeling it**.**

FCM clustering is the most accurate and widely method in the segmentation process of digital medical image processing compared to other methods. It acts as preprocessing, which aids classification and detection processes and sometimes use as a diagnosis process. Mentioning of fuzzy k-means (FKM) clustering is important because it used before FCMs clustering. Nowadays, using FCM is wider than FKM clustering, because it could solve these problems that had been happened in FKM clustering. FKM creates segmentations hierarchically, which involve each data point that can only be specified in one cluster, but FCM permits data points to be allocated into more than one cluster in which each data point has a degree of the endurance of belonging to each cluster as in fuzzy logic.

The mechanism of the process pass through these steps, as shown in Fig. 1; the first step is importing an MRI image that captured according to the specific technique of electromagnetic band imaging after that is converting the image in analog to same digital image size inside data acquisition step. Then preprocessing image begin by preparing images to the segmentation process, which involves image denoising and restoration. The process used median filter, as shown in Fig. 2. Then, FCM clustering implement and segment images into four parts. After that, the process converts the same image to same size analog image. Finally, export segmented image is as shown in Fig. 3.

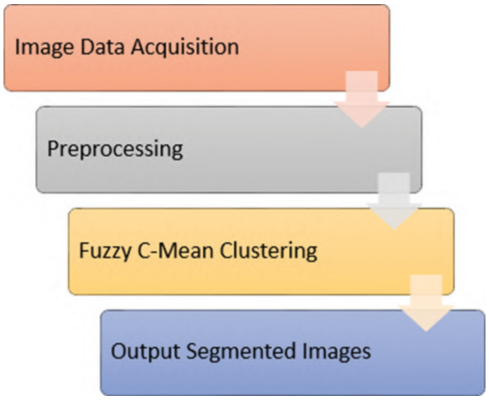


Fig. 1. Processes of FCM clustering.

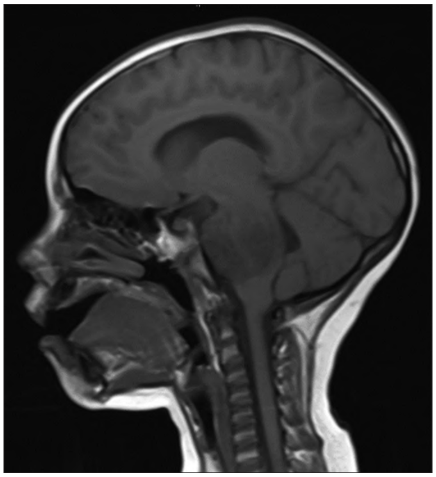


Fig. 2. Original image after the filtration process.

4.1. Mathematical Model of FCMs Clustering

In general, core of the FCM method is a mathematical model which relies on five steps until it can cluster similar features together inside one region. Because the first step of the work starts by taking values randomly from the membership matrix, so this example has depended on the imaginary value of two-dimensional image that illustrates the mathematical process of FCM clustering.

*m: fuzzification parameter; its range between (1.25 and 2).

while m: 2, i: first data point, j: first cluster, c: number of clusters.

4.1.1. First step

Randomly initialized the values as the membership matrix (Um) to the original image according to Equation.1, number of objects is 8, number of clusters is 4, The fuzziness parameter range (m) between (1.2 and 2), as shown in Table 1. Using this Equation 1.

$$\sum_{(j=1)}^{c}\mu_{j}(x_i) = 1 \text{ ......i} : 1,2,3,4, \text{ ......, k} \quad (1)\ [20]$$

4.1.2. Second step

After that, the model find constraints for every cluster. According to Equation 2.

$$C\_j = \frac{\sum_i \left(\mu_j(Xi)\right)^m X_i}{\sum_i \left(\mu_j(Xi)\right)^m} \quad (2)\ [20]$$

4.1.3. Third step

After that, the model finds distance (Di) for every cluster and find centroids, for instance, in the current example find; centroid (1), centroid (2), centroid (3), and centroid (4), using Euclidean rule, according to the Equation 3.

$$\mathrm{D_i} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (3)\ [20]$$

After that, the model finds distances for every cluster, as shown in Table 2.

4.1.4. Fourth step

Updating the membership values, according to the Equation 4.

$$\mu_j(x_i) = \left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}} / \sum_{k=1}^{c}\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}} \quad (4)\ [20]$$

In this case, m=2, i =first data point, j =first cluster, it is important new cluster values and put them in the new table so as to compare them with earlier cluster values to get new membership values, as shown in Table 3;

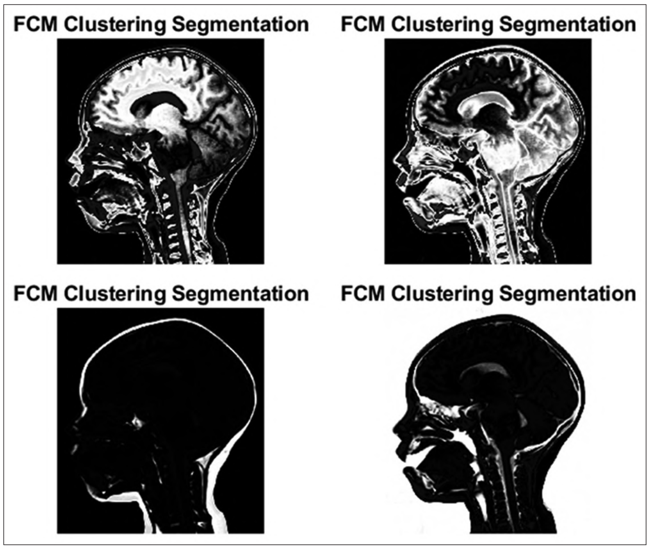

FCM Clustering Segmentation
FCM Clustering Segmentation
FCM Clustering Segmentation
FCM Clustering Segmentation

Fig. 3. All processes together.

TABLE 1: X and Y values of eight objects with initial values of four clusters of them

X	Y	C1	C2	C3	C4
2	8	0.1	0.2	0.3	0.4
4	6	0.3	0.4	0.6	0.8
6	4	0.5	0.6	0.9	0.2
8	2	0.7	0.8	0.2	0.6
1	7	0.2	0.1	0.5	0.1
3	5	0.4	0.3	0.8	0.3
5	3	0.6	0.5	0.1	0.5
7	1	0.8	0.7	0.4	0.7

4.1.5. Final step

This process iterates until it arrives correct centroids.

4.2. FCMs Algorithm

- Initialize the membership matrix of randomize values (fuzziness partition)
- Calculate the centroid vectors using the following equations

$$Cj = \frac{\sum_i \left(\mu_j(Xi)\right)^m X_i}{\sum_i \left(\mu_j(Xi)\right)^m} \quad (5)$$

And

$$\mathrm{D_i} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (6)$$

- Update partition matrix for new elements according to this equation

$$\mu_j(x_i) = \left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}} / \sum_{k=1}^{c}\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}} \quad (7)$$

TABLE 2: Data point and distance for all clusters

Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Data point	Distance	Data point	Distance	Data point	Distance	Data point	Distance
(2,8)	6.79	(2,8)	6.764	(2,8)	3.84	(2,8)	5.385
(4,6)	4	(4,6)	3.954	(4,6)	1.065	(4,6)	2.59
(6,4)	1.36	(6,4)	1.23	(6,4)	1.87	(6,4)	0.66
(8,2)	1.93	(8,2)	1.84	(8,2)	4.675	(8,2)	3.187
(1,7)	6.76	(1,7)	6.79	(1,7)	4.223	(1,7)	5.416
(3,5)	3.95	(3,5)	4	(3,5)	2.05	(3,5)	2.656
(5,3)	1.23	(5,3)	1.36	(5,3)	2.56	(5,3)	0.88
(7,1)	1.84	(7,1)	1.93	(7,1)	4.99	(7,1)	3.24

TABLE 3: Manually solved values of clusters

X	Y	Cluster 1	Cluster 2	Cluster 3	Cluster 4
2	8	0.198689327	0.199453065	0.351328263	0.250529346
4	6	0.136763286	0.138354361	0.513664923	0.21121743
6	4	0.204349796	0.225947742	0.148618034	0.421084428
8	2	0.326016177	0.34196262	0.134590635	0.197430568
1	7	0.206419945	0.205507927	0.330428327	0.257643801
3	5	0.185132797	0.182818637	0.356719292	0.275329273
5	3	0.26436788	0.23909742	0.127020505	0.369514195
7	1	0.346019973	0.329884326	0.127590531	0.19650517

- Repeating this process by checking convergence until it gets the same manually result in centroids if not the loop returns back to step 2.

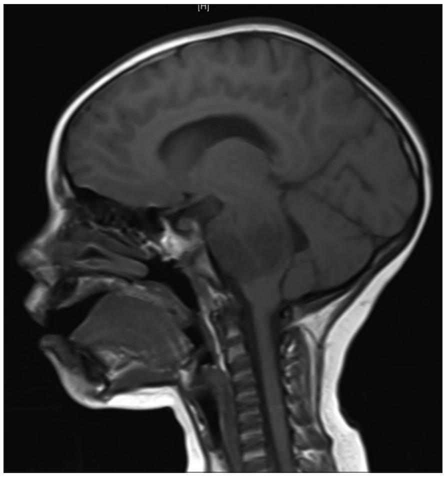


Fig. 4. Original image.

5. RESULTS AND DISCUSSION

The imperative thing in the procedure of digital medical image segmentation is the sensor or camera which use to capture images. This steps the primary step in the process and it has direct interaction with the physical organic things because the natural form of every signal is an analog signal. In medical imaging, it varies according to techniques of electromagnetic band imaging because any technique uses in the specified range to take an image. This study relied on using MRI because it is the most effective technique to diagnose brain tumor. Most of the time converting analog to digital happen inside the sensor. Data acquisition is the first and important process in methodology for the training and testing process, but FCM clustering method is unsupervised learning. Testing on this method did not depend on the specified dataset because it is only segmentation process and tested on five ready MRI images taken from the internet that properly applied on the proposed method, which could clustered images well, as shown in Fig. 4. After that, preprocessing step begins by image restoration using median filter and resizing because the segmentation process needs to find regions accurately and median filter does same process without damaging any edges. The next step is implementing FCMs algorithm which use to segmenting medical images accurately. The proposed method dispatches the images to some clusters according to areas and how much iteration necessary to execute the method until it can show the accurate results. The backbone of the issue is the results of the algorithm that is applied on and original gray scale image of 4 years child head MRI image in the left side. The target of this process is diagnosing an abnormal mass in the brain, as shown in Fig. 4.

The best method to solve these cases is FCM clustering because it simplifies the process of feature extraction. It separates different attributes according to these clusters that determine optionally. For instance, in the current method determined four clusters. The FCN algorithm depends on the fcn() method inside fuzzy logic toolbox inside the MATLAB tool. Every mathematical step automatically happens inside the ready method according to the called image, but it iterates inside the method automatically 100 times to find the values of fcn objects and cluster them. Matrix Laboratory known as MATLAB is a practical robust tool used to conduct procedures within the processing of digital images. It used to carry out mathematical computations with matrices and vectors, which is very straightforward to operate as it incorporates computation, visualization, and programming inside the system [21].

Many methods used to prevent the noise problem in digital image processing. However, the best filtering technique in this case is median filtering. In addition, different types of filters are applied to remove different types of noise. There are many types of filtering, such as median, mean (average), and Gaussian.

After that, implementing the FCM clustering method starts by applying preprocessed images to the FCM method. It involves grouping the similar images data and distributing it into N clusters with all data points inside the images. Different clusters exhibit different visions, as shown in Fig. 5. This method can illuminate some properties of the image. It makes clusters over that rule.

Fig. 6 shows the mass inside the head and it aids the medical professionals to decide accurately.

Fig. 7 shows the image separated into N clusters.

Fig. 8 shows the background of the image.

In general, all of these processes mixed in this method, as shown in Fig. 3.

The result of FCM is the most accurate compared to other unsupervised learning methods. FCM is used in a wide range of applications compared to other types, especially

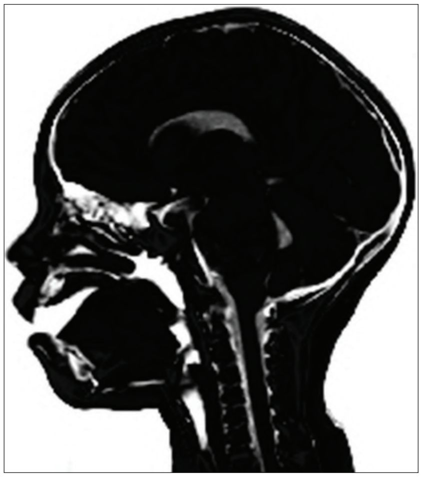


Fig. 5. Spaces inside the head magnetic resonance imaging.

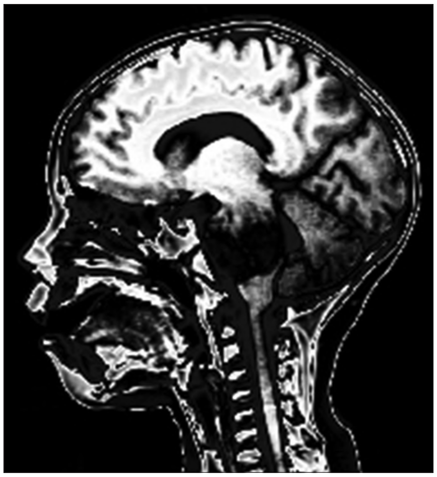


Fig. 6. Segmented tumor image.

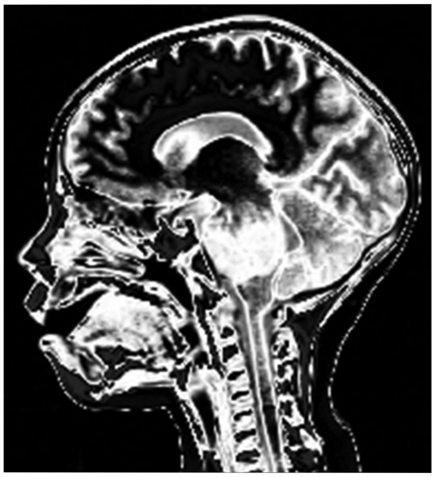


Fig. 7. Segmentation according to N clusters.

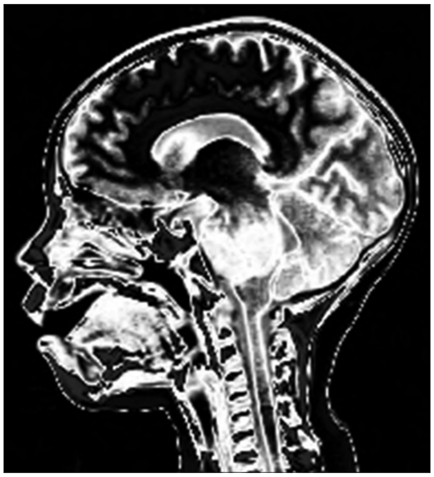


Fig. 8. Background of the image.

in diagnosing. Recently, it uses in most of the segmentation cases relate to medical images.

6. CONCLUSIONS

New image segmentation process using FCM clustering is very important to get every desired feature and making clusters to extract their patterns. In addition, the results of every step are very important for manually working to find what are the weak points capable to change and improve. Comparison between final values to initial values is done to realize the difference between them. The FCM clustering method performs this process automatically, but changing parameters are so beneficial criterion to get the best accurate segmented images.

FCM method is the most effective segmentation approach in unsupervised learning methods which used in medical image segmentation processes. The mathematical procedure of this approach is illustrated step by step. FCM clustering is an unsupervised learning method, which is an accurate segmentation process. It can rely in on diagnosis process, such as brain tumor, which tested. Then, the important role in finding the boundary of components of medical images and how it can find ROI and segment the organs and abnormal shapes. This approach supports medical professionals to take the correct decision.

REFERENCES

[1] A. Norouzi, M. S. M. Rahim, A. Altameem, T. Saba, A. E. Rad, A. Rehman and M. Uddin. “Medical image segmentation methods, algorithms, and applications”. *IETE Technical Review*, vol. 31, no. 3, pp. 199-213, 2014.

[2] R. C. Gonzalez and R. E. Woods. “*Digital Image Processing*”. 4th ed. Pearson Education, New York, 2018.

[3] R. Kumar, G. Satheesh and B. Nisha. “MRI brain image segmentation using Fuzzy C means cluster algorithm for tumor area measurement”. *International Journal of Engineering Technology Science and Research*, vol. 4, no. 9, pp. 929-935, 2017.

[4] T. Saikumar, P. Yugander, P. S. Murthy and B. Smitha. “Colour Based Image Segmentation Using Fuzzy C-Means Clustering”. In: *International Conference on Computer and Software Modeling*, Singapore, 2011.

[5] E. Doğanay, S. Kara, H. K. Özçelik and L. Kart. “A hybrid lung segmentation algorithm based on histogram-based fuzzy C-means clustering”. *Journal Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, vol. 6, no. 6, pp. 638-648, 2014.

[6] S. Naz, H. Majeed and H. Irshad. “Image Segmentation Using Fuzzy Clustering: A Survey”. In: *6th International Conference on Emerging Technologies*, Islamabad, Pakistan, 2010.

[7] G. Padmavathi, M. M. Kumar and S. K. Thakur. “Nonlinear image segmentation using fuzzy c means clustering method with thresholding for underwater images”. *IJCSI International Journal of Computer Science Issues*, vol. 7, no. 3, pp. 35-40, 2010.

[8] O. Jamshidi and A. H. Pilevar. “Automatic segmentation of medical images using Fuzzy c-means and the genetic algorithm”. *Journal Computational Medicine*, vol. 2013, p. 972970, 2013.

[9] G. Stephanie. “*Fuzzy Clustering Definition*”, 2016. Available from: https://www.statisticshowto.datasciencecentral.com/fuzzy-clustering. [Last accessed on 2019 Oct 01].

[10] L. Ma, H. Chen, K. Meng and D. Liu. “Medical Image Segmentation Based on Improved Fuzzy C-Means Clustering”. In: *International Conference on Smart Grid and Electrical Automation*, Changsha, China, 2017.

[11] M. M. Ahmed and D. B. Mohamad. “Anisotropic diffusion model segmentation of brain MR images for tumor extraction by combining kmeans clustering and Perona-Malik”. *International Journal of Image Processing*, vol. 2, no. 1, pp. 27-34, 2008.

[12] J. Quintanilla-Dominguezab, B. Ojeda-Magañaac, M. G. Cortina-Januchsab, R. Ruelasc, A. Vega-Coronab and D. Andinaa. “Image segmentation by fuzzy and possibilistic clustering algorithms for the identification of microcalcifications”. *Scientia Iranica*, vol. 18, no. 3, pp. 580-589, 2011.

[13] H. Jiang, Y. Liu, F. Ye, H. Xi and M. Zhu. “Study of clustering algorithm based on Fuzzy C-means and immunological partheno genetic”. *Journal of Software*, vol. 8, no. 1, p. 134, 2013.

[14] M. Yambal and H. Gupta. “Image segmentation using Fuzzy C means clustering: A survey”. *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 7, pp. I-5, 2013.

[15] N. E. A. Khalid, N. M. Noor and N. Ariff. “Fuzzy c-means (FCM) for optic cup and disc segmentation with morphological operation”. *Procedia Computer Science*, vol. 42, pp. 255-262, 2014.

[16] N. A. Ali, B. Cherradi, A. E. Abbassi, O. Bouattane, M. Youssfi. “GPU Fuzzy c-means algorithm implementations: Performance analysis on medical image segmentation”. *Multimedia Tools and Applications*, vol. 77, no. 16, pp. 21221-21243, 2018.

[17] O. Bandyopadhyaya, A. Biswasa and B. B. Bhattacharyaba. “Long-bone fracture detection in digital X-ray images based on digital-geometric techniques”. *Computer Methods and Programs in Biomedicine*, vol. 123, pp. 2-14, 2016.

[18] S. K. Adhikari, J. K. Sing, D. Kumar, B. M. Nasipuri. “Conditional spatial fuzzy C-means clustering algorithm for segmentation of MRI images”. *Applied Soft Computing*, vol. 34, pp. 758-769, 2015.

[19] G. S. Chuwdhury, M. Khaliluzzaman and M. Rashed-Al-Mahfuz. “MRI Segmentation using Fuzzy C-Means Clustering and Bidimensional Empirical Mode Decomposition”. In: *International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering*, 2016.

[20] E. N. Sathishkumar. “*Fuzzy c Means Manual Work*”. Lecturer at Periyar University, Salem, 2015.

[21] R. C. Gonzalez, R. E. Woods and S. L. Eddins. “*Digital Image Processing Using MATLAB*”. Pearson Education, London, United Kingdom, 2004.

ORIGINAL RESEARCH ARTICLE | UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Review Study on the Adoption of Cloud Computing for Higher Education in Kurdistan Region – Iraq

Abbas M. Ahmed[1], **Osamah Waleed Allawi**[2]

[1]Department of Business Administration, Sulaimani Polytechnic University, Sulaimani, Iraq, [2]Department of Computer Technology, Al-Hikma University College, Baghdad, Iraq

ABSTRACT

Cloud computing (CC) is considering as a popular computing model in the Western World. It is still not well understood by much higher education (HE) institutions in the developing world. CC will positively affect its consumers in executing their role in an economical way. It can be done using applications provided by the cloud specialist organizations. This study aims to evaluate the factors that influence the adoption of CC for HE within the Kurdistan Region in Iraq. The study was performed utilizing a non-experimental study exploratory research design. This exploratory study included an essential investigation into secondary data. The study development and modeling of secondary data to highlight the final results of the research. Through reviewing the literature of the existing frameworks in CC adoption, it is showed that there are limited institutions developed over the latest years. Moreover, HE in Kurdistan Region needs continued attention to get government support and redesign the educational system to cover all the core aspects in a better way. Here, at any time, there is a need to access the applications, software and hardware, platform, and infrastructure; the most required is to have the internet service.

Index Terms: Cloud Computing Adoption, Higher Education, Education Systems, Kurdistan Region – Iraq, Electronic Learning

1. INTRODUCTION

Higher education (HE) scenery all over the world is in a continuous state of influx and development, mainly as a result of essential challenges stemming from efforts in adopting new and growing technologies. Using technology will improve HE which will result in providing high-quality education and prepare the students to face the challenges of the 21st century [1]. Kurdistan regional government (KRG) could be a developing area in several faces, HE has been developed in this region, there are 28 universities, according to the Kurdistan ministry of HE (MHE) [2].

Access this article online

DOI:10.21928/uhdjst.v4n1y2020.pp59-70	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Cloud computing (CC) is a collection or group of hardware and software to human beings through the internet. CC provides many advantages such as steady, rapid, sample, suitability, and simultaneous accessibility of belongings at low cost in comparison with other techniques through the internet to the users. Resources can be requested by the consumers depending on their requirements. These requirements can be storing data, communication, data processing, and calculation cycles needed for their applications [3].

Each cloud has its own users. The services of the cloud can be accessed by the user to retain the increasing daily and safety systems in the CC environments. The specific role of

Corresponding author's e-mail: Abbas M. Ahmed, Department of Business Administration, Sulaimani Polytechnic University, Sulaimani, Iraq. E-mail: abbas.ahmed@spu.edu.iq

Received: 16-12-2019 Accepted: 18-03-2020 Publishing: 20-03-2020

the CC is that it can be used through exploiting the internet and the PCs in the data centers. The role of CC is important in the academic and industrial domain [4].

As mentioned above, HE is in continuous development according to the requirements of modern life. It is so-known that effectively using technology in HE is one of the key factors for providing high-quality education. The cost is the main reason for the slow adopting of new technologies in HE. The local societies and the whole world transformation demand huge funding and investment. These factors are difficult to come at the times of deep economic downturn and depleted budget reserves whether budgets of the government or the private institutions. The financial support provided to HE institutes has sharply decreased in times of recession, leading to financial difficulties in HE institutions (HEIs). To address their fiscal deficit, HE institutes have recourse to a variety of cost-cutting measures, including important cuts to information technology (IT) budgets [5].

However, in this paper, many studies have been reviewed, discussed, and critically analyzed to providing a solid literature review for future research. In addition, a wide range of case studies from past up to date is presented for a better understanding of the theory related to applying the CC in HEs. Articles, journals, books, and previous works had been listed in the following tasks.

Many well-established reviews and survey articles on applying CC in HE available in the literature such as Amron *et al.*, Qadri and Qadri, Rawajbeh *et al.*, Al-Shqeerat *et al.* [6]-[9], Singh and Baheti [10]. Amron *et al.* [6] reviewed three sectors in applying CC which will be the health-care sector, higher learning organization, and the public field. In the manner, Qadri and Quadri [7] reviewed numerous CC applications with emphases on the security aspect. In the work of Al Rawajbeh *et al.* [8], they clarify the roadmap of the successful adoption of CC in high education institutions. In the work of Al-Shqeerat *et al.* [9] provided baseline recommendations to avoid security risks efficiently when adopting CC in HE. Whereas, in the work of Singh and Baheti [10], they discussed the limitations and problems of the traditional education methods in additional education based on CC. Fig. 1 shows the number of reviewed and discussed articles in this work based on the years, note, and * represents the number of review articles.

1.1. Paraphrase

A total of 39 research articles and five review articles are covered in this work. The review emphasizes the CC service and its applications in HEs. Furthermore, our work distinguishes itself from the previous by it is studying the possibility of applying the CC service in HE of the Kurdistan Region – Iraq. This review approach allows us to improve the scope and shape the direction of HE based on IT.

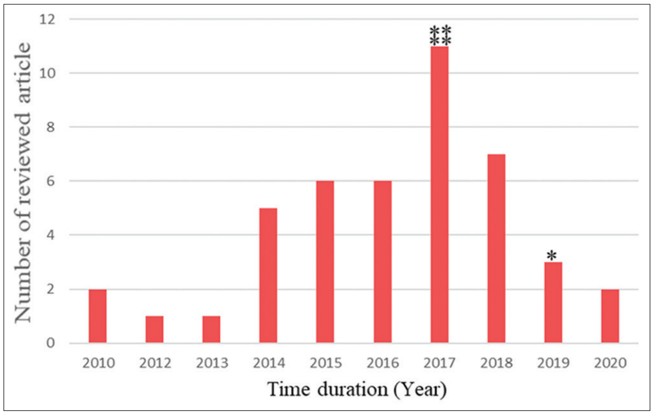


Fig. 1. Number of the reviewed article.

This paper segmented into four parts starting with the section of introduction which describes the CC service and it is an application in education. Furthermore, the related work has been discussed in section 2. Furthermore, in section 3 an overview of the CC service and it is an application in HEs. Finally, section 4 presents the conclusion of this paper.

2. RELATED WORK

A literature search is a pre-requisite for reviewing the literature on any subject, in general, this is done by scanning some prominent journals and conferences exclusively dedicated to the subject, concentrating on limited outlets cannot be considered as enough justification for a literature review on CC in HE as this is a recent phenomenon [2]. This is the reason why the publication channels till now are largely scattered for most of the concurrent phenomena, information science researchers and scholars are using online databases as their first literature collecting strategy.

The CC aspect is useful when was implemented for some universities such as California University (UC). They found that implementation of CC enhanced the development. Also, the implementation of software as a service (SaaS) applications CC which made the difference in increasing the advantages to students, such as make the exams online, have access to their exercises and solve it and resend it again, projects submitted by students, feedback facility between students and teachers. It also provides the facility of communication between students, using the applications by the students as well as the teachers without installing those applications on their computers and without store-related application files. Furthermore, the ability to access any computer from anywhere and at any time when the internet service is available can be possible too. The core concern related to applying the CC in HE fields is security issues. At the same time, this does not mean that there are some other obstacles related to trusting, trust, and assurance [11].

The educational cloud represents one of the most interesting applications of CC. To meet their most requirements, the private educational institutes are betaking toward using IT technology. The increasing dependence on IT requires the availability of the internet to students and institutes [12].

One of the main problems in Iraq is the lack of network infrastructure. Abdusalam *et al*. [13] mentioned useful information on the challenges and current status of the backbone infrastructure and internet in the KRG which provided by private companies from several countries, namely, Iraq, Iran, Turkey, and the others [14]. Universities in KRG are willing to use modern techniques in the education process and teaching methodologies. The cornerstone of the modern education system is using Information and Communication Technologies (ICT). Unfortunately, the MHE in Kurdistan suffers from a lack of ICT infrastructure in its governorates [15], and clearly; this means that establishing ICT infrastructure for universities requires extensive time, investment, and efforts. Furthermore, some researchers in Masud *et al*. [1] aimed to improve an instrument to investigate the factors of CC service based on the theory of planned behavior.

Some researchers focused on the materiality of each dimension and weight of each sub-dimension such as Thabit and Harjan [16]. The researchers checked the opinions of the Avicenna Center and Gihan University academic staff with 40 questionnaires. The questionnaire also contains some points related to developing the activities of the Avicenna center in Erbil. Thabit and Harjan [16] concludes that the e-learning Avicenna Center has to develop a new department of training the staff to deal with e-learning centers and improves the university students' skills to create a new generation compatible with e-learning technologies.

In addition, Riaz and Muhammad [17] proposed that the limitation of the education requirements in growing countries like Pakistan can be solved through the adoption of the CC. This action can guarantee that all the software resources and ICT based possessions can be shared among learners.

All the data confidentiality and integrity processed by the institutes can be defined as the data security according to many studies. This risk can be reduced through a model that complies with all the academic and administrative staffs' requirements and at the same time, all the users' devices are separated to reduce data theft chance. This action was done by controlling the data storage of each user device through different port from other users' data storage [18].

Amron *et al*. [6] reviewed three sectors in applying CC which will be the health-care sector, higher learning organization, and the public field. Five key factors completely outclassed all three sectors; technology preparedness, human readiness, organization assistance, environment, and security, as well as privacy. Factors of connection and feedback and access to the internet hereditary factors pertinent to the HE community, generally the study is motivated by curiosity about the dependability of CC to become the leader in information storage technology. Although several studies found the actual CC brings much more benefits than disadvantages, the particular negative effects of the applied CC should become also being noted especially in the aspect associated with safety and data personal privacy factors [6].

The study by Qadri and Quadri [7] has assessed the behavioral intention of the students of Iraq, being in its infancy in terms of internet adoption; thus, going through the transformation of traditional modes of learning into e-learning modes. The study has employed the modified form of "technology acceptance model" (TAM) model to assess the attitudinal behavior of the students of Iraqi HE toward the use of learning management system as the educational platform, the study has led to the conclusion that there exists a significant association between the variables under consideration. The standings of Iraqi HE are noted to be significantly improved with respect to the past statistics. Besides, the study also affirms the credibility of the TAM model in facilitating the assessment criteria for diverse technological deployments.

CC has considerable standing in the HEIs worldwide and locally. As well as in Saudi Arabia, typically the IT market is considered such as the largest sector in the Gulf area. The Saudi government offers allocated huge finance to improve the academic environment with the very best technological facilities. On the other hand, there are unique start-up universities in Saudi Arabia that absence e-learning tools in comparison to the elderly universities in SA, Saudi universities even now slowly seek to embrace CC in the HE atmosphere for distance studying and e-learning, while CC has been broadly used in universities within different countries to provide higher quality services to be able to HE and also CC enables HEIs to deal with the needs regarding software and hardware modifications rapidly at lower expenses. Therefore, the adoption involving CC into HE encourages students' academic level in addition to efficiency. The research came to the conclusion that there is a good urgent need to produce a new web software based on CC as well as cover some of the holes in existing web applications [8].

CC represents a great opportunity for universities so that you can take advantage of the actual enormous benefits involving cloud services and also resources in the educative process. However, the cloud end users remain concerned regarding security issues that symbolize the major obstacle which may prohibit the usage of CC on a large scale. Typically, the limitations of cloud support models were investigated in addition to challenges as well as risks threaten cloud processing. The study demonstrates that the stakeholders are usually not familiar with feasible security risks or operations used to protect data as well as a cloud application. Furthermore, this indicates that the many serious attacks may threaten cloud networks usually are denial of service and also phishing attacks [9].

The teaching materials may be made available through the cloud service workers to educate the customer on the available risk operations issues as it pertains to cloud usage. This shows how crucial it will be for educators who usually are cloud service users to be able to understand how to manage all their information used in often the cloud. To the students, this enhances their participation in studies, increases all their enthusiasm and motivation, therefore the time at that they study is raises while the cost is actually reduced. The students obtain limitless access to net-based teaching-learning sources needing little or absolutely no effort from the teacher. Studying is gradually made electronic as educational institutions transfer their resources, students info system, learning management techniques, knowledge management techniques to the cloud, with that, students are capable to access the needed sources from anywhere in a versatile way [19].

The goal of Sultana *et al.* [20] study is to determine the factors that will certainly influence CC adopting in university associated with Dhaka of Bangladesh. In this research, some significant factors possess been derived from information collection and data analysis in different functions of this university. The absence of proper infrastructure, services availability, and effectiveness in education are observed most important. Some other factors are resource require, cloud control ability as well as lack of training of employees. An educational institution may focus on these elements to increase the utilization of CC technologies to provide studying to the student. Singh and Baheti [10] were conducted their study to overcome the limitations of traditional education and learning system, CC solutions tend to be very useful for academic institutions especially with regard to HE institutes. Along with the involvement associated with CC in learning system, students can obtain access to various sources (i.e., textbooks, magazines video lectures, demonstrative video, and lab facilities) which are not achievable in traditional education and learning system. Teachers can assess students in a much better way; researchers can obtain all the facilities as well as infrastructures related to all their research field. Definitely, not only teachers and learners but administrators should also opt for equipment for administration purposes. Inside overall CC has different services that might be included in the actual traditional education and learning system.

A study conducted by Başaran and Hama [21] to investigate university faculty members' views toward the adoption of CC in HE. The current status of the faculty on CC usage in education and regional differences was discussed. The data were collected through an adopted questionnaire based on these frameworks and demographic information was answered by 300 faculty members from the northern parts of Cyprus and Iraq. The results showed that faculty members agreed mostly on the opportunities followed by an awareness of potential threats and weaknesses and finally they accept the strengths of adopting CC in education. The study brought to light on the comprehension of faculty members' views from comparative and integrated framework perspectives. In general sense, faculty members from the north part of Iraq seem to be slightly more optimistic about the adoption of CC in educational settings. This might result from either they less frequently use CC services as compared to faculty members from the north part of Cyprus who are younger in mean age and can be considered as being more capable consumers of cutting-edge technologies like cloud. Interestingly, both parties are aware of the problems which could be resulted from adopting such innovation.

With the number of works of literature reviewed above, it shows that a number of studies have been conducted on the adoption of CC at HE in Kurdistan. These studies are carried out in different environments, countries, and industries.

These studies showed that the CC is constantly evolving, and it became necessary for the different processes and activities within HE universities, and this requires the universities to apply it and use it. Despite the importance of CC usage and its role in activating the learning in Iraq to enhance education, there was a lack of researches and studies on the factors that affect the adoption of CC in Iraqi universities. As a result, there is a knowledge gap. This study aims to fill that gap. However, the analysis of the related work is shown in Table 1.

Based on the above discussion, it is observed that most of the studies emphases on the readiness factors of technological, HE, cost, educational, information security and cultural, as shown in Table 2 below:

TABLE 1: The analysis and comparison of the related work.

Ref.	Type of service	Country	Employment place	Advantage	Disadvantage
Kadhim [11]	Cloud computing	Iraq	Basic education	It can improve the educational sector It is suggesting the applying of decision-making a feature in education Internalize storage for confidential work	Limited by geographical scaling
Hashim *et al*. [12]	Cloud computing	North Iraq	Higher education	It gives the students an open and flexible environment by applying the VCL in Bayan University	The proposed system does not test yet The study is limited only on the Bayan University
Abdusalam *et al*. [13]	Cloud computing	North Iraq	Governments Organization	It assists in locating missteps in the implementation stage It is focused on the information security aspect	The study is limited on only three status which are Dhok, Erbil and Sulaymaniyah
Al-Hashimi *et al*. [14]	Cloud computing	North Iraq	Higher education	The possibility of applying the Cloud computing services in the North Iraq Universities Operations have been proposed for budget reductions	limited by geographical scaling
Abdulkadhim *et al*. [15]	Electronic document management system	Iraq	Governments Organization	It draws on the research results for the implications of IT managerial practice It provides enhance in managing the EDMS implementation process in government	The study focused only on government organizations
Asadi *et al*. [22]	Cloud computing	General	Higher education	demonstrated validity, reliability, simplicity, and functionality of the f the Theory of Planned Behavior – Cloud Computing Services use Questionnaire TPB – CCSQ	The proposed system is tested only in higher education
Thabit and Harjan [16]	Electronic learning	North Iraq	Higher education	Spread the culture of applying the E. Learning in Avicenna Center of Erbil Develop a new department of training the staff to deal with e-learning centers	The study focused only on the Avicenna Center of Erbil
Riaz and Muhammad [17]	Cloud computing	Pakistan	Higher education	It presents the usability evaluation of public cloud applications across three universities in Pakistan from stakeholders' perspective, i.e., (teachers and students)	They did not take into consideration the applying of Google sites to find out the effects of public cloud application in the education sector
Nofan and Sakran [18]	Cloud computing	General	Education	It is given a better understanding of the conception of cloud computing technology and its impact on teaching and learning in institutions	They did not take into consideration the information security aspects
Ariwa and Aiwa [19]	Cloud computing	Nigeria	Higher education	Cloud computing in Nigeria will transform the traditional education model to computer-based virtual applications with a focus on e-pedagogy	The study is limited only in Nigeria The study is limited only in higher education
Sultana *et al*. [20]	Cloud computing	Bangladesh	Higher education	It Identified the factors that will influence cloud computing adoption in University of Dhaka of Bangladesh	The scope of the study is limited only at the University of Dhaka
Başaran and Hama [21]	Cloud computing	Turkey, Iraq	Higher education	It offered education-specific solutions to institutions regarding cloud computing adoption	They did not take into consideration the information security aspects

VCL: Virtual computing laboratory

TABLE 2 : The most common readiness factors that have been used in the field of cloud computing service in higher education.

Readiness factors						
Fac.	Technological	HR	Cost	Educational	Info. security	Cultural
Ref.						
[11]	✓	x	✓	✓	✓	x
[12]	✓	✓	✓	✓	✓	x
[13]	✓	✓	✓	✓	✓	✓
[14]	✓	✓	✓	x	✓	✓
[15]	✓	✓	✓	x	✓	x
[22]	✓	✓	x	x	✓	x
[16]	✓	✓	✓	x	x	x
[17]	✓	✓	✓	✓	x	✓
[18]	✓	✓	✓	x	✓	x
[19]	✓	✓	✓	x	✓	x
[20]	✓	✓	x	x	✓	✓
[21]	✓	✓	x	x	x	✓

3. AN OVERVIEW OF CC AND ITS APPLICATION IN HE

In this section, an overview of CC and it is an application in HEs is given. Furthermore, it is consisting of four parts, which are CC definition, the benefits of CC, the applying of CC service in education, and the employing of CC service in HE, as shown in Fig. 2.

3.1. CC Definition

There are a set of important definitions and reviews on CC. The which is the National Institute of Standards and Technology defined the CC as a model which enables handy, a network access when required to a shared pool of configurable computing resources, for example, networks, servers, storage, applications, and services that can be swiftly provisioned and released with the minimum effort of management or service provider interaction [11]. On a more elementary level, CC can be a systematic way for managing a set of virtual computers somewhere automatically and control them in such a simple way to create, manage, or even destroy over the network, without human action [12].

There are various security types within the CC technique, which can include networks, databases, operating systems, resource scheduling, virtualization, transaction management, concurrency control, and memory management [4]. Hashim *et al.* [13] referred that the key benefits of the CC are the ability to allow users to access data and software anywhere whenever there is internet service available and the ability to smooth sharing of learning materials and data, while the concerns about the security and data privacy can be considered as the main obstacle in this field.

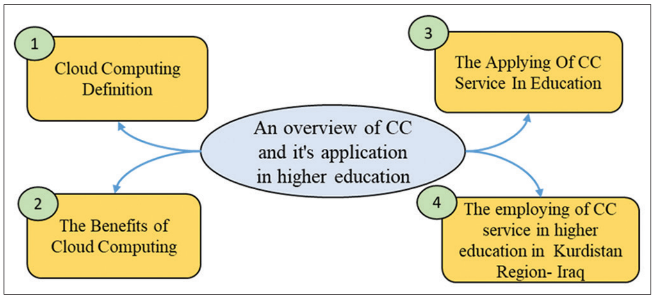

1
Cloud Computing Definition
3
The Applying Of CC Service In Education
An overview of CC and it's application in higher education
2
The Benefits of Cloud Computing
4
The employing of CC service in higher education in Kurdistan Region- Iraq

Fig. 2. An overview of cloud computing and its application in higher education.

The CC provides many types of services and when there is a full understanding of these services, it will be clear what this approach is all about. The main types of cloud services can be illustrated below:

- What is so-called Infrastructure as a Service (IaaS): Services provided by this level include the remote delivery (through the internet) of a full computer infrastructure (e.g., virtual computers, networks, and storage devices) [14]. The perfect example of this kind of service is Amazon1 which offers S3 for storage, EC2 for computing power, and simple queue service for network communication for limited businesses and individual consumers [23], just such as computer server and processing power [22].
- Platform as a Service (PaaS) which is PaaS: In this field, PaaS offered the ability to provide a software application without the need to install the software tool in consumer computers. The cloud development environment is the CC main access tool and their examples are operating systems, software testing tools [22].
- The other layer within these services is SaaS and under this layer, applications are delivered through

the medium of the internet as a service. This type of service is running on the provider's infrastructures and is accessed through the client's browser (e.g., Google Apps and Salesforce.com) [6]. To use the required software, it can be simply accessed through the internet and this action can nominate the need to install the software itself. The complete functionality of the applications is embedded within this type of cloud service and these functionalities are a variety from the productivity such as (office-type) applications to programs just like those for the management of enterprise-resource or which refers to Customer Relationship Management, for example, Billing Software, Image, and Video editor [22], [23]. Moreover, Fig. 3 shows the types of CC service.

3.2. The Benefits of CC

In fact, there are multi benefits or advantages for the solutions provided by CC. Some of these benefits over traditional technologies are illustrated below:

- Mobility: In general, the current orientation is to increase the dependency on the facilities provided by mobile devices. In the HE field, the students harnessed the mobile devices' facilities to access data whether these data were a textbook, researches, syllabi, or even have the privilege to do their own homework. The applications within the cloud-based classroom can be considered as the most efficient way to make the exchange between student and faculty easier [7].
- New Services: The cost of traveling (for the international students), as well as other difficulties related to attendance in the classrooms, motivate the need for starting virtual classrooms through online learning and video conferencing which is provided nowadays by many colleges and universities. The universities which offer the facilities to enable the students to join the classrooms from anywhere around the world using their own mobiles, computers, or tablets; could not provide such a service without the cloud servers [18].
- Storage: The actual usage of the CC by the universities provides them with the ability to quickly expand storage capabilities through scalable cloud storage. The data related to students are huge, starting from their own information, their marks, their medical records, and any other data. Here, the core risks are the chance to have a situation where these data can overwhelm traditional storage or even lost. The scalable cloud storage property alongside with business continuity and disaster recovery can be used to avoid such situations [9].
- Efficiency: In HE, there is always a striving by the universities to improve their organizations. Almost 55% of the higher learning institutions looking forward to increase the efficiency and they trust in CC as the best way to achieve this goal [5].

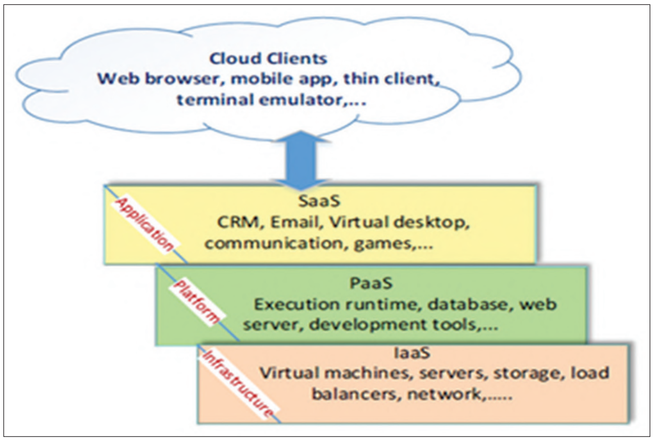

Cloud Clients
Web browser, mobile app, thin client, terminal emulator,...
SaaS
CRM, Email, Virtual desktop, communication, games,...
Application
PaaS
Execution runtime, database, web server, development tools,...
Platform
IaaS
Virtual machines, servers, storage, load balancers, network,.....
Infrastructure

Fig. 3. Service models of cloud computing [22].

Like everything in this life, there is advantages and benefits while in the other side there is risks and limitation and the CC is not an exception (Table 3 shows these risks and limitations), multiple cloud CC can the HE institutes choose but they have to take in their considerations the real need and the institute strategy itself [24].

3.3. The Applying of CC Service in Education

A rudimentary understanding of information and communications technology in the education field is one of the motivations and key factors for what can be seen as fast-changing technology. It is a necessary issue for HE actors to have a full understanding of how the cloud CC is adopted as well as involved. To transform the HE systems to be cloud-based systems, knowledge use and creation are a critical factor to ensure the full social, economic, and cultural transformation [24].

Coping with rapidly changing software and hardware needs at a lower cost in HE motivates many researchers to migrate from the classical systems toward the CC technique. The HE corporations planning to use 20% of the information techniques budget allocated for them, this will be done by shifting their applications toward the cloud. The challenges that would face this transition should be addressed within inclusive CC strategy; on the other hand, this step will ensure a smooth transition as well as optimal results to increase the institute's organizational efficiency [5].

TABLE 3 : The benefits and limitations of applying cloud computing in higher educations [24].

Benefits	Limitations
It is available anywhere and anytime	Not all applications run in the cloud
Support for teaching and learning	Data protection and security issues
Low cost, since it is free or pay depend on the use	Organizational support
Opening to the business environment and advanced research	Dissemination politics, intellectual property
Using green technologies to protect the environment	Maturity of solutions
Increased openness of students to new technologies	Lack of confidence
Increasing functional capabilities	Standards adherence
Offline usage with further synchronization opportunities	Speed/lack of internet can affect work methods

In HE, it is obvious that the main beneficiaries are students and all faculty staff (academic and administrative). These users have access to the data alongside the control on those data through the internet. The privileges and activities to all the users who connected to the cloud are a variety from uploading lectures, assignments, and tests (for teachers) as well as accessing those lectures, assignments, and tests (for students) re-upload the assignments and test if necessary. The main requirement to access the cloud anywhere and anytime is the availability of internet service [12], [25]-[27].

However, what so-called “the intelligent education” or the “Electronic education” can get its entire requirement to be efficient like the application software itself as well as the required database alongside with email management from the SaaS. On the other hand, the shortage or the breakdown here is concentrated in the dependence of this technique on the internet. Since that, the internet is a key factor for the permanence of the CC; all the HE users (staff and students) have to ensure the continuity of the internet service as well as ensuring that the internet connection is fast enough to have the full access to the cloud services at any time [28].

Moreover, in the work of Al-Khayat and Al-Othman [29] tried to make the design of the educational CC comprehensive and complete. The proposed generic CC model is to implement many frameworks for improving the quality of education of students and academic staff besides saving time. Although the using of modern ICT are the cornerstone of modern teaching and learning in the engineering colleges and institutes in Iraq, the extensive use of educational technologies and investing time and efforts in buying and maintaining infrastructure was disrupting the aim of establishing effective teaching and learning environment. To face this big problem and to an emphasis on quality of education, there should be an awareness about the CC benefits on cost-effectively providing better education services in addition to making a real investment of CC in providing both SaaS and infrastructure [29].

The CC resources can be accessed whenever required and with the minimum effort needed for managing these resources. The goal of applying the CC on HE is differing from one country to another and from one region within the world to another. In Africa, the main goal is establishing systems that can provide students with services like e-library. From the perspective of the organizations themselves, the goal can be summarized in reducing the cost and improve their IT capabilities. At the same time, fear and uncertainty still exist from applying the CC in the HE system. Taking into consideration the risks and struggles in adopting the CC in the HE system in Africa and comparing them to the benefits, the applying of CC is inevitable [13].

The education system will make it possible for teachers to highlight the weakness areas where the students used to make mistakes, this activity can be done through the students’ records analyzing. This analysis will enable teachers to improve or even change their teaching methods. Applying this technique will allow the students to have access to the lectures during the classes or even at home. Sharing learning materials and hardware (servers) by all the university colleges will reduce the operation cost for the universities effectively through the Utilization of cloud CC systems [30]. Furthermore, Fig. 4 shows the users as well as the way that these users interact with the CC.

3.4. The Employing of CC Service in HE in the Kurdistan Region – Iraq

The strategy of clouds must be related to the strategy of various academic institutions or universities. The transformation to CC- based establishments requires full knowledge about how it can function in different aspects and principles associated with organizational structure and relations between universities and institutions alongside with advantages and risks, security issues, and policies. Recently, the cloud CC researches illustrated the best usage practices of CC contain the following phases [31], [32].

- Evaluating the current level of the various institutions from the perspective of the IT requirements, framework,

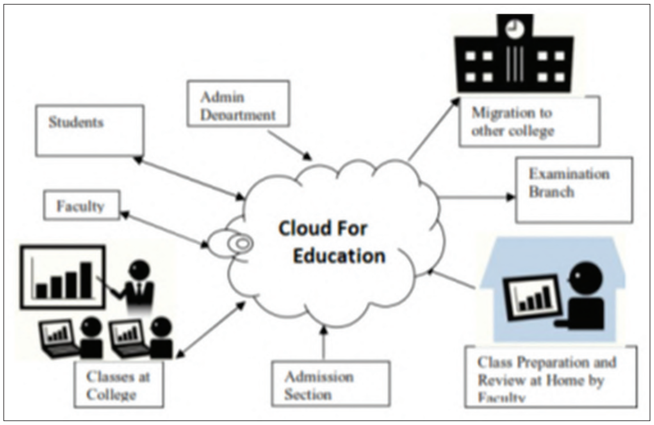


Fig. 4. Cloud computing service application in higher education's [30].

TABLE 4: Cloud computing types [43].

Type	Description
Public cloud	Services provided by organizations and customers pay for what they actually use, in terms of being cost-effective, public cloud is considered superior over the other, on the other hand, it raises other issues such as security, privacy, and levels of controls
Private cloud	Services provided to and managed by the organizations' staff themselves or any third party vendors, this cloud service is not provided to the general public, the private cloud could be implemented locally or remotely
Community cloud	Type of cloud provided to a specific target group of people, the services are shard exclusively among the members of this group only
Hybrid cloud	Combination of two or more cloud – the private clouds, public cloud, and the community

and usage: This step includes the understanding of various institutions' IT structure.

- Experimenting with the CC solutions: The applications and projects cannot be transformed to be cloud-based applications or projects suddenly, this shifting has to be step by step starting from experimenting with the CC technique on the pilot application and then apply it on other chosen applications. To do so, setting cloud goals just such as development and testing the environment or storing some data within the cloud and continue processing the internal processes is required [33], [34].
- Selecting the CC solution: Within this step, determining the data and applications, structure, functions, and core processes within the academic institutions is done. They may be grouped according to teaching, research, and administrative support. It also contains the cloud model which has been chosen (private, public, community, and hybrid) for the specified processes, functions, and applications [35], [36]. However, Table 4 shows the CC types.

3.5. Brief History of HE in Kurdistan

Kurdistan is a federal state located in the north of Iraq that has its own law and legislation, with a populace of 5.2 million and expanding the three governorates of Erbil, Slemani, and Duhok, cover roughly 40,000 km^2 [37]. KRG has realized how HE is important to upgrade the federal state infrastructure. KRG starts allocating a big portion slightly from its budget for the education field in general and HE in specific, just like in the 2013 budget where 16% of the budget was allocated to this issue [38]. Although, there was only one university in Kurdistan until 1992; in general, highly valued and has a special space in society. Gradually, the KRG policy has adopted huge investment in the H.E field which results in opening new HE institutes [39]. Kurdistan Region oversees 33 universities. These include 14 public and 15 state-recognized private universities, two universities with different ownership and two institutions [40], [41].

Some universities provide virtual computing labs which can be considered as a virtual environment where the students have the ability to reserves PCs [42]. The reserved PCs have their own specialized hardware as well as software. This environment enables the students when they have an internet service; to access those reserved computers from anywhere. Therefore, within the suggested technique, IaaS provides virtual machines (VMs) when they are needed for students of the university. The main reason for using these VMs is divided into two parts; the first one is to make courses and lab exercises. The other sub-reason is to build virtual labs. However, the advantages of using such an environment can be summarized in the ability of users (students and university staff) to use the resources. On the other hand, economic incomes can be achieved for the university. The work of Hashim *et al.* [12] showed that Bayan University achieved the task of using the system of cloud education through the technology of virtualization which illustrated the main point of virtual computing laboratory. This system allows the university to provide a flexible environment for their students to have access to the computers available within the university labs as well as reducing jams of using the computer hardware.

A new step in developing HE in Iraq has done when the MHE in Kurdistan applied an online registration system which enabled the students all over the republic of Iraq to select their universities and colleges. This action and many other effective steps done by the KGR, raise the number of the HE students to 94,700 according to latest statistics 48%. Of this number are female students. The academic degrees offered by the universities within the Kurdistan Region are diploma (2 years), bachelor (4 years), master (2 years), and doctorate (Ph.D.) (3-5 years) in multiple scientific and administrative academic fields and others [44].

The educational system in Kurdistan can be considered as an unstructured system; alongside using, the technologies are real problems to the MHE of Kurdistan. A survey has been conducted to determine using the extent of CC in KR universities; the questions within the survey (which included 222 academic staff and students from 14 universities) were varying to cover the requirements needed to apply the CC in their universities. The researchers in Ahmed *et al*. [3] highlights the main reasons to use the CC within the KR universities just such as reducing the cost, enhancing the university structure, develop the performance, and other related issues. On the other hand, the researchers illustrated the drawback and challenges confront applying the CC in KR just like lack of ICT infrastructure, security issues, privacy, and shortage of current systems, and data and documents ownership. To reduce the drawbacks of public cloud and get the advantages of CC HEIs and universities in KRG have to change their strategies from using public clouds to use their owned clouds [3].

4. CONCLUSION

The applying of CC in HE is providing many advantages such as steady, rapid, sample, suitability, and simultaneous accessibility of belongings at low cost in comparison with other techniques through the internet to the users. In this study, the existence of CC adoption frameworks for universities in developing countries along with Iraq has been discussed briefly. A review of these studies showed that the university within Kurdistan Region in Iraq needs continued attention to get government support, CC within Iraqi HE universities has limited developed over the latest years in the private and public universities. The findings indicate that interventions designed to increase the CC adoption need to include a focus on the practice level because that is decision-making regarding adoption occurs, in addition, to help IT, managers, within institutions to change their workflow to obtain the most services, along with addressing privacy concerns and explicitly acknowledging.

In addition, the study offered a variety of university settings to ensure higher generalizability associated with the outcomes. All these results can be mainly relevant and timely concerning the decision maker who presently faces the obstacle of CC adoption in the Iraqi education environment. The limitations of this study include that there was single-source bias, as the collection of information was from secondary sources only. Furthermore, the study has more of a judgmental conclusion, as there is no post data assessment. HE universities should figure out how to rationalize their students' needs and priorities, applications, and their own premise information, and after that merge their framework accordingly. Finally, the most related work to this study has discussed to attempt to fill a gap in the current research to develop an adoption model that can help Iraqi HE universities to adopt CC. Therefore, it is recommended for future researchers to conduct a field survey by collecting primary data and conducting statistical tests on the variables implicated in the findings of this study. Furthermore, due to the bold role of the internet and cyberspace in human life and its impact on behavior, lifestyle, it is suggested in future works monitor the role of social media in the use of CC in education.

REFERENCES

[1] A. H. Masud, X. Huang and J. Yong. "Cloud Computing for Higher Education: A Roadmap". In: 2020: *International Conference on Computer Supported Cooperative Work in Design*, pp. 552-557, 2012.

[2] Z. A. Ahmed and M. I. Ghareb. "An online course selection system: A proposed system for higher education in Kurdistan region government". *International Journal of Scientific and Technology Research*, vol. 7, no. 8, pp. 145-160, 2018.

[3] Z. A. Ahmed, A. A. Jaafar and M. I. Ghareb. "The ability of implementing cloud computing in higher education-KRG". *Kurdistan Journal of Applied Research*, vol. 2, pp. 39-44, 2017.

[4] Q. K. Kadhim, R. Yusof, H. S. Mahdi, S. S. Al-shami and S. R. Selamat. "A review study on cloud computing issues". *Journal of Physics: Conference Series*, vol. 2018, p. 12006, 2018.

[5] V. H. Pardeshi. "Cloud computing for higher education institutes: Architecture, strategy and recommendations for effective adaptation". *Procedia Economics and Finance*, vol. 11, pp. 589-599, 2014.

[6] M. T. Amron, R. Ibrahim and S. Chuprat. "A review on cloud computing acceptance factors". *Procedia Computer Science*, vol. 124, pp. 639-646, 2017.

[7] M. N. Qadri and S. Quadri. "A study of mapping educational institute with cloud computing". *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 2. pp. 59-66, 2017.

[8] M. Al Rawajbeh, I. Al Hadid, J. Aqaba and H. Al-Zoubi. "Adoption of cloud computing in higher education sector: An overview". *Indian Journal of Science and Technology*, vol. 5, no. 1, pp. 23-29, 2019.

[9] K. H. Al-Shqeerat, F. M. Al-Shrouf and H. Fajraoui. "Cloud computing security challenges in higher educational institutions-a survey". *International Journal of Computer Applications*, vol. 161, pp. 22-299, 2017.

[10] U. Singh and P. K. Baheti. “Role and service of cloud computing for higher education system”. *International Research Journal of Engineering and Technology*, vol. 9, p. 10, 2017.

[11] T. A. Kadhim. “Development a teaching methods using a cloud computing technology in Iraqi schools”. *Journal of University of Babylon*, vol. 26, pp. 18-26, 2018.

[12] E. W. A. Hashim, M. O. Hammood and M. T. I. Al-azraqe. “*A Cloud Computing System Based Laborites’ Learning Universities: Case Study of Bayan University’s Laborites-Erbil*”. Book of Proceeding, p. 538, 2016.

[13] A. S. Abdusalam, D. Faiq Abd, Z. A. Hamid, Z. A. Kakarash and O. H. Ahmed. “Study of challenges and possibilities of building and efficient infrastructure for Kurdistan Region of Iraq”. *UHD Journal of Science and Technology*, vol. 2, pp. 15-20, 2018.

[14] M. Al-Hashimi, M. Shakir, M. Hammood and A. Eldow. “Address the challenges of implementing electronic document system in Iraq e-government-Tikrit city as a case study”. *Journal of Theoretical and Applied Information Technology*, vol. 95, pp. 3672-3683, 2017.

[15] H. Abdulkadhim, M. Bahari, H. Hashim and A. Bakri. “Prioritizing implementation factors of electronic document management system (EDMS) using topsis method: A case study in Iraqi government organizations”. *Journal of Theoretical and Applied Information Technology*, vol. 88, pp. 375-378, 2016.

[16] T. H. Thabit and S. A. Harjan. “Evaluate e-learning in Iraq applying on Avicenna center in Erbil”. *European Scientific Journal*, vol. 11, pp. 1-14, 2015.

[17] S. Riaz and J. Muhammad. “An evaluation of public cloud adoption for higher education: A case study from Pakistan”. In: *Mathematical Sciences and Computing Research* (*iSMSC*), *International Symposium*, pp. 208-213, 2015.

[18] M. W. Nofan and A. A. Sakran. “The usage of cloud computing in education”. *Iraqi Journal for Computers and Informatics*, vol. 42, pp. 68-73, 2016.

[19] K. C. Ariwa and E. Aiwa. “Engineering sustainability and cloud computing in higher education-a case study model in Nigeria”. *International Journal of Computing and Network Technology*, vol. 5, pp. 65-75, 2017.

[20] J. Sultana, N. Nipa, and F. A. Mazmum. “Factors affecting could computing adoption in higher education in Bangladesh: A case of university of Dhaka”. *Applied and Computational Mathematics*, vol. 6, pp. 129-136, 2017.

[21] S. Başaran and G. O. Hama. “Exploring Faculty Members Views on Adoption of Cloud Computing in Education. In: *Proceedings of the International Scientific Conference*. vol. 5, p. 237, 2018.

[22] Z. Asadi, M, Abdekhoda and H. Nadrian. Cloud computing services adoption among higher education faculties: Development of a standardized questionnaire. *Education and Information Technologies*, vol. 25, no. 1. pp. 175-191, 2020.

[23] P. R. Maskare and S. R. Sulke. “Review paper on e-learning using cloud computing”. *International Journal of Computer Science and Mobile Computing*, vol. 3, pp. 1281-1287, 2014.

[24] M. M. Seke. “Higher education and the adoption of cloud computing technology in Africa.” *International Journal on Communications*, vol. 4, p. 1, 2015.

[25] M. S. Abdullah and M. Toycan. “Analysis of the factors for the successful e-learning services adoption from education providers’ and students’ perspectives: A case study of private universities in Northern Iraq”. *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 14, pp. 1097-1109, 2017.

[26] N. Sultan. “Cloud computing for education: A new dawn”? *International Journal of Information Management*, vol. 30, pp. 109-116, 2010.

[27] H. S. Hashim, K. Conboy and L. Morgan. “Factors influence the adoption of cloud computing: A comprehensive review”. *International Journal of Education and Research*, vol. 3, pp. 295-306, 2015.

[28] A. O. Akande and J. P. Van Belle. “Cloud Computing in Higher Education: A Snapshot of Software as a Service”. In: *Adaptive Science and Technology, IEEE 6th International Conference*, pp. 1-5, 2014.

[29] M. S. Al-Khayat and M. S. Al-Othman. “A proposed cloud computing model for Iraqi’s engineering colleges and institutes”. *ZANCO Journal of Pure and Applied Sciences*, vol. 28, pp. 1-5, 2016.

[30] P. Darus, R. B. Rasli and N. Z. Gaminan. “A review on cloud computing implementation in higher educational institutions”. *International Journal of Scientific Engineering and Applied Science*, vol. 1, pp. 459-465, 2015.

[31] A. Barnwal, D. Kumar. “Using cloud computing technology to improve education system”. *Asian Journal of Technology and Management Research*, vol. 4, pp. 68-72, 2014.

[32] D. F. Fithri, A. P. Utomo, and F. Nugraha. “Implementation of SaaS cloud computing services on E-learning applications (case study: PGRI foundation school)”. *Journal of Physics: Conference Series*, vol. 1430, no. 1, p. 012049.

[33] T. Bozzelli. “*Will the Public Sector Cloud Deliver Value? Powering the Cloud Infrastructure*”. Available from: http://www.cisco.com/web/strategy/docs/gov/2009_cloud_public_sector_tbozelli.pdf. [Last accessed on 2010 Oct 05].

[34] K. Njenga, L. Garg, A. K. Bhardwaj, V. Prakash and S. Bawa. “The cloud computing adoption in higher learning institutions in Kenya: Hindering factors and recommendations for the way forward”. *Telematics and Informatics*, vol. 38, pp. 225-246, 2019.

[35] I. Arpaci. “A hybrid modeling approach for predicting the educational use of mobile cloud computing services in higher education”. *Computers in Human Behavior*, vol. 90, pp. 181-187, 2019.

[36] M. R. Mesbahi, A. M. Rahmani and M. Hosseinzadeh. “Reliability and high availability in cloud computing environments: A reference roadmap”. *Human-Centric Computing and Information Sciences*, vol. 8, no. 1, pp. 20, 2018.

[37] A. A. Jaffar, M. I. Ghareb and K. F. Sharif. “The challenges of implementing E-commerce in Kurdistan of Iraq”. *Journal of University of Human Development*, vol. 2, 2016.

[38] R. Avci and N. Doghonadze. “The challenges of teaching EFL listeningin Iraqi (Kurdistan Region) Universities”. *Universal Journal of Educational Research*, vol. 5, pp. 1995-2004, 2017.

[39] D. S. Atrushi and S. Woodfield. “The quality of higher education in the Kurdistan Region of Iraq”. *British Journal of Middle Eastern Studies*, vol. 14, pp. 11-16, 2018.

[40] N. Ahmed. “*Performance Appraisal in Higher Education Institutions in the Kurdistan region: The case of the University of Sulaimani*”. Cardiff Metropolitan University, Wales, 2016.

[41] S. Razzaghzadeh, A. H. Navin, A. M. Rahmani and Hosseinzadeh, M. “Probabilistic modeling to achieve load balancing in expert clouds”. *Ad Hoc Networks*, vol. 59, pp. 12-23, 2017.

[42] H. P. Breivold and I. Crnkovic. “Cloud Computing education strategies”. In: *IEEE 27th Conference Software Engineering Education and Training*, pp. 29-38, 2014.

[43] M. A. Wahsh and J. Dhillon. “A systematic review of factors affecting the adoption of cloud computing for E-government

implementation". *Journal of Engineering and Applied Sciences*, *ARPN Journal of Engineering and Applied Sciences*, vol. 23, pp. 17824-17832, 2015.

[44] J. F. Kakbra and H. M. Sidqi. "Measuring the impact of ICT and e-learning on higher education system with redesigning and adapting MOODLE system in Kurdistan Region Government, KRG-Iraq". In: *Proceedings of the 2nd e-learning Regional Conference State of Kuwait*, p. 13, 2013.

REVIEW ARTICLE

UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Review of Properties and Functions of Narrowband Internet of Things and its Security Requirements

Zana Azeez Kakarash[1,2], Farhad Mardukhi[2]

[1]Department of Information Technology, University of Human Development, Sulaymaniyah, Iraq, [2]Department of Computer Engineering and Information Technology, Faculty of Engineering, Razi University, Kermanshah, Iran

ABSTRACT

Internet of Things (IoT) is a new web sample based on the fact that there are many things and entities other than humans that can connect to the Internet. This fact means that machines or things can automatically be interconnected without the need for interacting with humans and thus become the most important entities that create Internet data. In this article, we first examine the challenges of IoT. Then, we introduce features of NB-IoT through browsing current international studies on Narrowband IoT (NB-IoT) technology, in which we focus on basic theories and key technologies, such as the connection number analysis theory, the theory of delay analysis, the coating increase mechanism, low energy consumption technology, and the connection of the relationship between signaling and data. Then, we compare some functions of NB-IoT and other wireless telecommunication technologies in terms of latency, security, availability, and data transfer speed, energy consumption, spectral efficiency, and coverage area. Finally, we review and summarize NB-IoT security requirements that should be solved immediately. These topics are provided to overview NB-IoT which can offer a complete familiarity with this area.

Index Terms: Internet of Things, Narrow Band, Internet of Things, Narrowband Internet of Things

1. INTRODUCTION

Internet of Things (IoT) is a long-term stream that we are currently at its earliest stage. We can consider three primary phases to achieve the first phase of IoT. In the first phase, things can be identified for us and others and gradually assign a specific address on the network for themselves. In this phase, each object keeps certain information in it, but these are people who need to take out this information using tools like their smartphones [1], [2], [3]. In the second phase, each device has the ability to send information to the user at a specified time. After completing the relationship between objects and humans, it is time to relate things to each other. In the third phase, objects are associated with each other without human interference. Completing these three phases will finish the first phase of IoT evolution [4], [5].

Access this article online

DOI:10.21928/uhdjst.v4n1y2020.pp71-80	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

At the end of the first phase, there is a world of ideas in front of developers. The problem is that each device has some information that is available on the network by other objects and its owner and developers can use their own creativity to make better use of this information; Telecommunication networks communicate with each other based on technologies, spectra, and different frequency band. This technology in recent years has been more widely considered with the advent of IoT technology/Internet of

Corresponding author's e-mail: Zana Azeez Kakarash, Department of Information Technology, University of Human Development, Sulaymaniyah, Iraq. E-mail: zana.azeez@uhd.edu.iq

Received: 07-03-2019 Accepted: 20-03-2020 Published: 22-03-2020

everything and the expansion of devices and communication networks with specific requirements [6], [7], [8].

Narrowband IoT (NB-IoT) is a low power radio network (low consumption) in a wide range (low power wide area network [LPWAN]), which is designed and developed to allow the connection of a large number of devices or services using cellular telecommunication band (cellular network) [9], [10].

The NB of IoT focuses on network coverage in a closed space, less cost, and more battery life and has the ability to connect a large number of connected devices. The NB technology of IoT can be found in the spectrum in-band of the long-term evolution (LTE) network or the fourth generation in the frequency blocks of a fourth-generation operator or unused blocks (guard band) of a fourth-generation operator. It can also be used alone for the deployment of a specific range. It is also appropriate for new combinations (re-farming of [global system for mobile (GSM) communication] spectrum) [11], [12].

The NB was first introduced and developed by Sig Fax (2009). This company faced the 3rd generation partnership project (3GPP) institute, which defines cellular/mobile telecommunication standards with three challenges which have the ability to answer with a NB. The challenge is that there is a vibrant market for devices that:

1. Do not have a lot of abilities
2. They want to be very cheap
3. They have a low power consumption
4. Require high range (cover).

It can be said that the NB of IoT can exist in the following three conditions:

- Completely independent network
- In unused bands of 200 kHz, which previously used in GSM networks
- Second and third generations of mobile/communications
- At fourth-generation stations that can assign a block (frequency) to NB of the IoT or can be placed in (guard band) [13], [14], 15].

Finally, it can be said that the establishment of a NB of a network of IoT depends on the geographic conditions of the country and region as well as facilities and conditions of telecommunication and mobile operators of these countries. For example, in the United States, Verizon companies (Verizon and AT and T) can use LTE-M1 because both companies have invested in their fourth generation of the network; therefore, they probably do not want to create an independent network, and they want to have a network based on their current fourth-generation network [13], [14].

In front of areas of the world that have a wider GSM network than the fourth-generation network, it is rational to use an independent NB-IoT network. For example, T-Mobile operators in the United States and Sprint eventually have turned their attention toward the deployment of a NB network of IoT on the frequency spectrum of GSM network [13], [14], [15].

This paper recommends NB-IoT applicable models for application in many places to solve many problems (smart white goods, smart coordination’s, smart power metering, and smart road lighting) and provides a comprehensive overview of the design changes brought in the NB-IoT standardization along with the detailed research advancements from the viewpoints of security requirements and the practical presentation of NB-IoT as far as successful throughput.

The rest of the paper is organized as follows. Section 2 describes some background concepts relevant to our review. Section 3 describes the challenges of IoT. Significant features of NB-IoT are described in Section 4. Section 5 presents NB-IoT and different wireless communication technologies. In Section 7 describes basic requirements for NB-IoT security and in Section 8 discusses the conclusion.

2. BACKGROUND

2.1. Brief Review of NB-IoT

NB-IoT is a guideline based low control wide zone (LPWA) innovation created to empower a wide scope of new IoT gadgets and administrations. NB-IoT essentially improves the power utilization of client gadgets, framework limit, and range effectiveness, particularly in profound inclusion. The battery life of beyond what 10 years can be upheld for a wide scope of utilization cases.

New physical layer flag and channels are intended to meet the requesting necessity of broadened inclusion – rustic and profound inside – and ultra-low gadget multifaceted nature. The introductory expense of the NB-IoT modules is required to be tantamount to GSM/General Packet Radio Services (GPRS). The basic innovation is anyway a lot more straightforward than the present GSM/GPRS and its expense is relied on to diminish quickly as interest increments.

By supporting all major equipment such as mobile equipment, chipset, and module producers, NB-IoT can exist together with 2G (second-generation), 3G (third-generation), and 4G (forth-generation) versatile systems. It likewise profits by all the security and protection highlights of versatile systems, for example, support for client character classification, element confirmation, privacy, information respectability, and portable hardware distinguishing proof.

2.2. Benefits and Constraints of NB-IoT

The main properties on NB-IoT technology, as defined in Rel-13 3GPP TR 45.820 [10], are given in Table 1.

We have to survey the basic points of interest and consequent restrictions in regards to the inalienable capacities of the NB-IoT innovation to investigate the end-gadget activity and its incorporation with the IoT application [11], [12], [13], [14]. As planned ease of NB-IoT module presents no requirements and just brings benefits contrasting with other LPWA arrange arrangements, it would not be talked about further.

2.2.1. Wide coverage and deep signal penetration

This component gives a chance to the new application class of indoor and underground applications which incorporate information securing and control of gear situated in sewer vents, cellars, pipelines, and different conditions in which the existing correspondence foundation is inaccessible. Regardless of the improvement of sign entrance, the gadgets are relied upon to work on the lower limits of signature gathering. Hence, support for the vehicle of dependable information ought to be given as a piece of the availability arrangement.

2.2.2. Low power consumption of NB-IoT modules

The chance of battery-controlled structure or potential vitality collecting for end-gadget arrangements, which brings about long life remain solitary activity, is considered as the quick advantage of the low force property. Since gadgets are required to work for quite a while, at that point, reconfigurability is an ideal limit which features the requirement for sporadic, however, solid two-way correspondence. The two-way correspondence necessity is likewise seen by 3GPP in their rush hour gridlock model.

TABLE 1: NB-IoT main properties [42]

Range	<35 km
Battery life	>10 years
Frequency bands	LTE bands
Bandwidth	200 kHz or shared
Modulation	DL: OFDMA with 15 kHz subcarrier spacing UL: Single tone transmissions – 3.75 and 15 kHz, multi-tone SC-FDMA with 15 kHz subcarrier spacing
Max throughput	<56 kbps UL, <26 kbps DL
Link budget	164 dB
Capacity	+50 k IoT devices per sector

NB-IoT: Narrowband Internet of Things, OFDMA: Orthogonal frequency division multiple access, LTE: Long-term evolution, SC-FDMA: Single carrier frequency division multiple access

2.2.3. Massive connectivity

The inactive limit of NB-IoT supporting foundation is the gigantic availability coming about in up to 50 k gadgets per cell, which relies on inclusion mode and traffic blend gadgets are utilizing. Since a huge number of gadgets are proposed to be coordinated into conveyed applications, unbounded remote help reaction time is normal, which is considered as one of the regular issues progressively enormous scope combinations. The correspondence measurements which are influenced incorporate the persistence of information correspondence, models for automatic repeat demand and stream control, and guaranteed unwavering quality (nature of administration) [38], [39], [40], [41].

3. CHALLENGES OF THE IOT

On IoT, we face a world in which makers supply their goods with their standards, and it is not clear, with the continuation of this variety, billions of devices that make up IoT, where will lead future of networks. We examine two challenges of IoT in this section. One of them is standard conflicts, and the other one is the security that puts the future of IoT in disorderly conditions [16], [17].

3.1. Lack of Standard Unit

The IoT of today has a different world. When the Internet standards were created, people controlled this standard that their true desire was to formulate global standards. Standards are equally accessible to everyone, but the Internet of today is in control of companies that each wants to use these standards to defeat competitors and benefit from them. Furthermore, the Internet is in the hands of governments that basically want to supervise everything. How do governments and companies in this situation want to agree on global standards? In the IoT, standard means everything.

Each device must announce to other devices what it wants to do. Without these standards, they cannot do any of these. Add this truth to challenge that equipments connected to IoT are very different and variant. Many companies and organizations try to set standards, and all see union, industrial Internet consortium, IPSO Union, and the open interconnect consortium are of the main institutions. In the IoT landscape, there are not spots at which all agree over a series of global standards [15], [16], [17], [18].

3.2. Security

A recent discovery of a bug called Bash or Shellshock uncovered a serious security issue on the IoT. The bug is a bunch of codes that allow hackers to run on UNIX and Linux operating systems, as shown in Fig. 1.

The bug is announced by the National Institute of Standards and Technology as a high-level security threat. The seriousness of the threat comes from the fact that hackers do not need to have prior knowledge of the attacked system before they add their code to the Bash bug. The bug does not affect the IoT only, but all devices connected to it are at risk of being attacked. Devices that are attacked by the bug remain to be uncatchable and vulnerable. This discovered threat suggests that there might be many unaddressed security issues, which is good news to hackers and Internet criminals and raise questions about the effectiveness and usability of IoT in the future.

Another aspect of IoT as contributing to security issues is its complexity, which makes it hard to identify security gaps. These gaps have been realized by researchers, as they have concluded that the connected world has many hidden risks that require intensive research to find suitable solutions [18], [19], [20]. Many devices through various channels can connect to IoT, and as yet no mechanism has been put forward to alert device users of security threats and the way they can prevent attacks from bash-like bugs.

4. SIGNIFICANT FEATURES OF NB-IOT

NB-IoT is another rapidly developing remote connectivity 3GPP cell innovation standard introduced in Release 13 that corresponds to IoT's preconditions for the LPWAN. It is developing rapidly as the top-level driving innovation in LPWAN to enable a wide range of new IoT devices, including smart parking, utilities, wearables, and modern facilities.

Main features of NB-IoT are shown in Fig. 2 and briefly described below:

4.1. Low Energy Consumption

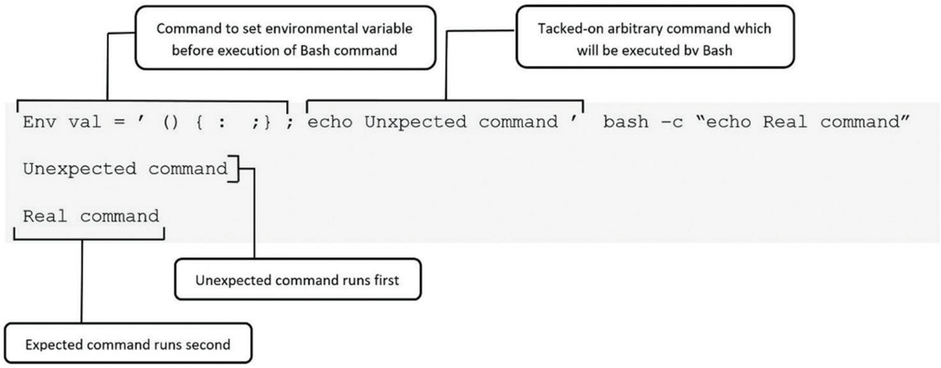
Using power saving mode (PSM) and infrequently developed receive (extended discontinuous receive [e-DRX]) Longer standby time can be observed in NB-IoT. In this context, PSM technology has been added lately to ReL12, in which terminal power-saving mode is still being recorded online, but it cannot achieve to saving energy by sending a signal to put the terminal in a deep sleep for a longer time [20], [21].

4.2. Improved Coverage and Low Latency Sensitivity

Given the reproduced information TR45.820, it very well may be affirmed that the intensity of the covering NB-IoT can find a good pace autonomous arrangement mode. Recreation try for both in-band organization and watchman band sending is finished. So as to advance the inclusion, systems, for example, remobilization (multiple times) and low recurrence tweak by NB IoT was endorsed. At present, NB-IoT support from quadrature amplitude modulation 16 is still under discussion. To lose blending 164 dB, if a dependable information move gave, due to re-change of mass information, dormancy increments [13], [14], [15], [16], [17], [18].

4.3. Transition Mode

As it is shown in Table 2, NB-IoT development is based on LTE. Correction is mainly based on LTE-related technologies due to unique NB-IoT features. Radiofrequency bandwidth from NB-IoT physical layers is 200 kHz. At the bottom link, NB-IoT with quadrature phase-shift keying (QPSK) modem and orthogonal frequency-division multiple access technologies is compatible with a distance under carrier 15 kHz. In the uplink, binary phase-shift keying or QPSK modem and single-carrier frequency division multiple access innovations, including single sub-bearer and different



```
Env val = ' () { :  ;} ; echo Unxpected command '  bash -c "echo Real command"

Unexpected command

Real command
```

Command to set environmental variable before execution of Bash command
Tacked-on arbitrary command which will be executed bv Bash
Unexpected command runs first
Expected command runs second

Fig. 1. How the function of code bash is vulnerable in the environment.

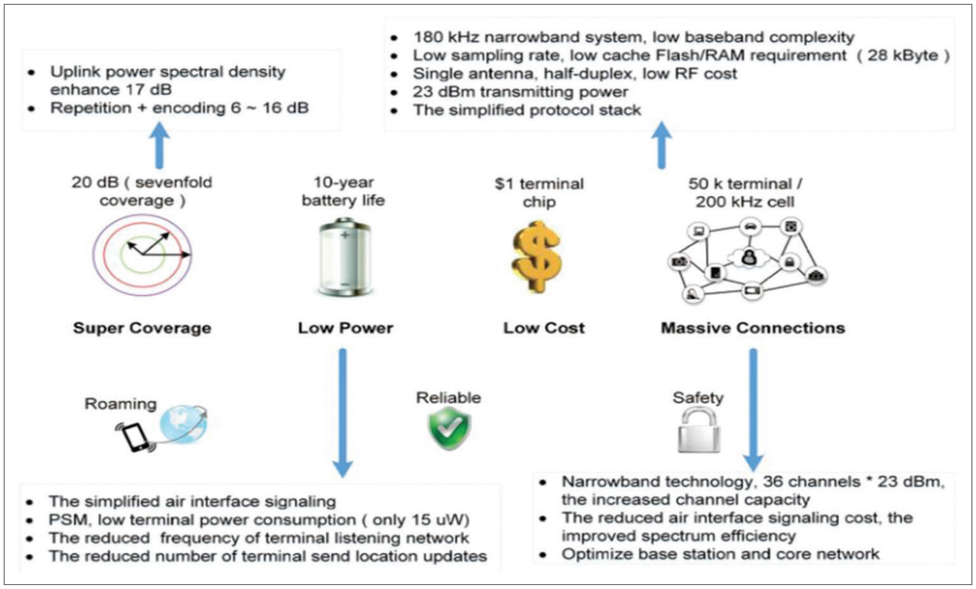

Uplink power spectral density enhance 17 dB
Repetition + encoding 6 ~ 16 dB
180 kHz narrowband system, low baseband complexity
Low sampling rate, low cache Flash/RAM requirement (28 kByte)
Single antenna, half-duplex, low RF cost
23 dBm transmitting power
The simplified protocol stack
20 dB (sevenfold coverage)
10-year battery life
$1 terminal chip
50 k terminal / 200 kHz cell
Super Coverage
Low Power
Low Cost
Massive Connections
Roaming
Reliable
Safety
The simplified air interface signaling
PSM, low terminal power consumption (only 15 uW)
The reduced frequency of terminal listening network
The reduced number of terminal send location updates
Narrowband technology, 36 channels * 23 dBm, the increased channel capacity
The reduced air interface signaling cost, the improved spectrum efficiency
Optimize base station and core network

Fig. 2. Main features of Narrowband Internet of Things [42].

TABLE 2: Main NB-IoT technical characteristics

Layer	Technical feature		
Physical layer	Uplink	BPSK or QPSK modulation	
		SC-FDMA	Single carrier, the subcarrier interval is 3.75 kHz and 15 kHz the transmission rate is 160 kbit/s – 200 kbit/s Multi-carrier, the subcarrier interval is 15 kHz, the transmission rate is 160 kbit/s – 250 kbit/s
	Downlink	QPSK modulation	
		OFDMA, the subcarrier interval is 15 kHz, the transmission rate is 160 kbit/s – 250 kbit/s	
Upper layer	LTE-based protocol		
Core network	S1 interface based		

BPSK: Binary phase-shift keying, NB-IoT: Narrowband Internet of Things, QPSK: Quadrature phase shift keying, LTE: Long-term evolution, OFDMA: Orthogonal frequency division multiple access, SC-FDMA: Single carrier frequency division multiple access

subcarrier, are embraced. A solitary sub-bearer innovation with the sub-bearer separating of 3.75 kHz and 15 kHz is appropriate to IoT terminal with ultra-low rate and ultra-low force utilization. The convention of NB-IoT high layer (the layer above physical layer) is figured through modification of a few LTE highlights, for example, multi-association, low force utilization also, not many information. The center system of NB-IoT is associated through S1 interface [16], [17], [18], [19], [20], [21], [22].

4.4. Spectrum Resources

IoT is a core service that attracts a larger user group in the communication services market for the future. Hence, NB-IoT development supported by four major telecom operators in China, as shown in Table 3, which is the owner of FVHD NB-IoT relevant spectrum source.

4.5. Deployment Supported by NB-IoT

According to RP-151621 regulations, NB-IoT is currently only foreign demand draft transfer mode with a bandwidth of 182 kHz and three types of deployment model shown in Fig. 3:

- Independent deployment (standalone mode), which utilizes a free recurrence band that has no cover with the LTE recurrence band
- Guard band deployment (protective band mode), which uses edge band frequency
- In-band deployment (in-band mode), which uses an LTE frequency band for deployment, and takes one physical resource block from LTE frequency band source for deployment [22], [23].

4.6. Structure and Framework

The bottom link in NB-IoT eNodeB supports from the wireless framework of E-UTRAN one frame structure

TABLE 3: Spectrum classification for NB-IoT by telecom operators

Operator	Uplink frequency band/MHz	Downlink frequency band/MHz	Bandwidth/MHz
China unicorn	909–915	954–960	6
	1745–1765	1840–1860	20
China telecom	825–840	870–885	15
China mobile	890–900	934–944	10
	1725–1735	1820–1830	10
SARFT	700	700	Undistributed

NB-IoT: Narrowband Internet of Things

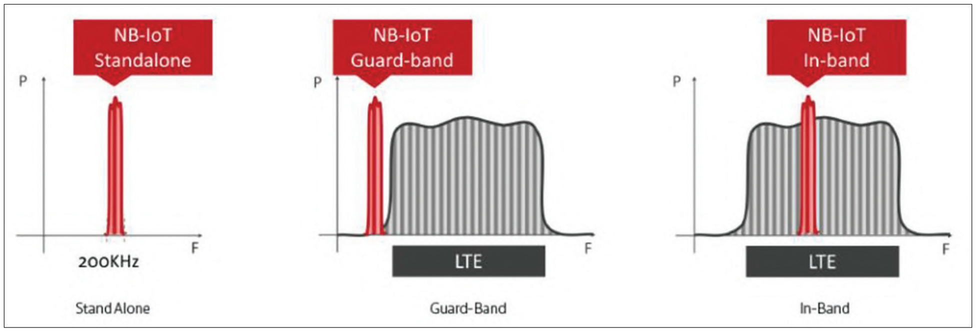


Fig. 3. Three deployments supported by Narrowband Internet of Things.

(FS1), as shown in Fig. 4. Upper link supports FS1 for under carrier spacing of 15 kHz. However, for spacing under carrier 3.75 kHz, a new type of framework is defined Fig. 5.

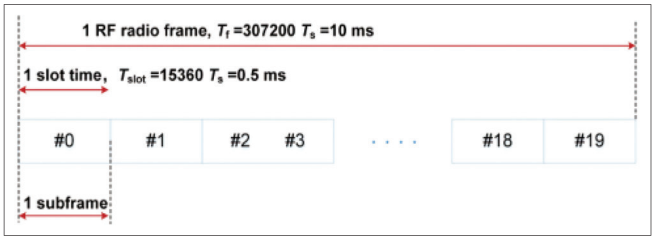

1 RF radio frame, T_f =307200 T_s =10 ms
1 slot time, T_{slot} =15360 T_s =0.5 ms
#0
#1
#2
#3
. . . .
#18
#19
1 subframe

Fig. 4. Narrowband Internet of Things framework structure for spacing under carrier of 15 kHz for upper and lower links [42].

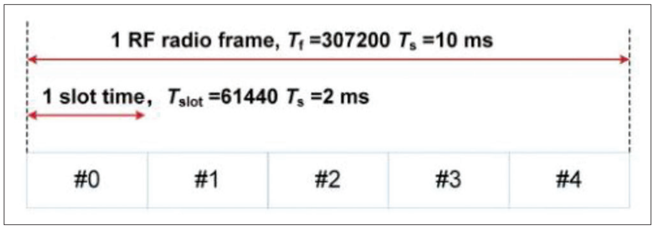

1 RF radio frame, T_f =307200 T_s =10 ms
1 slot time, T_{slot} =61440 T_s =2 ms
#0
#1
#2
#3
#4

Fig. 5. Narrowband Internet of Things framework structure for spacing under carrier of 3.75 kHz for upper link [42].

5. KEY TECHNOLOGY OF NB-IOT

5.1. Connection Analysis Theory

3GPP analyzes several connections that NB-IoT can access it when network supports from terminal periodic reporting service and network command reporting service. It is assumed that services are distributed within a day and NB-IoT can support 52547 connectivity per cell. Indeed, this assumption is too ideal, which almost ignores the business of NB-IoT service. As a result, it is difficult to generalize it in other application scenes. At present, there are few studies in NB-IoT business service. However, the research results of LTE-M (machine type communications [MTC]) and enhanced MTC are still valuable to learn. To overcome LTE network access overhead at a time a lot of MTC terminals enter the network at the same time, researchers have focused their analysis on LTE random access channel (RACH) load pressure and additional load control mechanisms. Researches typically coordinate service entering process as a homogeneous/hybrid process with the same distribution. The users retransitions the number of packet in queue head or channel position in a specific time slot as position variables for obtaining a stable graphical plot with the assumption of completing multi-channel S-ALOHA static mode performance analysis.. The graphing plan can be used for LTE RACH optimal design. However, when a lot of MTC terminals enter the network simultaneously, a large number of MTC terminals are sent simultaneously to the network to request a quick meeting in a short time to respond the same incident or monitoring the relevant components. This feature can be hardly described by classical homogeneous/hybrid Poisson process which forms direct application of network performance analysis method based on stable state hypotheses. Hence, a transient functional analysis method is essential for multi-channel S-ALOHA of non-Poisson services [24], [25], [26].

5.2. The Latency Analysis Theory

Besides the numbers of connecting analysis, 3GPP indicates that there is a need for a theoretical latency model capable of addressing the latency of synchronization, random access, resource allocation, and data transmission to access the upper link. Some of these latencies are concerned with signal detection and service behavior as researchers in the field have concentrated on mean and random-access latency variance and little attention has been paid to other curtail features such as probability density function (PDF) of latency. Researchers such as Rivero-Angeles *et al.* [26] and Liu *et al.* [27] have used the Markov process to produce probability generation function from PDF. However, the complex nature of computing has made it difficult for researchers to find a mechanism to lower latency and increase communication probability.

5.3. Covering Enhancement Mechanism

Slender band adjustment and sub-GHz arrangement from NB-IoT can upgrade, getting affectability to build inclusion capacity. Besides, 3GPP suggests another advancement component dependent on coverage classes, which is another idea presented for NB-IoT by 3GPP.

5.4. Very Low Energy Technology

A major issue with IoT is energy consumption. Researchers have simulated the energy consumption for terminal services within NB-IoT with the aim to identify an area for improvements and the result showed that if the information is transmitted once a day, the life expectancy of a 5 Wh battery could be much prolonged. This leads to the suggestion that an evaluation mechanism for energy efficiency is required to ensure that lower energy consumption for IoT is achieved. Some researches, such as Liu *et al.* [27] and Balasubramanya *et al.* [28], on energy consumption in DRX focuses on single terminals between control signaling states and terminal operating mode. However, more work is needed to find a holistic mechanism that is seen as one of the tasks of 3GPP R14.

5.5. Connectivity between Signaling and Data

Coupling simulation between data and signaling is another concern in IoT that companies such as Huawei technology have indicated needs to be addressed. This is because in many simulation tools, data and signals are separated and simulation tests are done for each independently. This leads to a result where issues in connecting the two cannot be understood which makes it difficult to simulate real network capacity, for example, when access to MTC terminals are requested [29], [30].

6. NB-IOT AND DIFFERENT WIRELESS COMMUNICATION TECHNOLOGIES

LPWA technology is gaining popularity as IoT services grow rapidly. The technology is used to deliver smart services with low data speed, which can be utilized in IoT intelligent applications. These applications are classified into three groups by Hekwan Wu in the 2016 International Internet Conference in China, as shown in Table 4.

Fig. 6a illustrates the position of (LPWAN) in comparison to other communication technologies in terms of inclusion zone and information transmission rate. This type of technology is most suited to applications that require high bandwidth and short-range transmission speed such as Bluetooth and ZigBee [31], [32], [33].

TABLE 4: Distribution statistics for IoT smart connection technology in 2020

Global M2M/IoT connection distribution in 2020	Category	Network connection techniques	Fine-grained market opportunity
10%	High data rate (>10 Mbps), e.g., C- CTV, eHealth	3G: HSPA/EVDO/TDS 4G: LTE/LTE-A WiFi 802.11 technologies	Big profit margin for car navigation/ entertainment system
30%	Medium data rate (<1 Mbps), e.g., POS, smart home, M2M backhaul	2G: GPRS/CDMA2KIX MTC/eMTC	2G M2M could be replaced by MTC/ eMTC techniques
60%	Low data rate (<100 Kbps), e.g., sensors, meters, tracking logistics S-mart parking, smart agriculture	NB-IoT SigFox LoRa Short distance wireless connection, e.g., Zigbee	Various application cases; Main market for LPWA; Market vacancy

NB-IoT: Narrowband Internet of Things, GPRS: General packet radio services, eMTC: Enhanced machine-type communications

Fig. 6b shows the position of NB-IoT that makes use of both 4G/5G attributes and low power radio technology and advantages of low-energy consumption remote correspondence advances (e.g., ZigBee innovation) to be specific concentrated transmission and minimal effort.

We have further investigated the technology and compared it with LoRa, which is a type of WAN communication technology, as shown in Table 5.

7. REQUIREMENTS FOR NB-IOT SECURITY

Security requirements for NB-IoT are similar to that of traditional IoT with a number hardware, energy consumption, and network connection mode differences. Traditional IoT normally has a robust computing power with strong internal security design but with high energy consumption [34], [35], [36]. There are IoT technologies equipped with low-power hardware, but in return, it offers a low computing power with high-security risk which may lead to service denial. As a consequence, any security violation, even on a small scale, may leave a negative lasting effect as terminals are simpler and easier for attackers to obtain information. Researchers in Chen *et al.* [35], Li *et al.* [36], Mangalvedhe *et al.* [37], and Koc *et al.* [38] have analyzed NB-IoT security requirements, which is distributed over three layers, as shown in Fig. 7.

The below explanation introduces the security prerequisites of NB-IoT planning to the 3-layer design comprised perception layer, transition layer, and application layer.

7.1. Perception Layer

Perception layer is NB-IoT base layer that shows fundamental and establishment of administration and engineering higher layers. NB-IoT observation layer, for example, regular discernment layer, will, in general, be under latent and dynamic assaults. Uninvolved assault implies trespasser ransacks data with no redress. The fundamental highlights incorporate listening in, rush hour gridlock investigation, etc. As NB-IoT depended on an open remote system, trespassers may discover data about NB-IoT terminals with strategies, for example, information connect theft and traffic properties examination to focus on a progression of resulting assaults.

7.2. Transition Layer

Contrasted with the traditional layer in customary IoT, NB-IoT changes complex system organization that implies hand-off entryway gathers data and afterward sends it to the

TABLE 5: Comparison of NB-IoT and LoRa

Item	NB-IoT	LoRa
Power consumption	Low (l0 years battery life)	Low (l0 years battery life)
Cost	Low	Lower than NB-IoT
Safety	Telecom level security	Slight interference
Accuracy rate	High	High
Coverage	<25 km (resend supported)	<11 km
Deployment	Rebuild supported based on LTE FDD or GSM	Inconvenience

NB-IoT: Narrowband Internet of Things, GSM: Global system for mobile, FDD: Foreign demand draft

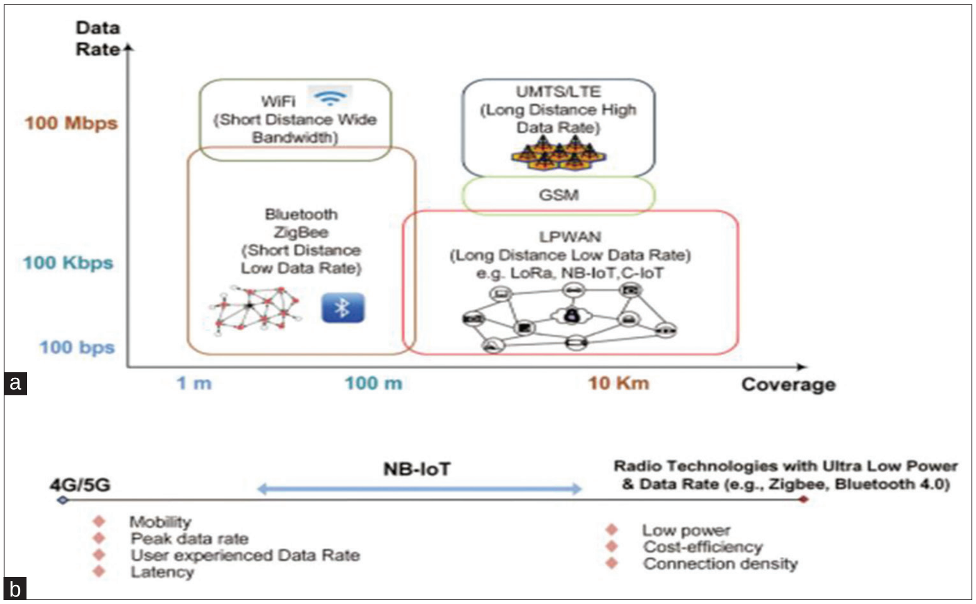

Data Rate
100 Mbps
100 Kbps
100 bps
WiFi
(Short Distance Wide Bandwidth)
UMTS/LTE
(Long Distance High Data Rate)
GSM
Bluetooth
ZigBee
(Short Distance Low Data Rate)
LPWAN
(Long Distance Low Data Rate)
e.g. LoRa, NB-IoT,C-IoT
a
1 m
100 m
10 Km
Coverage
4G/5G
NB-IoT
Radio Technologies with Ultra Low Power & Data Rate (e.g., Zigbee, Bluetooth 4.0)
Mobility
Peak data rate
User experienced Data Rate
Latency
Low power
Cost-efficiency
Connection density
b

Fig. 6. Correlation between Narrowband Internet of Things (NB-IoT) and different wireless communication technologies (a) comparison of various wireless communication technologies. (b) NB-IoT design exchanges.

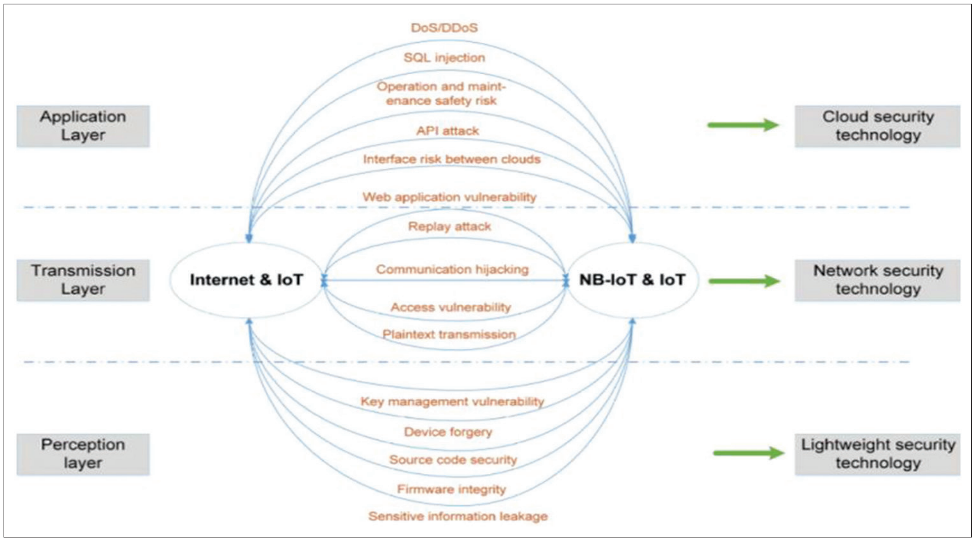


Fig. 7. Similarity between traditional Narrowband Internet of Things (IoT) and IoT in terms of security requirements.

base station for sustaining. Subsequently, numerous issues, for example, multi-organizing, significant expense, and battery with a high limit, are illuminated. A system for the entire city can carry simplicity of upkeep and the board with advantages, for example, advantageous tending to and establishment through detachment from property administration.

7.3. Application Layer

The purpose of the NB-IoT application layer is to store, analyze, and manage data efficiently. After the perception and transfer layer, a large amount of data converges in the application layer. Then, vast resources are formed to provide data support from different applications. Compared to the traditional IoT application layer, the NB-IoT application layer carries more data [37], [38], [39], [40].

8. CONCLUSION

In this paper, we reviewed the basic properties, benefits, and background and the latest scientific findings of NB-IoT. The general background of the IoT was introduced. The benefits, features, basic theory, and NB-IoT key technologies such as connection analysis, latency analysis, and coverage enhancement analysis were provided. Subsequently, we focused on differences between NB-IoT and different types of communication technologies. Finally, we made a comparison between NB-IoT and other wireless communication technologies and we examine NB-IoT security requirements from three levels; perception layer, transition layer, and application layer. There are many future research paths for this study. We continue to investigate a visible network model that can visually reflect the status of NB-IoT network operation. Such a model should complete each of the operational modules and do link-level open type simulation and NB-IoT confirmation form pellet.

REFERENCES

[1]. P. Reininger. “3GPP Standards for the Internet-of-Things”. Technologies Report, Huawei, Shenzhen, China, 2016.

[2]. “*Feasibility Study on New Services and Markets Technology Enablers for Massive Internet of Things*”. Document TR 22.861, 3GPP, 2016.

[3]. M. Chen, Y. Qian, Y. Hao, Y. Li and J. Song. “Data-driven computing and caching in 5G networks: Architecture and delay analysis”. *IEEE Wireless Communications*, vol. 25, no. 1, pp. 70-75, 2018.

[4]. 3GPP. “*Standardization of NB-IoT Completed*”, 2016. Available from: http://www.3gpp.org/news-events/3gpp-news/1785-nb_IoT_complete. [Last accessed on 2018 Oct 01].

[5]. A. Rico-Alvarino, M. Vajapeyam, H. Xu, X. Wang, Y. Blankenship, J. Bergman, T. Tirronen and E. Yavuz. “An overview of 3GPP enhancements on machine to machine communications”. IEEE Communications Magazine, vol. 54, no. 6, pp. 14-21, 2016.

[6]. Ericsson. “*Cellular Networks for Massive IoT*”. Technologies Report, Ericsson, Stockholm, Sweden, 2016.

[7]. Y. L. Zou, X. J. Ding and Q. Q. Wang. “Key technologies and application prospect for NB-IoT”. *ZTE Technology Journal*, vol. 23, no. 1, pp. 43-46, 2017.

[8]. A. Laya, L. Alonso and J. Alonso-Zarate. “Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives”. *IEEE Communications Surveys and Tutorials*, vol. 16, no. 1, pp. 4-16, 2014.

[9]. RIoT. “*Low Power Networks Hold the Key to Internet of Things*”. Technologies Report, Berlin, Germany, 2015.

[10]. X. Ge, X. Huang, Y. Wang, M. Chen, Q. Li, T. Han and C. X. Wang. “Energy-efficiency optimization for MIMO-OFDM mobile multimedia

communication systems with QoS constraints". *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2127-2138, 2014.

[11]. P. Osti, P. Lassila, S. Aalto, A. Larmo and T. Tirronen. "Analysis of PDCCH performance for M2M traffic in LTE". *IEEE Transactions on Vehicular Technology*, vol. 63, no. 9, pp. 4357-4371, 2014.

[12]. G. C. Madueno, Č. Stefanović and P. "Popovski. Reengineering GSM/GPRS Towards a Dedicated Network for Massive Smart Metering". In: *IEEE International Conference on Smart Grid Communications* (*SmartGridComm*), pp. 338-343, 2014.

[13]. W. Liu, J. Dong, N. Liu, Y. L. Chen, Y. B. Han and Y. B. Ren. "*NB-IoT key Technology and Design Simulation Method*". Telecommunications Science, China, pp. 144-148, 2016.

[14]. M. Centenaro and L. Vangelista. "A Study on M2M Traffic and Its Impact on Cellular Networks". In: *2015 IEEE 2nd World Forum on Internet of Things* (*WF-IoT*), pp. 154-159, 2015.

[15]. G. Y. Lin, S. R. Chang and H. Y. Wei. "Estimation and Adaptation for Bursty LTE Random Access", vol. 65. In: *IEEE Transactions on Vehicular Technology*, pp. 2560-2577, 2016.

[16]. Q. Xiaocong and M. Mingxin. "NB-IoT standardization technical characteristics and industrial development". *Information Research*, vol. 5, pp. 523-526, 2016.

[17]. M. T. Islam, M. T. Abd-elhamid and S. Akl. "A Survey of Access Management Techniques in Machine Type Communications". Vol. 52. IEEE Communications Magazine, Piscataway, pp. 74-81, 2014.

[18]. G. H. Dai and J. H. Yu. "Research on NB-Io T background, standard development, characteristics and the service". *Mobile Communications*, vol. 40, no. 7, pp. 31-36, 2016.

[19]. M. A. Khan and K. Salah. "IoT security: Review, blockchain solutions, and open challenges". *Future Generation Computer Systems*, vol. 82, pp. 395-411, 2018.

[20]. V. Kharchenko, M. Kolisnyk, I. Piskachova and N. Bardis. "Reliability and Security Issues for IoT-based Smart Business Center: Architecture and Markov Model". In: *2016 Third International Conference on Mathematics and Computers in Sciences and in Industry* (*MCSI*), pp. 313-318, 2016.

[21]. J. J. Nielsen, D. M. Kim, G. C. Madueno, N. K. Pratas and P. Popovski. "A Tractable Model of the LTE Access Reservation Procedure for Machine-type Communications". In: *2015 IEEE Global Communications Conference (GLOBECOM)*. pp. 1-6, 2015.

[22]. C. H. Wei, R. G. Cheng and S. L. Tsao. "Performance analysis of group paging for machine-type communications in LTE networks". *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 3371-3382, 2013.

[23]. M. Koseoglu. "Lower bounds on the LTE-A average random access delay under massive M2M arrivals". *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 2104-2115, 2016.

[24]. S. Persia and L. Rea. "Next generation M2M cellular networks: LTE-MTC and NB-IoT capacity analysis for smart grids applications". In: *2016 AEIT International Annual Conference* (*AEIT*), pp. 1-6, 2016.

[25]. T. M. Lin, C. H. Lee, J. P. Cheng and W. T. Chen. "PRADA: Prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks". *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2467-2472, 2014.

[26]. M. E. Rivero-Angeles, D. Lara-Rodriguez F. A. Cruz-Perez. "Gaussian approximations for the probability mass function of the access delay for different backoff policies in S-ALOHA". *IEEE Communications Letters*, vol. 10, no. 10, pp. 731-733, 2006.

[27]. J. Liu, J. Wan, B. Zeng, Q. Wang, H. Song and M. Qiu. "A scalable and quick-response software defined vehicular network assisted by mobile edge computing". *IEEE Communications Magazine*, vol. 55, no. 7, pp. 94-100, 2017.

[28]. N. M. Balasubramanya, L. Lampe, G. Vos and S. Bennett. "DRX with quick sleeping: A novel mechanism for energy-efficient IoT using LTE/LTE-A". *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 398-407, 2016.

[29]. K. Lin, D. Wang, F. Xia, H. Ge. "Device clustering algorithm based on multimodal data correlation in cognitive internet of things". *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2263-2271, 2018.

[30]. G. Naddafzadeh-Shirazi, L. Lampe, G. Vos and S. Bennett. "Coverage enhancement techniques for machine-to-machine communications over LTE". *IEEE Communications Magazine*, vol. 53, no. 7, pp. 192-200, 2015.

[31]. F. Xu, Y. Li, H. Wang, P. Zhang and D. Jin. "Understanding mobile traffic patterns of large scale cellular towers in urban environment". *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 1147-1161, 2017.

[32]. Y. Li, F. Zheng, M. Chen and D. Jin. "A unified control and optimization framework for dynamical service chaining in software-defined NFV system". *IEEE Wireless Communications*, vol. 22, no. 6, pp. 15-23, 2015.

[33]. X. Ge, J. Yang, H. Gharavi and Y. Sun. "Energy efficiency challenges of 5G small cell networks". *IEEE Communications Magazine*, vol. 55, no. 5, pp. 184-191, 2017.

[34]. X. Yang, X. Wang, Y. Wu, L. P. Qian, W. Lu and H. Zhou. "Small-cell assisted secure traffic offloading for narrow band internet of thing (NB-IoT) systems". *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1516-1526, 2018.

[35]. L. Chen, S. Thombre, K. Järvinen, E. S. Lohan, A. Alén-Savikko, H. Leppäkoski, M. Z. Bhuiyan, S. Bu-Pasha, G. N. Ferrara, S. Honkala and J. Lindqvist. "Robustness, security and privacy in location-based services for future IoT: A survey". *IEEE Access*, vol. 5, pp. 8956-8977, 2017.

[36]. Y. Li, X. Cheng, Y. Cao, D. Wang and L. Yang. "Smart choice for the smart grid: Narrow band Internet of Things (NB-IoT)". *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1505-1515, 2018.

[37]. N. Mangalvedhe, R. Ratasuk and A. Ghosh. "NB-IoT deployment Study for Low Power Wide Area Cellular IoT". In: *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications* (*PIMRC*), pp. 1-6, 2016.

[38]. A. T. Koc, S. C. Jha, R. Vannithamby and M. Torlak. "Device power saving and latency optimization in LTE-A networks through DRX configuration". *IEEE Transactions on Wireless Communications*, Vol. 13, no. 5, pp. 2614-2625, 2014.

[39]. R. Cheng, A. Deng and F. Meng. "*Study of NB-IoT Planning Objectives and Planning Roles*". China Mobile Group Design Inst. Co., Technical Reports Telecommunications Science, 2016.

[40]. Y. Hou, and J. Wang. "LS-SVM's No-reference Video Quality Assessment Model Under the Internet of Things". In: *2017 IEEE Smart World, Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation* (*Smart World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI*), pp. 1-8, 2017.

[41]. R. Aleksandar, P. Ivan, P. Ivan, B. Đorđe, S. Vlado and R. Miriam. "Key Aspects of Narrow Band IoT Communication Technology Driving Future IoT Applications". Conference: In: *2017 IEEE Telecommunication Forum* (*TELFOR*), 2017.

[42]. C. Min, M. Yiming, H. Yixue, A. K. HWANG. "Narrow band internet of things". *IEEE Access*, vol. 5, pp. 20557-20577, 2017.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Hematological Impacts of the Traffic Emissions in Sulaymaniyah Governorate

Dunya Hars Bapir[1], Salih Ahmed Hama[1,2]

[1]Department of Biology, College of Science, University of Sulaimani, Kurdistan Region, Sulaymaniyah, Iraq,
[2]Department of Medical Laboratory Sciences, College of Health Sciences, University of Human Development, Kurdistan Region, Sulaymaniyah, Iraq

ABSTRACT

The current study was achieved to evaluate the essential hematologic impacts of traffic emission. Ninety-six cases were studied that included both exposures and controls. The focal point was on Raparin District in Sulaymaniyah Governorates. A questioner form was depended for collecting the information about each case. Fresh venous blood (5 ml) was collected aseptically from both exposures and controls. Hematologic autoanalyzer (Coulter-Automated Counter) was used for hematologic investigations. It was appeared that the mean leukocyte counts were higher among exposures in comparison to controls; the period of exposure and smoking was significantly effective on total white cells. Lymphocyte counts were significantly declined among exposures. It was appeared that the distance from the emission gas sources, smoking, and period of exposure was significantly effective on the total lymphocyte counts ($P < 0.05$). No valuable effects of traffic emission were noticed on granulocytes in general ($P > 0.05$), although the neutrophil counts were significantly higher among exposure. Moreover, the study revealed that there were noticeable effects of traffic emission, on the total platelet counts between exposures and controls. Finally, the distance from the emission sources was significantly effective on platelet counts among exposures themselves ($P < 0.05$).

Index Terms: Traffic Emission, Hematology, Lymphocytes, Complete Blood Count

1. INTRODUCTION

The term pollution was defined as exposing to the harmful pollutants or products in the environment that appeared to have a measurable effect on the man or other animal health as well as on vegetation or other materials [1], [2]. There are five types of pollutants that are hydrocarbon, carbon monoxide, particulate matter, nitrogen dioxide, and sulfur oxides. These tend to be the worst quality content found in Iraqi fuel, which are emitted from the combustion of sulfur containing fossil fuels such as coal, metal smelting, motor vehicle operations, and other industrial processes. Urban air pollution is a significant cause of global mortality, pre-mature deaths, which are the causes of seven hundred thousand deaths worldwide according to data from the WHO [3], [4]. Several reports have indicated that exposure to aromatic hydrocarbons such as benzene, toluene, and styrene-butadiene has significant alterations in different hematologic parameters. The noticeable effects include a decline in circulating erythrocytes, hemoglobin (HGB), platelets, total white blood cells (WBCs), and absolute numbers of lymphocytes, as well as neutrophils [4]-[7]. The adverse effects may be on bone marrow and stem cells at both production and differentiation levels. Moreover, it may have effects on Hepcidin's sustained and chronic upregulation that is an iron regulatory protein, which may lead to HGB and red blood cell (RBC) production diminishing. Consequently, anemia can occur [8]-[12]. Leukopenia, thrombocytopenia, and reduction in bone marrow-derived mesenchymal stem cells also may be common side effects [13].

Access this article online

DOI:10.21928/uhdjst.v4n1y2020.pp81-86

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: dunyabiologist064@gmail.com/dunya.bapir94@gmail.com

Received: 07-03-2020 Accepted: 23-04-2020 Published: 25-04-2020

The aims of the current study were to investigate the significant traffic emission impacts of various hematological parameters and to study the effect of some risk factors and their relations to traffic emissions hematologic consequences.

2. MATERIALS AND METHODS

Ninety-six persons (males, and females) were studied included (48 exposes and 48 control cases), both sexes involved. The laboratory investigations were done from May 15, 2019, to September 25, 2019. The hematological tests were performed in Azadi Laboratory in Ranya city. Five fresh venous blood were collected from all cases and directly transferred to the lab for investigations. The hematologic examinations were included in the study; leukocyte profiles (total WBC, granulocyte, neutrophil, and lymphocyte) counting. Red cell profiles (RBC, red cell distribution width [RDW], hematocrit [HCT], HGB, mean Cell HGB MCH, mean cell HGB concentration mean corpuscular hemoglobin concentration [MCHC], and mean cell volume [MCV]) counting, as well as the platelet profiles (platelet [PLT], mean platelet volume [MPV], platelet distribution width [PDW], plateletcrit [PCT], and large platelet cell ratio [LPCR]) counting.

Three factors were studied for their relation with the emission impacts on the studied cases, which are exposure period (short-term – <10 years – and long-term – more than 10 years); distance from the emission sources (<500 m and more than 500 m); and finally smoking (smokers and non-smokers).

An automated hematologic analyzer (Coulter; KT6200, of OEM) was depended in achieving the above tests. The obtained data were tabulated and statistical analyses were done using GraphPad prism 6 software (Mann–Whitney *t*-test).

3. RESULTS AND DISCUSSION

From the current study, it was appeared that different hematologic parameters were affected negatively by exposure to the chemical compounds produced from the traffic emissions.

3.1. WBC Measurement WBC

Studying total WBC counts revealed that the mean value of WBC counts of exposures was 7100 cells/μl ±1.5, whereas the mean value among control cases was (6780 cells/μl ±1.8), which was lower than that of exposures (Fig. 1). Statistical analysis showed that there were no significant differences between the total WBC counts from exposures and controls ($P > 0.05$).

Due to further analysis, there were significant differences between those with (5–10 years) exposure history (mean=6700 cells/μl ±1), and those with more prolonged exposure (10–20 years), (mean=7400 cells/μl ±1.9) ($P < 0.05$). Furthermore, significant differences were observed among smoker exposures (mean=7500 cells/μl ±2.1) and non-smoker exposures (mean=6800 cell/μl ±1.2) ($P < 0.05$). Moreover, it was reported by the current study that the total numbers of lymphocytes were also affected by exposure to traffic emissions. It was noticed that the mean value of the lymphocyte counts among exposures was lower (1100 cells/μl ±3.7), when compared with controls (1900 cells/μl ±3.9). There was a significant difference between exposures and controls regarding the mean value of lymphocyte count levels ($P < 0.05$) (Fig. 1). The mean value of lymphocytes among exposures whose home distances about 100–200 m far from traffic contamination sources were 1052 cells/μl ±3.6, while it was slightly higher (1107 cells/μl ±3.7) among exposures whose home distance was far from the first group (500–1000 m) from the sources of traffic gases. It was appeared that the distance from the emission

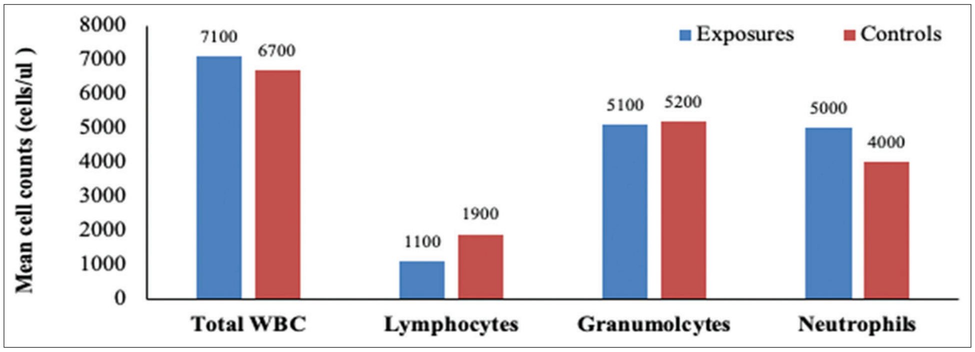


Fig. 1. Leukocyte measurements for traffic emission exposures and controls.

gas sources has a significant effect on the mean values of lymphocytes among exposures themselves ($P < 0.05$).

The total lymphocyte counts among exposed smokers were 1002 cells/μl ±3.5, whereas among non-smoker exposures were higher (1161 cell/μl ±3.8), the statistical analyses indicated that there were valuable effects of smoking on the lymphocyte counts especially when integrated with traffic emission gases ($P < 0.05$). In addition, the effects of the duration of emission exposure on lymphocytes counts showed that the mean value for exposures with about 5–10 years of exposing history was 1137 cells/μl ±3.4. For those with more prolonged exposure history (10–20 years) lymphocytes were relatively lower (1056 cells/μl ±3.7), which indicated that the exposure duration plays a significant effect on the total lymphocyte counts (P < 0.05).

The lower levels of lymphocyte count among exposures may be due to the toxic effects of the chemical contents of the traffic emissions. Similar observations were recorded by other investigators who found that the mean value of lymphocyte counts was reduced as a result of exposure to chemicals raised from fuel-burning [9]. Integration of the smoking effects with emission gases among exposures confirmed the impact of traffic emissions of the WBCs in general and on lymphocyte numbers, especially the mean value of lymphocytes was declined among non-smokers and significantly different from controls.

The observations reported by the current study were parallel to the results mentioned by other investigators [10] who noticed a decline in the total numbers of white cells and lymphocytes among mice, which were exposed to the traffic emissions. Changes in granulocyte counts also were studied. The mean value of granulocytes was 5100 cells/μl ±7.8 among exposures, and 5200 cells/μl ±9.7 among controls. No significant difference was seen between exposures and controls considering granulocytes ($P > 0.05$) (Fig. 1). No valuable effects of smoking and exposure duration were reported ($P > 0.05$), which may indicate that any decline in the granulocyte numbers was not due to the smoking effects, as in the case of lymphocytes. Unlike the above observations, the mean value of neutrophil counts was significantly higher among exposed cases (5100 cells/μl ±1) when compared to that of control cases (4000 cells/μl ±9.4) (Fig. 1). Smoking and exposure periods showed no noticeable effects among exposures themselves ($P > 0.05$). The current observations relatively confirm the impact of chemical products of the traffic emission, especially when the effects of smoking were adverse, as other investigators talked about the negative effects of smoking on blood parameters, including granulocyte. Different investigators reported that the neutrophil count was raised among emission exposures when they compared their observations to the control groups [5].

Moreover, the results of the current study were parallel to the observations recorded by other study that showed higher neutrophil counts exposures when compared to controls [6].

3.2. RBCs Measurement RBC

In general, the total numbers of RBCs were almost similar between exposures (5700 × 102 cells/μl ±0.83) and controls (5600 × 102 cells/μl ±1), and no valuable variations were seen between them ($P > 0.05$) (Fig. 2). The RBC counts for those whose home was (100–200 m) far from the sources of traffic emissions was (5400 × 102 cells/μl ±0.5), while it was higher for those whose home was more far (500–1000 m) that was (5900 × 102 cells/μl 0.9). Smoker exposures showed lower RBC counts (5500 × 102 cells/μl ±0.7) when compared with non-smoker exposures (5800 × 102 cells/μl ±1.2), although it was not significant ($P > 0.05$). It was noticed that RBC count for exposures with (5–10 years) exposure history was higher (5600 × 102 cells/μl ±1.1), in compared

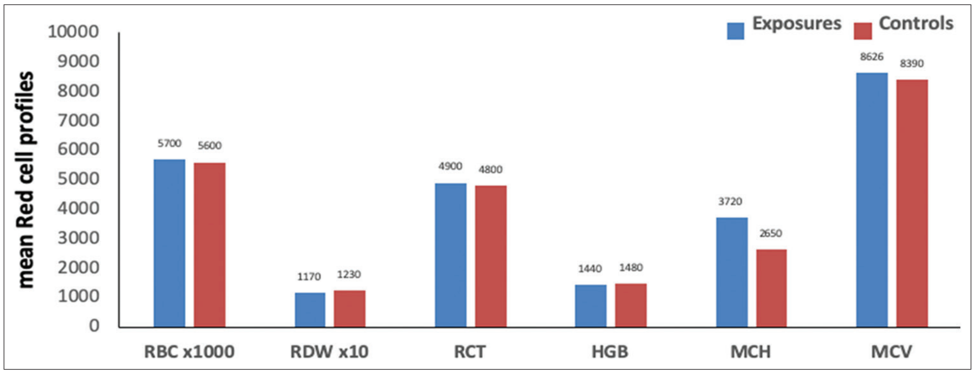


Fig. 2. Mean values of red blood cells for traffic emission exposures and controls.

to those with more prolonged exposure history (10–20 years) (5400 × 102 cells/μl ±0.81). However, the differences were not valuable ($P > 0.05$). Among the factor that may explain the above observation may be due to the sample collection season (summer), where the traffic gases may be less effective on exposures compare to cold and dry weather. However, other scientific works reported significant effects of traffic gases on RBC counts and showed elevated RBC counts among exposures does not agree with the current observations [8]. Moreover, another factor may play a role, which is the presence of relatively low levels of $PM_{2.5}$ in the Iraq fuel, as previously noticed that the high $PM_{2.5}$ may be responsible for elevations in RBC counts [8], [9], [15].

The RDW for exposures was lower (11.7%, ±0.74) when compared with that of controls (12.3%, ±0.94), although it was not significant ($P > 0.05$) (Fig. 2). It was noticed that smoking, home distance, and exposure duration have no significant effects on RDW for exposures themselves ($P > 0.05$). The low levels of RDW in the current study may be due to the slight reductions in RBC counts, especially the RDW can be considered as a marker for RBC counts and sizes. The results of the current study not agreed with the previous reported by other investigators who showed decline RBC counts significantly [22].

Furthermore, it was reported that the HCT for exposures (49%, ±0.98) was not different significantly from that of controls (48%, ±0.76) ($P > 0.05$) (Fig. 2). Statistical analysis indicated that smoking, home distance, and exposure duration have no significant effect on HCT ($P > 0.05$). The changes in the HCT among exposures may be due to the limited effects of traffic emission, as mentioned earlier, especially HCT that can reflect alterations in red cell count and functions. The current observations were not agreed with the results reported by other studies that showed the elevated HCT among traffic emission exposures [8], [15]. Moreover, results showed that emission exposure has no significant effects on Hb of exposures (14.4 g% ±4) compared to controls (14.8 g% ±2.9) ($P > 0.05$) (Fig. 2). Smoking, home distance, and exposure duration showed no valuable effects on Hb ($P > 0.05$). The current results indicated that due to non-valuable decline in Hb, the value of HCT was not changed significantly ($P > 0.05$). The above results were not agreed with the studies that reported by other researchers in the past that observed the pollutants could lead to anemic conditions, which consequently cause a reduction in HCT [6], [11], [23]. Although the results of the current study were supported to the observations that reported by some investigators who studied the effects of emissions on traffic polices and in Pakistan, and claimed that the traffic emission has no significant effects of HGB HB. While they reported that smoking was effective on HB, which was not agreeing with the current observation [16], [17].

MCH also was among the hematologic parameters which were not significantly varied between exposures and controls ($P > 0.05$). Furthermore, it was noticed that smoking, home distance, and exposure duration have no valuable effects on MCH and MCV ($P > 0.05$). Similarly, the observations were recorded for MCHC ($P > 0.05$) (Fig. 2), which was agreed with results obtained in a study done on mice in the past considering MCHC, where the level was reduced [14]. As our statement, the lack of effects of the traffic of emission on the vast majority of RBC profiles may be due to the saturated environment with O_2, especially the study area is rich in forest

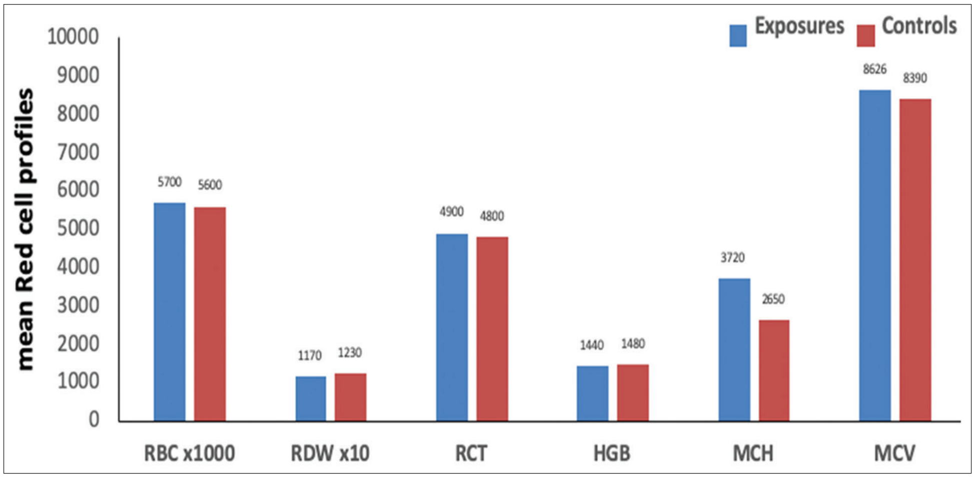


Fig. 3. Mean value of platelet profile for traffic emission exposures and controls.

and green spaces are at an excellent level. Low levels of O_2 can negatively affect RBC profiles; especially there is a strong relation between RBC and oxygen transportation.

3.3. Platelet Measurement PLT

The mean values of platelet count among exposures were higher (1800 cells/µl ±6.5) than that of controls (1690 cells/µl ±3.9). When the results analyzed, it has appeared that significant differences were found among exposures and controls regarding platelet counts ($P < 0.05$) (Fig. 3). Moreover, it was concluded that home distance and exposure duration have significant effects on platelet counts, respectively ($P < 0.05$). Smoking showed no valuable effects, which may confirm that all outcomes are due to the long-term exposure to the chemical components of traffic emission, not to the smoking contents. Other researcher found similar results on experimental animals and humans [18], [19], [20], [21]. They suggested elevation in platelet counts concerning emission air pollutants.

The current study revealed that there were no noticeable differences between exposures and controls regarding MPV ($P > 0.05$). Furthermore, it has appeared that home distance, smoking, and exposure duration have no significant effects of MPV ($P > 0.05$) (Fig. 3). Similarly, no valuable variations were observed between exposures and controls regarding PDW Smoking, home distance, and exposure duration showed no noticeable effects on PDW ($P > 0.05$) (Fig. 3), which might be due to the relations of changes in both MPV and PDW [24]. In addition, it was concluded from the current study that traffic emission has no significant effect on PCT and LPCR ($P > 0.05$). This study revealed that smoking, home distance, and exposure duration showed no valuable effects on each of PCT and LPCR ($P > 0.05$) (Fig. 3). In a study, it was noticed that the LPCR effects due to chemical exposure have a significant role in the discrimination between hyper-destructive and hypo-productive thrombocytopenia [25]. However, the PCT levels were fewer, especially among traffic emission exposures; however, it may increase in acute cholecystitis patients with PDW and lowered MPV [24].

4. CONCLUSION

Traffic emission gases showed no significant effects on the vast majority of the hematologic parameters, although, valuable elevation has been seen in neutrophils and platelets due to the traffic emission. The results of the current study suggested links between inflammatory and cardiovascular diseases among emission exposures. Future researches must be considered to investigate these relations.

REFERENCES

[1] M. Franchini and P. M. Mannucci. "Thrombogenicity and cardiovascular effects of ambient air pollution". *Journal of Blood,* vol. 118, no. 9, p. 2405, 2011.

[2] R. Khan and A. Agarwal. "Modulatory effect of Vitamine E and C on nitrogen dioxide induced hematotoxicity in both the sexes of wistar rats". *International Journal of Interdisciplinary Research*, vol. 3, no. 3, pp. 46-50, 2016.

[3] N. Boussettaa, S. Abedelmalekc, H. Mallekd, K. Alouie and N. Souissiaa. "Effect of air pollution and time of day onperformance, heart rate hematologicalparameters and blood gases, following theYYIRT-1 in smoker and non-smoker soccerplayers". *Science and Sports*, vol. 33, no. 6, pp. 1-14, 2018.

[4] P. Ahlawat. "Effect of sulphur dioxide exposure on haematological parameters in albino rats". *Journal of Scientific and Engineering Research*, vol. 3, no. 6, pp. 58-60, 2016.

[5] C. Tan, Y. Wang, M. Lin, Z. Wang, L. He, Z. Li, Li, Y and K. Xu. "Long-term high air pollution exposure induced metabolic adaptations in traffic policemen". *Environmental Toxicology and Pharmacology*, vol. 58, no. 16, pp. 156-162, 2018.

[6] R. M. Kartheek and M. David. "Modulations in haematological aspects of wistar rats exposed to sublethal doses of fipronil under subchronic duration". *Journal of Pharmaceutical, Chemical and Biological Sciences*, vol. 5, no. 3, pp. 187-194, 2017.

[7] C. Jephcote and A. Mah. "Regional inequalities in benzene exposures across the European petrochemical industry: A Bayesian multilevel modelling approach". *Environment International*, vol. 132, no. 104812, pp. 1-17, 2019.

[8] Bahaoddini and M. Saadat. "Hematological changesdue to chronic exposure to natural gasleakage in polluted areasof Masjid-i-Sulaiman (Khozestan province, Iran)". *Ecotoxicology and Environmental Safety*, vol. 58, no. 2, pp. 273-276, 2004.

[9] Kamal, A. Cincinelli, T. Martellini and R. N. Malik. "Linking mobile source-PAHs and biological effects in traffic police officers and drivers in Rawal pindi (Pakistan)". *Ecotoxicology and Environmental Safety*, vol. 127, pp. 135-143, 2016.

[10] G. M. Farris, S. N. Robinson, B. A. Wong, V. A. Wong, W. P. Hahn and R. Shah. "Effects of benzene on splenic, thymic, and femoral lymphocytes in mice". *Toxicology*, vol. 118, no. 2-3, pp. 137-148, 1997.

[11] T. Honda, C. V. Puna, J. Manjourides and H. Suhb. "Anemia prevalence and hemoglobin levels are associated with long-term exposure to air pollution in an older population". *Environment International*, vol. 101, no. 4, pp. 125-132, 2017.

[12] A. Masih, A. Lall, A. Taneja and R. Singhvi. "Exposure profiles, seasonal variation and health risk assessment of BTEX in indoor air of homes at different microenvironments of a terai province of northern India". *Chemosphere*, vol. 176, no. 2, pp. 8-17, 2017.

[13] M. Abu-Elmagd, M. Alghamdi, M. Shamy, M. Khoder, M. Costa, M. Assidi, R. Kadam, H. Alsehli, M. Gari, P. N. Pushparaj, G. Kalamegam and M. H. Al-Qahtani. "Evaluation of the effects of airborne particulate matter on bone marrow-mesenchymal stem cells (BM-MSCs): Cellular, molecular and systems biological approaches". *International Journal of Environmental Research and Public Health*, vol. 14, no. 4, p. 440, 2017.

[14] J. Reisa and S. Martel. "Acute exposure guideline levels for selected airborne. In: *Acute Exposure Guideline Levels*, National Academy of Sciences/National Research Council (US) Committee, Washington DC, USA, p. 178, 2014.

[15] L. Ton. “Platelet neutrophil interactions as drivers of inflammatory and thrombotic disease”. *Cell and Tissue Research*, vol. 371, pp. 567-576, 2018.

[16] E. Wigenstama, L. Elfsmarka, A. Buchta and S. Jonassona. “Inhaled sulfur dioxide causes pulmonary and systemic inflammation leading tochemical-induced lung injury”. *Toxicology*, vol. 368-369, no. 4, pp. 28-36, 2016.

[17] Ö. Etlik and A. Tomur. “The oxidant effects of hyperbaric oxygenation and air pollution in erythrocyte membranes (hyperbaric oxygenation in air pollution)”. *European Journal of General Medicine*, vol. 3, no. 1, pp. 21-28, 2006.

[18] P. Poursafa, R. Kelishadi, A. Amini, Amini, A. M. Amin, M. Lahijanzadeh and M. Modaresi. “Association of air pollution and hematologic parameters in children and adolescents”. *Jornal de Pediatria*, vol. 87, no. 4, pp. 350-356, 2011.

[19] Gorriz, S. Llacuna, M. Riera and J. Nadal. “Effects of air pollution on hematological and plasma parameters in apodemus sylvaticus and mus musculus”. *Archievs of Environmental Contamination and Toxicology*, vol. 31, no. 1, pp. 153-158, 1996.

[20] G. L. Walter. “Effects of carbon dioxide inhalation on hematology, coagulation, and serum clinical chemistry values in rats”. *Toxicologic Pathology*, vol. 27, no. 2, pp. 217-225, 1999.

[21] Q. Sun, X. Hong and L. E. Wold. “Cardiovascular effects of ambient particulate air”. *Circulation Journal*, vol. 121, no. 25, pp. 2755-2765, 2010.

[22] M. Kargarfard, A. Shariat, B. Shaw, I. Shaw, T. Lam, A. Kheiri, A. Eatemadyboroujeni and S. M. Tamrin. “Effects of polluted air on cardiovascular and hematological parameters after progressive maximal aerobic ex”. *Lung Journal*, vol. 193, no. 2, pp. 275-281, 2015.

[23] M. Nikolić, D. Nikić and A. Stanković. “Effects of air pollution on red blood cells in children”. *Polish Journal of Environmental Study*, vol. 17, no. 2, pp. 267-271, 2008.

[24] M. Zain and S. Aitte. “Study of changes in blood parameters and calculation of PCT, MPV and DPW for the platelets of laboratory females and males of albino mice during exposure to doses of pyrethriodpesticide (alphacypermethrin)”. *IOSR Journal of Pharmacy and Biological Sciences*, vol. 14, no. 2, pp. 71-78, 2019.

[25] Y. Budak, M. Polat and K. Huysal. “The use of platelet indices, plateletcrit, mean platelet volume and platelet distribution width in emergency non-traumatic abdominal surgery: A systematic review”. *Biochemical Medicine* (*Zagreb*), vol. 26, no. 2, pp. 178-193, 2016.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Proposed Fully Homomorphic for Securing Cloud Banking Data at Rest

Zana Thalage Omar[1,2]*, **Fadhil Salman abed**[1,2]

[1]University of Human Development, College of Science and Technology, Department of Computer Science, Sulaymaniyah, Kurdistan Region of Iraq, Iraq, [2]University of Sulaimani, College of Science, Computer Department, Sulaymaniyah, Kurdistan Region of Iraq, Iraq

ABSTRACT

Fully homomorphic encryption (FHE) reaped the importance and amazement of most researchers and followers in data encryption issues, as programs are allowed to perform arithmetic operations on encrypted data without decrypting it and obtain results similar to the effects of arithmetic operations on unencrypted data. The first (FHE) model was introduced by Craig Gentry in 2009, and it was just theoretical research, but later significant progress was made on it, this research offers FHE system based on directly of factoring big prime numbers which consider open problem now, The proposed scheme offers a fully homomorphic system for data encryption and stores it in encrypted form on the cloud based on a new algorithm that has been tried on a local cloud and compared with two previous encryption systems (RSA and Paillier) and shows us that this algorithm reduces the time of encryption and decryption by 5 times compared to other systems.

Index Terms: Cloud Computing Security, Encryption, Decryption, Cloud Storage, Homomorphic Encryption

1. INTRODUCTION

Computing technology is seeing significant progress and significant interest, especially when the computation outsourcing has been outsourced to a third party as the cloud is the most frequently used form [1]. That is why many companies no longer trust to store their sensitive data in the cloud, which uses traditional unsecured encryption systems [2]. From this, the need to use homomorphic encryption for banking data is coming, which is a new approach that can help banks to increase data security and management [3]. There are two types of homomorphic cryptosystems: Partially homomorphic systems and fully homomorphic systems [4]. Partially homomorphic schemes support one of the additions or multiplication operations, these systems are divided into two parts according to the process that supports like the RSA, where it only supports the multiplication process and does not support the addition process, for example, if we have two numbers M1, M2 and they are encrypted by the RSA, then its value becomes C1, C2 and on obtaining the product of multiplying the two encrypted values C1 * C2 = C3 and then we decrypt the encrypted output C3, we will get a result similar to M1 * M2 = M3, but if we add the two values C1 + C2 = C4 and when decrypting the result C4 we do not get a result similar to M1 + M2 = M4. On the contrary, when the two values are encrypted using Paillier, we find that only the result of C1+C2=C5 is similar to M1+M2=M3 and C1*C2=C6 do not equal to M1+M2=M4. Therefore, we say that the two algorithms (RSA and Paillier) are not a fully homomorphic systems [5], [6]. The first FHE was given in 2009 by Craig Gentry [7]. Researchers first researched a (FHE) system in the late last century, specifically at the end of the seventies, and soon after, in 1987, RSA was published, the RSA algorithm became a leading approach by many researchers because at that time there was no idea of the public key cipher

Access this article online

DOI:10.21928/uhdjst.v4n1y2020.pp87-95

E-ISSN: 2521-4217

P-ISSN: 2521-4209

Corresponding author's e-mail: Zana Thalage Omar, University of Human Development, College of Science and Technology, Department of Computer Science, Sulaymaniyah, Kurdistan Region of Iraq, Iraq, University of Sulaimani, College of Science, Computer Department, Sulaymaniyah, Kurdistan Region of Iraq, Iraq. E-mail: zana.th.omar@gmail.com

Received: 13-03-2020 Accepted: 10-05-2020 Published: 12-05-2020

that was presented during the RSA scheme for the first time [5]. Because this kind of encryption allows the key to decrypt the encrypted data, and thus one can read and know all the data, and for this reason, if one does not have the secret key, the data become useless. Therefore, a question and an issue were asked: Can mathematical operations apply to encrypted data without decrypting it, and from this, the idea of using fully homomorphic systems (FHE) was raised. After that, several attempts were made to develop these systems, but most of the research did not succeed as they received partially homomorphic schemes such as RSA and Goldwasser-Micali [8]. The algorithm that achieves the addition and multiplication properties can be considered as FHE, as it is regarded as a special algorithm that contains the feature of performing mathematical operations (addition and multiplication) on data without decrypting it and obtaining correct results [9]. FHE is an encryption technology that allows calculations to be performed on encrypted data without decrypting it, and this results in an encrypted result where when this result is decrypted we get a result similar to the result of the calculations on the data without encrypting it [9]. The world of computing is in constant progress, and the main challenge is to create a guarantee and trust among customers when storing their sensitive data on the cloud to ensure and respect their privacy. This is a new approach that cloud providers follow to encrypt users' data, upload it to the cloud, and perform operations on it without decrypting it to ensure the integrity of customer data [10]. This paper presents a fully homomorphic system (the correct numbers and texts) based on a new algorithm that will be explained later in this paper as this scheme relies on data encryption and operations performed on it without decrypting and reducing computational complications and the time used to encrypt and decrypt data and reduce energy consumption. Most of the previous research in this field deals with data when encrypting after converting it to the binary system and this means more time. As for our current research, data operations are encrypted without the need to convert them to binary representation and this reduces mathematical operations and there is a reduction in the time of encryption and encryption solution, as well as a mathematical model has been suggested that deals with the inverse calculation and the process of raising to the exponential and increases the complexity of attacking the new system.

2. LITERATURE REVIEW

C. Gentry *et al.* (2012), this paper introduces contrast/orientation techniques to transfer the elements of plain text across these vectors very efficiently so that they are able to perform general calculations in a batch way without the need to decrypt the text and also make some improvements that can accelerate the normal homomorphic, where you can make homogeneous evaluation of arithmetic operations using multi-arithmetic head only [11].

J. Fan *et al.* (2012), this paper concludes two copies of the redefinition that lead to a quick calculation of homogeneous processes using the parameter transformation trick, as this paper conveys Brakerski's fully homomorphic scheme based on the learning with errors (LWE) problem to the ring-LWE [12].

Z. Brakerski *et al.* (2012), this paper introduces squash and bootstrapping techniques to convert a somewhat symmetric encryption scheme into an integrated symmetric encryption scheme [13].

X. Cao *et al.* (2014), this paper presents a completely symmetric encryption scheme using only a basic unit calculation as it relies on the technique of using multiplication and addition instead of using ideal clamps on a polynomial loop [14].

C. Xiang *et al.* (2014), this paper presents an entirely symmetric encryption scheme on integers, as it reduces the size of the public key using the square model encryption method in public key elements instead of using a linear model based on a stronger variant of the approximate-GCD problem [15].

M. M. Potey *et al.* (2016), this paper presents a completely symmetric encryption system where it focuses on storing customer data on the cloud in an encrypted form so that customer data remain safe and when any data modification is made the system loads data on the customer's device and modifies it and then stores it again on the cloud in encrypted form [16].

K. Gai *et al.* (2018), this paper proposes a new solution for mixing real numbers on a novel tensor-based FHE solution that uses tensor laws to reduce the risk of unencrypted data storage [17].

S. S. Hamad *et al.* (2018), these heirs offer a completely symmetric encryption system, as it relies on the principle of encryption a number from the plain text with another number using a secret key without converting to binary format and then comparing the result with a DGHV and SDC system [18].

S. S. Hamad *et al* (2018), this paper presents a fully homomorphic encryption system based on Euler's theory and the time complexity has been calculated and compared with other systems with an encrypt key size up to 512 bits while the size of the key in our proposed scheme reaches more than 2048 bits and the encrypting process is done through more complex and powerful mathematical equations [19].

V. Kumar *et al* (2018), this paper presents fully homomorphic encryption system with probabilistic encrypting and relies on Euler's theory. The encrypting process is done through the following mathematical equation ($C=M^{k* \mu (n) +1}$ mod x) while in our proposed scheme a more complex and difficult mathematical algorithm is used which helps to stand more against hacker attacks and deter them [20].

R. F. Hassan *et al.* (2019), this paper proposes a blueprint for building asymmetric cloud-based architecture to save user data in the form of unusual text. This pattern uses the elliptic curve to create the secret key for data encryption. This pattern is a new algorithm that reduces processing time and storage space [21].

3. STATEMENT OF THE PROBLEM

Cloud providers provide many services, including applications and storage many companies and users do not trust the providers of these services due to security concerns. Where the user does not upload his personal data to the cloud because the cloud providers are able to read and modify every bit loaded on the cloud and use it for personal purposes, and this thing does not comply with respecting the user's privacy. Furthermore, some cloud providers still use traditional security techniques that are not secure with low-security level to protect user privacy. Some of the cloud providers have started to use high-level technologies to protect the privacy of users and the security of their data, but there remains a problem that the provider of the cloud itself is still able to access user data, and this is not safe for users. This problem can be solved when following FHE systems when storing data on the cloud where these systems can encrypt the data and store it in the cloud in an encrypted form and thus the cloud provider or others cannot see the data and use it, so the privacy of users and the security of their data are protected.

4. PROPOSED FHE SYSTEM

The proposed scheme works as follows:
Generating the encryption key and then encrypting the numbers and texts and storing them in encrypted form on the cloud. In our work, we use a local cloud and experiment with the proposed scheme on it. The purpose of this process is to save the data encrypted on the cloud so that no one can view the data and use it for personal purposes Therefore, when the data owner needs to perform an amendment of the encrypted data on the cloud, an encrypted request is sent to the server, and the server performs mathematical operations on the encrypted data and returns an encrypted result where this encrypted result can only be decrypted through the private encryption key which is with the owner of data only so that he can decrypt the encrypted result and see his data. In this way, we have been able to maintain the privacy and security of the data when stored in the cloud. These procedures go through three stages. Generation the encryption key stage, the encryption stage, and the decryption stage. The model of the proposed scheme is given in Fig. 1, and the flowchart of the proposed scheme is given in Fig. 2. The proposed scheme performs several random examples with multiplication and addition as follows:

A. Key Generation:
  1. Generate two large Prime number p and q
  2. Compute n = p*q
  3. Calculate $L=((P^{-1} \bmod q)*p)+((q^{-1} \bmod p)*q)$
  4. Select r: Where r is a big random integer

B. B. Messages Encryption

The conditions:

$$(M_1\&M_2), (M_1+M_2), \text{and } (M_1*M_2) < n$$

Where M1 and M_2 are the Messages.

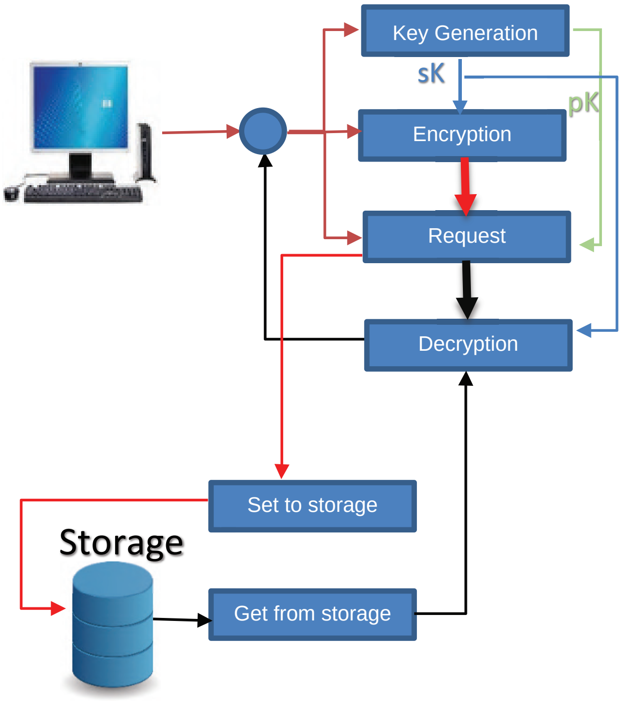


Fig. 1. Model of proposed FHE scheme.

The schema of message encryption is:

$$C = L * M^{r \mu (p) +1} \bmod n. \tag{1}$$

Where μ (p) = (p-1), Euler function and C, the ciphers text.

C. Message Decryption

The schema of cipher decryption is:

$$M=C \bmod p \tag{2}$$

Where M is the number or text that will be encrypt and C is the result of the encrypted number or text we named it cipher text

D. Euler's Theorem

All of us know that Euler's Theorem contains two-part they are:

1. $M^{\mu (p)} \equiv 1(\bmod p)$, when p and m are prime to each other.
2. $M^{r* \mu (n) +1} \equiv M \ (\bmod n)$, when r is an integer, M<n and n=p*q where q and p are two primes number.

E. A simple example of how to make an amendment to encrypted data

We have two values M1 = 3, M2 = 5. We encrypt them through a simple encryption equation that is multiplied by each value, so we get C1 = M1 * 2 and C2 = M2 * 2, so C1 = 6, C2 = 10 when we add the two values C1 + C2 = C3 so C3 = 16 We decrypt C3 so we get the result 16/2 = 8 which is the same result when we add M1 + M2 = M3 where 3 + 5 = 8 as shown in Fig. 3.

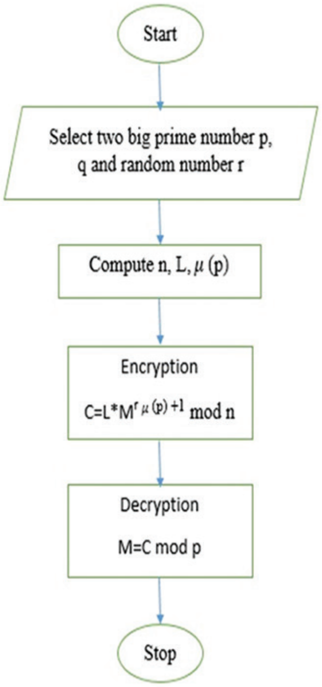

Start
Select two big prime number p, q and random number r
Compute n, L, μ (p)
Encryption
$C=L*M^{r \mu (p) +1} \bmod n$
Decryption
M=C mod p
Stop

Fig. 2. Flowchart of proposed FHE scheme.

5. THE PROVE OF OUR SCHEMA IS FHE

We choose two numbers M_1, M_2 and encrypt them to get two encrypted or (ciphers) C_1 and C_2, respectively, and then we combine $C_1 + C_2$ to get a new ciphered result we name it C_3 then we decrypt C_3 and compare the result with M_3 which is the result of combine $M_1 + M_2$ we also multiply $C_1 * C_2$ to get C_4 and compare it to M_4, which is the result of $M_1 * M_2$.

A. The Prove of Additive Homomorphic

If the following condition is fulfilled, it becomes clear to us that the proposed scheme additive homomorphic:

$$M_1+M_2=\text{dec}\ [\text{enc}\ (M_1) + \text{enc}\ (M_2)] \tag{4}$$

Where dec is the decryption function and enc is the encryption function

Proof:

$$C_1=L*(M_1^{r \mu (p) +1} \bmod n).$$
$$C_2=L*(M_2^{r \mu (p) +1} \bmod n).$$
$$C_1+C_2= L*(M_1^{r \mu (p) +1} \bmod n) + L*(M_2^{r \mu (p) +1} \bmod n).$$
$$\text{dec}\ (C_1+C_2) = (C_1+C_2) \bmod p$$
$$= [L*(M_1^{r \mu (p) +1} \bmod n) + L*(M_2^{r \mu (p) +1} \bmod n)] \bmod p$$
$$= [(L \bmod p) + ((M_1^{r \mu (p) +1} \bmod n) \bmod p) + (L \bmod p) + ((M_2^{r \mu (p) +1} \bmod n) \bmod p)]$$
$$= [(M_1^{r \mu (p) +1} \bmod p) \bmod n + (M_2^{r \mu (p) +1} \bmod p) \bmod n]$$

We know $M_1^{r \mu (p) +1} \bmod p = M_1$ and $M_2^{r \mu (p) +1} \bmod p = M_2$ by Euler's Theorem so

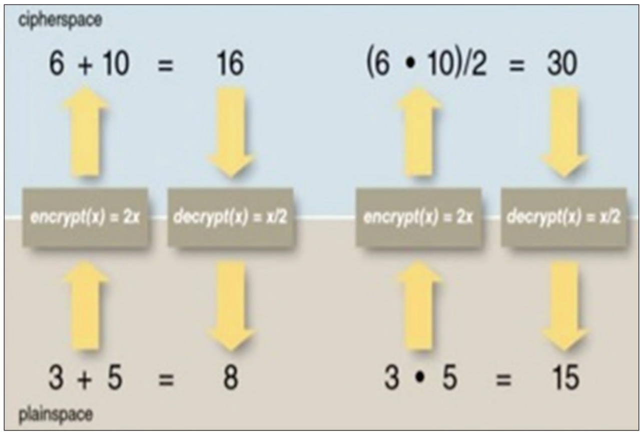


Fig. 3. A simple example of how modify encrypted data.

$= (M_1 \text{ mod } n) + (M_2 \text{ mod } n)$
$= (M_1+M_2) \text{ mod } n$
Because M_1+M_2 less than $<$ (n)
$= M_1+M_2$
dec $(C_1+C_2) = M_1+M_2$ so the condition is fulfilled

B. The Prove of Multiplicative Homomorphic

If the following condition is fulfilled, it becomes clear to us that the proposed scheme multiplicative homomorphic:

$$M_1*M_2=\text{dec }[\text{enc }(M_1) * \text{enc }(M_2)] \quad (5)$$

Where dec is the decryption function and enc is the encryption function

Proof:
$C_1=L*(M_1^{\,r\,\mu\,(p)\,+1} \text{ mod } n).$
$C_2=L*(M_2^{\,r\,\mu\,(p)\,+1} \text{ mod } n).$
$C_1*C_2= (L*(M_1^{\,r\,\mu\,(p)\,+1} \text{ mod } n)) * (L*(M_2^{\,r\,\mu\,(p)\,+1} \text{ mod } n)).$
dec $(C_1*C_2) = (C_1*C_2) \text{ mod } p$
$= [(L*(M_1^{\,r\,\mu\,(p)\,+1} \text{ mod } n)) * (L*(M_2^{\,r\,\mu\,(p)\,+1} \text{ mod } n))] \text{ mod } p$
$= [(L \text{ mod } p) * ((M_1^{\,r\mu\;(p)\,+1} \text{ mod } n) \text{ mod } p) * (L \text{ mod } p) * ((M_2^{\,r\,\mu\,(p)\,+1} \text{ mod } n) \text{ mod } p)]$
$= [(M_1^{\,r\,\mu\,(p)\,+1} \text{ mod } p) \text{ mod } n * (M_2^{\,r\,\mu\,(p)\,+1} \text{ mod } p) \text{ mod } n]$
We know $M_1^{\,r\,\mu\,(p)\,+1} \text{ mod } p = M_1$ and $M_2^{\,r\,\mu\,(p)\,+1} \text{ mod } p = M_2$ by Euler's Theorem so
$= (M_1 \text{ mod } n) * (M_2 \text{ mod } n)$
$= (M_1*M_2) \text{ mod } n$
Because M_1*M_2 less than $<$ (n)
$= M_1*M_2$
dec $(C_1*C_2) = M_1*M_2$ so the condition is fulfilled

6. REAL EXAMPLE

Let us choose two different number $M_1= 10$, M2 = 40, select two big prime numbers p=523, q=617, select random number r=100 and compute n, L where n = p*q and L = $((P^{-1} \text{ mod } q)*p) + ((q^{-1} \text{ mod } p)*q)$, as in Fig. 1, so n = 322691 and L = 322692 now we will compute C_1, C_2 as shown in Fig. 4 where

$C_1= L*(M_1^{\,r\,\mu\,(p)\,+1} \text{ mod } n)$
$C_1= 322692 * (10^{100*(522)\,+1} \text{ mod } 322691)$
$C_1= 84555952836$
$C_2= L*(M2^{\,r\,\mu\,(p)\,+1} \text{ mod } n)$
$C_2= 322692 * (40^{100*(522)\,+1} \text{ mod } 322691)$
$C_2= 70220360736$

A. Check the Additive Homomorphism

As shown in Fig. 5, Let us define C_3 is the result of C_1+C_2

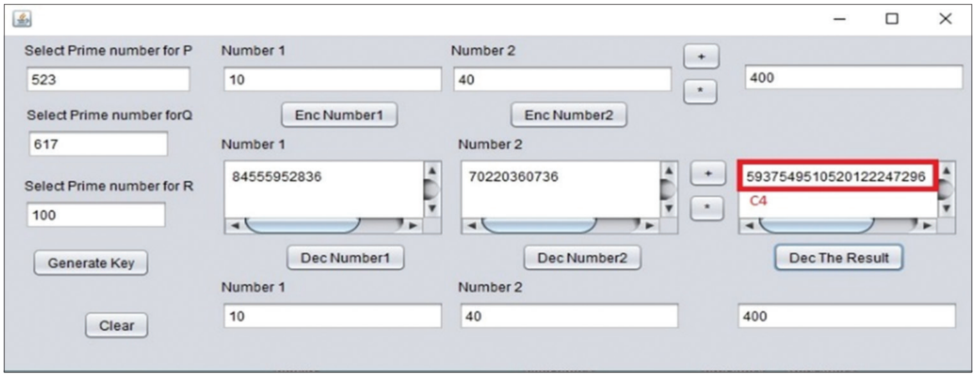
$C_3= C_1+C_2$
$C_3= 84555952836+70220360736$
$C_3= 154776313572$
$M_3= C_3 \text{ mod } p$
$M_3= 154776313572 \text{ mod } 523$
$M_3= 50$, which is the same of $M_1+M_2 = 10 + 40 =50$

B. Check the Multiplication Homomorphism

As shown in Fig. 6, Let us define C_4 is the result of C_1*C_2
$C_4= C_1*C_2$
$C_4= 84555952836*70220360736$
$C_4= 5937549510520122247296$
$M_4= C_4 \text{ mod } p$
$M_4= 5937549510520122247296 \text{ mod } 523$
$M_4= 50$, which is the same of $M_1*M_2 = 10 * 40 =400$

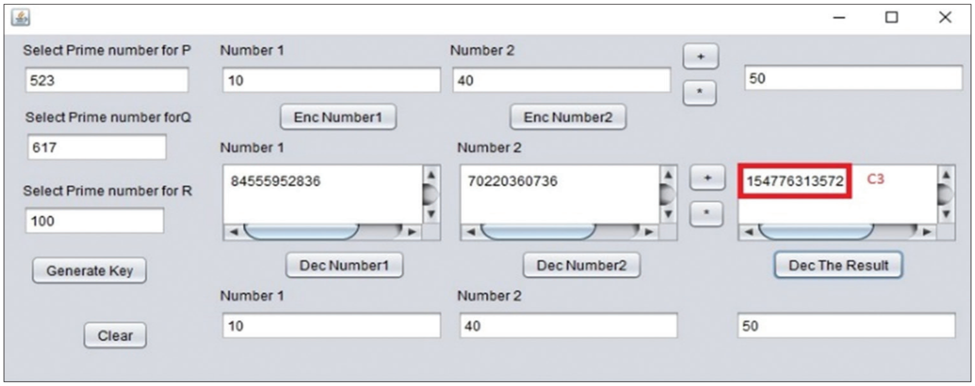
7. RESULTS

Our proposed method has been applied in Java Language on a laptop that has these characteristics Intel (R) core (TM)



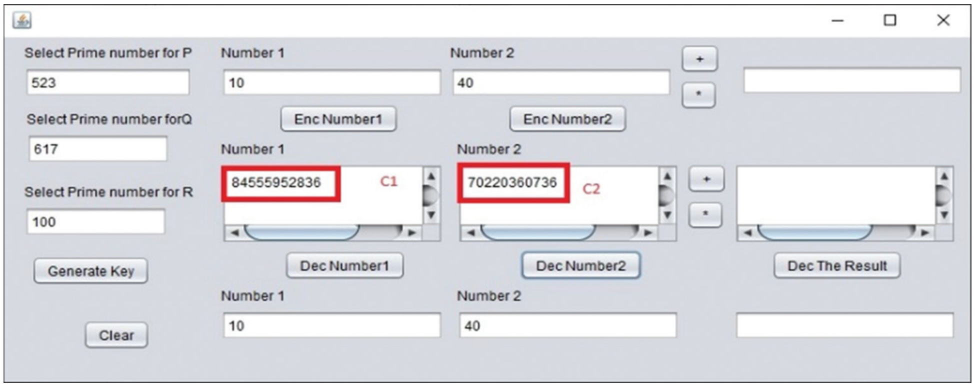
Select Prime number for P: 523
Select Prime number forQ: 617
Select Prime number for R: 100
Number 1: 10
Number 2: 40
400
Number 1: 84555952836
Number 2: 70220360736
5937549510520122247296
C4
Number 1: 10
Number 2: 40
400

Fig. 4. Verification that the proposed scheme supports multiplicative homomorphic system.



Select Prime number for P: 523
Select Prime number forQ: 617
Select Prime number for R: 100
Number 1: 10
Number 2: 40
Result: 50
Number 1: 84555952836
Number 2: 70220360736
C3: 154776313572
Number 1: 10
Number 2: 40
Result: 50

Fig. 5. Verification that the proposed scheme supports additive homomorphic system.



Select Prime number for P: 523
Select Prime number forQ: 617
Select Prime number for R: 100
Number 1: 10
Number 2: 40
Number 1 (C1): 84555952836
Number 2 (C2): 70220360736
Number 1: 10
Number 2: 40

Fig. 6. A real example of generating an encryption key and encrypting two different numbers.

i7-8550U CPU @1.80GHz 2.00GHz, 8 GB Ram, 64-bit Operating System,x64-based processor, Windows 10 and Big Integer library of java is used. We have previously seen that in section 5 our scheme achieves the two properties (addition and multiplication) on the correct numbers when encrypt in contrast to the two systems (RSA and Pailler) that produce a single property either multiplication or addition, we took as an example of 2048 bit A message containing several languages: English, Kurdish, Arabic and Chinese, to indicate that our scheme works in all languages. The message was:

The language considered at the university is English

ئەو زمانەی كە لە زانكۆدا بە پەسەندە زمانی ئینگلیزییە

اللغة المعتبرة في الجامعة هي الانكليزية

大學考慮的語言是英語

P=1696360000174411201884514721530052513663798696916067418627001848350512356831638156323463645139635084178686272336909107427580252972488626713138540811001148507598647266852628765047404770278555262771396058882448569487514849713569452550255019774552242450237953291087748249243514503675865792656993937298380518114 51

Q=14287104767416750569830145755718405618234346454308656139478728164897664839386811178089812726550570830431779982604070801091951359063000033551412304502727764286919185422688044176382868441653979152025714562637864986328594073222682892055143305181268765546606290203967765789214797094090430829344744371724839023970 7

R=36578651863691264477604206506364490015712370834320716872902633782216844351341184895528113792015879819081222933850012467453478465007419488451522718279819 43

N=2423607304574691011057220168339429745775510876856825573501591215207476748560343071696011805583892432344285336810918650824789494269717337049640187576258686356978941473841479446251805425174083399

0064335766935428553190201256143433654205627755229
6773381542471828455818813480172653876483980783338
1523057539397742755088141082360135822895062302531
9405062251415063552873019444449238666440140085803
2829153319755489679960430558612883401366594381416
5468112883656495673094721811758386521739451237520
5070768701405826931878983152614067930454176175622
2924904444160392437762620644204922911348434700560
07271825256265091103199457484857

L=24236073045746910110572201683394297457755108
7685682557350159121520747674856034307169601180
5583892432344285336810918650824789494269717337
0496401875762586863569789414738414794462518054
2517408339900643357669354285531902012561434336
5420562775522967733815424718284558188134801726
5387648398078333815230575393977427550881410823
6013582289506230253194050622514150635528730194
4444923866644014008580328291533197554896799604
3055861288340136659438141654681128836564956730
9472181175838652173945123752050707687014058269
3187898315261406793045417617562229249044441603
9243776262064420492291134843470056007271825256
26509110319945748485

The Message after encryption (Cipher Text)

541476558809409702391337896786206578891371305860
691083047075666735305343010133163455201330600329
818224876828649953919527566237735831578747495518
661815006015094327373599324058140501637682523981
366081263444402953878958225004102881404987245214
085192130463968623162036132714218988345866733882
828903027959438577677193858956252126893602243322
002345822997903630750182808060329693726890973821
429052022147058264305295245097017754099269475380
968046201854139181624798301373478600684536391994
135042539217304792283425928429438405414943114956
731879603950076538717093967938918097476473355425
283428257417215267662967218064104960563636218183
044111151212122457871341575675158274986166996526
006578968820402465601212584511978294298514268554
125554995603375526132322574633145472359908234720
133081143881121000520379674740198817341417761860
826872691325817210768306765600237104658826101240
831563114649492567258100255788974674414548062825

TABLE 1: Computation encryption time of various schema

Key Size	Proposed method	RSA	Pailler
64 Bit	88 ms	59 ms	103 ms
128 Bit	139 ms	100 ms	182 ms
256 Bit	218 ms	149 ms	727 ms
512 Bit	1141 ms	397 ms	4212 ms
1024 Bit	6058 ms	2185 ms	55139 ms
2048 Bit	65876 ms	29820 ms	263303 ms
Average	12253 ms	5451 ms	53944 ms

TABLE 2: Computation decryption time of various schema

Key Size	Proposed Method	RSA	Pailler
64 Bit	31 ms	72 ms	193 ms
128 Bit	43 ms	116 ms	330 ms
256 Bit	50 ms	315 ms	1429 ms
512 Bit	60 ms	1450 ms	11441 ms
1024 Bit	131 ms	7829 ms	122628 ms
2048 Bit	260 ms	79609 ms	976289 ms
Average	95 ms	14898 ms	185385 ms

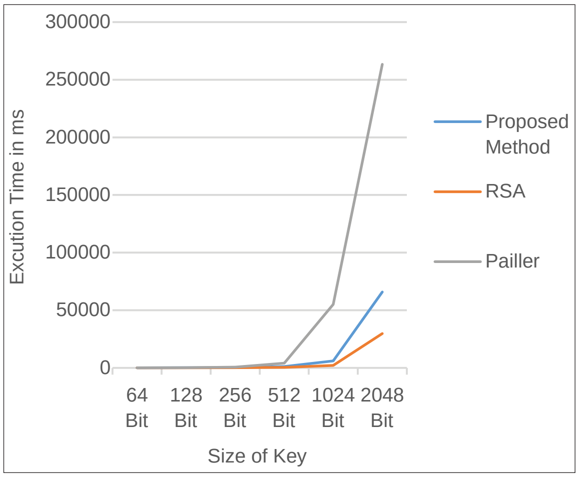


Fig. 7. Computation encryption time of various schema.

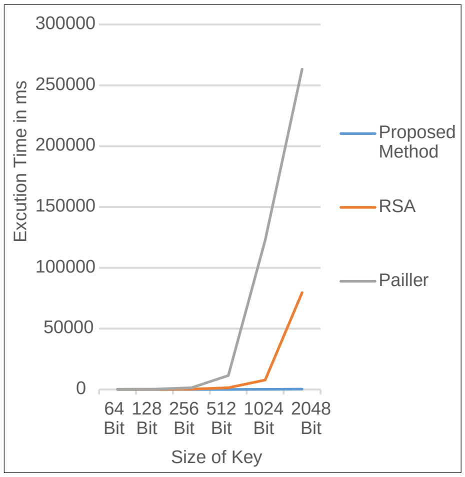


Fig. 8. Computation decryption time of various schema.

2892837234662839061292113517475051240283024822857039832150288766224552917207474813575344328669795611610517942144953845555647644200227716739046080269214238010929862344527946869053956142504282153978829913279741016316926632254422850165388901441231774699848323143262753973952981234912026166080929309059343417238287374915366024637186651090446025995427032842858325561760100762479528333312842498083186883479202672771030666642010819171038712008353233977703555339349166794021508312097161373132603109611666109415743990715529774034556154583520372539433462549143084673593882815487336243261126911242981325891250006136188595483928019489540285506606523583489298137160845149207589839264683606383242087579161421012774684088722206157675920392220322437888837467761391646974013621593727999527387894145533554657005640988111761561242777691841460412436817297935149248403437793923291041969716726718988314898193850389144973734527764417056337441280540889889965231589793043301722177856967321141588234755398782709859264037093772068861826447323185329396490549555672427762431169794565317121037175050358312647042605790539753324457714637549871900468942240262274576522420220671065577816480533078578928195481985808140541026441726779572492306966870609990207l….etc. Due to the length of the encrypted text (Cipher Text), which reaches 67 pages, it has been truncated. Where the encrypted time was 1572 ms, and the decrypted time was 31 ms. We have also tested it on text with 8KB in its size and several different keys in terms of size, and we compared the results with the planners from. In terms of velocity, we obtained the following results: As shown in Tables 1 and 2, respectively, and the graph in Figs 7 and 8.

8. CONCLUSION

Our scheme relies on FHE on whole numbers, texts, and supports all languages such as English, Arabic, Kurdi, and Chinese and others. Very large prime numbers (up to 617 digits, 2048 bit) represent the strength for the attack of our scheme because the proposed system depends on the problem of factorization to the primary factors, which are considered mathematical problems under discussion at the present time when taking the time. We have come to the conclusion that our scheme is very effective in relation to the time when encrypt and decrypt numbers and texts in comparison with other techniques and approaches that are circulated and used at the present time.

REFERENCES

[1] L. A. Tawalbeh and G. Saldamli. “Reconsidering big data security and privacy in cloud and mobile cloud systems”. *Journal of King Saud University Computer*, vol. 40, pp. 1-7, 2019.

[2] J. Domingo-Ferrer, O. Farràs, J. Ribes-González and D. Sánchez. “Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges”. *Computer and Communications*, vol. 140-141, no. 2018, pp. 38-60, 2019.

[3] S. Sakharkar, S. Karnuke, S. Doifode and V. Deshmukh. “A research homomorphic encryption scheme to secure data mining in cloud computing for banking system”. *International Journal for Innovative Research in Multidisciplinary Field*, vol. 4, no. 4, pp. 276-280, 2018.

[4] J. H. Cheon, A. Kim, M. Kim and Y. Song. “Homomorphic encryption for arithmetic of approximate numbers”. In: *Lecture Notes in Computer Science*. Vol. 10624. Springer Science+Business Media, Berlin, Germany, pp. 409-437, 2017.

[5] P. Sha and Z. Zhu. “The modification of RSA algorithm to adapt fully homomorphic encryption algorithm in cloud computing”. In: *Proceeding 2016 4th IEEE International Conference Cloud Computing and Intelligence Systems*. pp. 388-392, 2016.

[6] L. Chen and Z. Zhang. “*Bootstrapping Fully Homomorphic Encryption with Ring Plaintexts Within Polynomial Noise*. Vol. 2. Conference Paper, pp. 285-304, 2017.

[7] C. Gentry. “*A Fully Homomorphic Encryption Scheme*”. Dissertation, p. 169, 2009.

[8] V. Kumar and N. Srivastava. “Chinese Remainder Theorem based Fully Homomorphic Encryption over Integers”. *International Journal of Applied Engineering Research*, vol. 14, no. 2, pp. 203-208, 2019.

[9] M. A. Mohammed and F. S. Abed. “A symmetric-based framework for securing cloud data at rest”. *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 1, pp. 347-361, 2019.

[10] K. Gai, M. Qiu, Y. Li and X. Y. Liu. “Advanced fully homomorphic encryption scheme over real numbers”. In: *Proceeding 4th IEEE International Conference Cyber Secur Cloud Computing CSCloud 2017 3rd IEEE Intertnational Conference Scalable Smart Cloud*, *SSC 2017*, pp. 64-69, 2017.

[11] C. Gentry, S. Halevi and N. P. Smart. “Fully homomorphic encryption with polylog overhead”. In: *Lecture Notes in Computer Science.* Vol. 7237. Springer Science+Business Media, Berlin, Germany, pp. 465-482, 2012.

[12] J. Fan and F. Vercauteren. “Somewhat practical fully homomorphic encryption”. In: *Proceeding 15th International Conference Practice Theory Public Key Cryptogr*, pp. 1-16, 2012.

[13] Z. Brakerski and V. Vaikuntanathan. “Fully homomorphic encryption from ring-LWE and security for key dependent messages”. In: *Lecture Notes in Computer Science*. Vol. 6841. Springer, Berlin, Germany, pp. 505-524, 2011.

[14] X. Cao, C. Moore, M. O’Neill, N. Hanley and E. O’Sullivan. “High-speed fully homomorphic encryption over the integers”. In: *Lecture Notes in Computer Science*. Vol. 8438. Springer, Berlin, Germany, pp. 169-180, 2014.

[15] C. Xiang and C. M. Tang. “Improved fully homomorphic encryption over the integers with shorter public keys”. *International Journal of Security and its Applications*, vol. 8, no. 6, pp. 365-374, 2014.

[16] M. M. Potey, C. A. Dhote and D. H. Sharma. “Homomorphic encryption for security of cloud data”. *Procedia Computer Science*,

vol. 79, pp. 175-181, 2016.

[17] K. Gai and M. Qiu. “Blend arithmetic operations on tensor-based fully homomorphic encryption over real numbers”. *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3590-3598, 2018.

[18] S. S. Hamad and A. M. Sagheer. “Design of fully homomorphic encryption by prime modular operation”. *Telfor Journal*, vol. 10, no. 2, pp. 118-122, 2018.

[19] S. S. Hamad and A. M. Sagheer. “Fully homomorphic encryption based on Euler’s theorem”. *The International Journal of Information Security*, vol. 9, no. 3, p. 83, 2018.

[20] V. Kumar, R. Kumar, S. K. Pandey and M. Alam. “Fully homomorphic encryption scheme with probabilistic encryption based on euler’s theorem and application in cloud computing”. In: *Advances in Intelligent Systems and Computing.* Vol. 654. Springer, Berlin, Germany, pp. 605-611, 2018.

[21] R. F. Hassan and A. M. Sagheer. “A proposed secure cloud environment based on homomorphic encryption”. *International Advanced Research Journal in Science, Engineering and Technology*, vol. 6, no. 5, pp. 166-175, 2019.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Enabling Accurate Indoor Localization Using a Machine Learning Algorithm

Haidar Abdulrahman Abbas[1]*, Kayhan Zrar Ghafoor[2]

[1]Department of Computer¸ College of Science, University of Sulaimani, Sulaymaniyah, Iraq, [2]Department of Software Engineering, University of Salahaddin, Erbil, Iraq

ABSTRACT

In this paper, fingerprint referencing methods based on wireless fidelity Wi-Fi received signal strength (RSS) have used for indoor positioning. More precisely, Naïve Bayes, decision tree (DT), and support vector machine (SVM) one-to-one multi-classes and error-correcting-output-codes classifier are to enable accurate indoor positioning. Then, normalization is used to reduce positioning error by reducing the fluctuation and diverse distribution of the RSS values. Different devices are used in this experiment; the training dataset is not included in the main dataset. Nonetheless, the learned model by the SVM algorithm cannot be affected by the elimination of train datasets of the test device. The efficiency of DT is lower than the other machine learning algorithms, because it performs by Boolean function, and it provides the low accuracy of prediction for dataset than the algorithms. Naïve Bayes technique based on Bayes Theorem is better than DT and close to SVM for positioning approves that 1–1.5 m positioning accuracy for indoor environments can be achieved by the proposed approach which is an excellent result than traditional protocol.

Index Terms: Received Signal Strength, Wireless Access Points, Wireless Fidelity Fingerprinting, Indoor Localization, Decision Tree, Naïve Bayes, Support Vector Machine

1. INTRODUCTION

In recent years, indoor localization became very popular due to the extensive range of applications [1]. Global navigation satellite systems and global positioning systems are used to enable accurate location, but they failed in indoor environments because of the low received signal power and satellite visibility in such places, underwater, inside building, caves, and tunnels [2]. These technologies need an open environment to work properly [3]. Signal strengths of wireless fidelity (Wi-Fi) Apps to fingerprint the location can be used in Wi-Fi fingerprint-based localization systems, these signals are collected in the location and become the mainstream results for indoor localization. Wi-Fi fingerprint-based localization methods have two main phases offline fingerprinting phase and online localization phase Fig. 1. The offline phase is utilized in building a Wi-Fi fingerprint map by a site survey and save it at a database, and the online localization phase is used to locate the mobile devices by examining the received Wi-Fi signals with the fingerprint map [4].

Access this article online

DOI:10.21928/uhdjst.v4n1y2020.pp96-102	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Recently, users utilize mobile devices (e.g., smartphones) to access Wi-Fi networks in indoor environments (e.g., shopping malls). The investigation of indoor localization methods utilizing signals has increased widely [5]. Moreover, these methods are profitable because it does not require extra tools. One of the best advantages of location fingerprinting is capable of taking the benefits of multipath and non-line of sight problems in an indoor environment, as they truly assistant RSS to be distinct at dissimilar points of the area [3]. While there are several valuable features in fingerprint-based

Corresponding author's e-mail: Haidar Abdulrahman Abbas, Department of Computer¸ College of Science, University of Sulaimani, Sulaymaniyah, Iraq. E-mail: haidar.abbas@univsul.edu.iq

Received: 08-05-2020 Accepted: 18-06-2020 Published: 27-06-2020

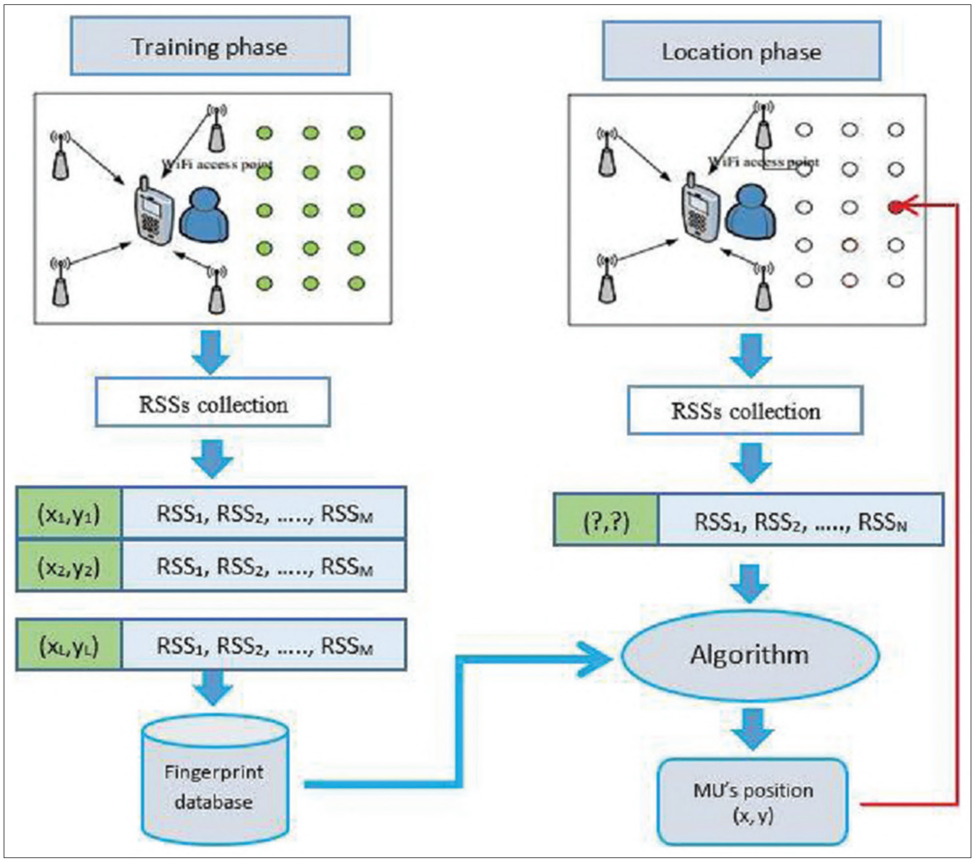


Fig. 1. Constructing the wireless fidelity fingerprinting map [7].

localization, building fingerprint landmarks for a huge area requires an important amount of time and human resources. Database fingerprint can be altered by environmental influences, time, and different devices, so it is necessary to update frequently.

Important studies have been dedicated to the online localization phase; decimeter-level localization efficiency can be obtained by utilizing advanced algorithms which are used to collect online Wi-Fi signals with the fingerprint map [6]. This field of study has been concerned by the researcher in both industry and academia. To collect RRS and fingerprint to assessment, the target location machine learning algorithms can be used such as deep learning and K-Nearest-Neighbor (K-NN) [7].

In this study, fingerprint methods utilizing the Wi-Fi strength signal is presented for indoor positioning. To decrease the positioning errors, Naive Bayes, decision tree (DT), and support vector machine (SVM) one-to-one multi-classes and error-correcting output codes (ECOC) classifiers are proposed and the contrast among these methods. Normalization is used to reduce errors positioning in the values of RSS because of instability and diverse distributions values of RSS. Different devices are used in this experiment when the train data set is not involved in the main dataset. Nonetheless, the learned model by the SVM algorithm cannot be affected by the elimination of train datasets of the test device. The efficiency of DT is lower than the other machine learning algorithms, because it performs by Boolean function, and it provides the low accuracy of prediction for dataset than the other algorithms. Naïve Bayes techniques based on Bayes Theorem is better than DT and close to SVM for positioning accuracy. SVM error positioning approves that 1–1.5 m positioning accuracy for indoor environments can be achieved by the proposed approach which is an excellent result than traditional protocol.

2. RELATED WORK

Recently with the development of computing and the popularity of location-based services, many types of research have considered the improvement of the indoor localization system. Some of these researches focus on the designing system for specific applications that requires a high efficiency (e.g., in the order of centimeters) [8]. Normally, developing these systems need devoted hardware with a huge application cost. Contrarily, several kinds of research have focused on general location-based services where the necessity of accuracy in the form of meters.

Wi-Fi strength signal is used by the fingerprinting method for indoor positioning. To decrease the positioning errors, the improved form of nearest neighbor algorithm is suggested which is called NK-NN, multipath, and RSS variations created the new form of NK-NN, which are utilized the basic KNN and it is variant. In the RSS testing sample, the noise can be removed by compared each testing sample to each fingerprint and based on the minimum distance, the sample is chosen for the position's calculation. After that, the process of classification is operated on the Kth-nearest training sample of diverse reference points which assistance to trim the noise of RSS training and preventing them from the localization. In the experimental outcome, the NK-NN method has better performance than other similar methods [7].

Other studies used Convolutional Neural Network (CNN)-based Wi-Fi fingerprinting for indoor localization. It can be seen that the achievement in image classifications, the suggested method can be potent to minor changes of received signals as it uses the radio map topology as well as signal strengths. In the suggested method, based on the one-dimensional Wi-Fi signals, the two-dimensional (2-D) virtual radio map is built (e.g., received signal strength indicator values) and later a CNN utilizing 2-D radio map is designed as inputs. Consequently, the proposed method is learned the signal strengths as well as the topology radio map. To enhance the efficiency of the suggested method utilizing different improving techniques as feature scaling, dropout, data balancing, and ensemble [6].

To enhance the accuracy of positioning systems, many approaches have been studies and focused on long short-term memory (LSTM) networks. A deep neural network is utilized to improve the efficiency of positioning methods which is acceptable for handling sequential datasets. Therefore, LSTM modes are used because they can recognize the dependency of long-term that is existing in the Wi-Fi data that can be seen from the deep recurrent model's performance. The architecture of RNN and LSTM can recognize the dependency of long-term and utilizing them for later prediction. It is good to examine the previous landmark position to an exact estimate of points on the radio map. The main aim of implementing RNN is to guarantee for providing better performance of recurrent networks on the Wi-Fi dataset. Vanilla LSTM is the primary model that has a good enhancement by 47.8% over the KNN and 10.2 enhancements over RRN utilizing the complete dataset. The efficiency of Vanilla LSTM is even developed after updated to 3-Stacked LSTM. The improvement of 3-Stacked LSTM is 74.4% over the KNN and 18.1% over Vanilla LSTM [9].

There is a rich theoretical basis that is prepared by the Statistical Learning Theory for developing the model, starting a set of examples. In a specific Wi-Fi, the wireless has a signal strength measurement for standard functioning mode so that no particular hardware is desired. SVM is designed and compared to other approaches examined in the scientific literature on the equivalent data set. Experiments executed in the real-world environment illustrate that the outcomes are comparable, with the benefits of low algorithms complication in the standard functioning phase. Furthermore, the algorithm performed better than the other techniques which are mainly appropriate for classification [10].

3. METHODOLOGY

The localization algorithm is fundamentally used for making RSS related radio maps in a designated indoor environment as well as converting localization problem into an optimization problem: Obtaining RSS value measurements of an undisclosed location, the function can help in estimating the location when used in reverse order. While using a fingerprinting technique in the online phase, identical smartphones should be used in both buildings, the RSS dataset as well as testing. Using different smartphones would worsen the accuracy of the calculated position. To eliminate this problem, we propose a DT, Naive Base, and SVM model or adapt the nature of the calculated RSS values among multi-smartphones. The model is directed at various types of smartphone measurements by adopting a machine learning algorithm.

First, gathered RSS values at all the identified close positions are normalized, the normalization is achieved by subtracting the mean value of the gadgets engaged in training the model and then dividing the results by the standard deviation of the aforementioned gadgets. Before normalizing the RSS of wireless access points (AP), the error positioning high because of fluctuation and heterogeneous distribution of the RSS values and applying normalization to decrease the variation of the value and rescale the RSS value within a uniform distribution. RSS vector will be filled with zeroes for those APs. The normalized RSS values can then be applied to train and test the algorithms.

3.1. DT

DTs are a non-parametric method that belongs to the supervised learning algorithms family. It is for classification and regression [11]. In this algorithm, a binary DT is developed from the training data set. In the beginning, basic decision rules derived from the data features are learned. It operates in three nodes; root node, internal node, and a leaf or terminal node. The terminal node has a single receiving edge and zeroes an outgoing edge. The internal has two edges, one for incoming and one for outgoing, while the root node can have zero or more outgoing edges but does not have any incoming edges. Each leaf node is given a class label. Each node is related to a decision performed on the inputs. Next, the node is split into new subsets, one for each of the node's sub-trees, in such a way that the same target location is in the same subsets [11]. The algorithm halts upon finding a pure decision meaning each node's data subset has a single target location and when uncertainty is inefficient.

3.2. Na ve Bayes

The crux of this theorem is derived from the Bayes theorem. According to the Naïve Bayes theorem, all features are independent of each other. While this assumption is usually not true in real-world applications, yet Natives Bayes have had positive results in certain scenarios, mostly when there is a small number of training samples [12]. With an end goal to accomplish high accuracy while decreasing pre-deployment trials, we select this strategy for processing the probabilities of the locations' given measurements. The event with the most elevated likelihood is considered as the candidate. Naive Bayes classifier depends on two fundamental assumptions: (1) The features do not affect each other and (2) the prominence of all the features is equal [13].

3.3. SVM

SVMs [14], [15] are non-parametric supervised learning models with related learning algorithms that analyze data used for pattern recognition problems. SVMs are applied in the localization system by training the support vectors on a radio map that consists of grid points. SVMs study the association between the trained fingerprints and their grid points by taking into account each grid point as a class. This method can be expanded to multiple class classification rather than just two classes. In our training dataset, we have 105 classes, so we used the ECOC one-to-one SVM which is used to classification when classes are more than two after representing the training data by mapping the data to the feature space. The SVM algorithms identify hyperplane, which separates the support vector trained with a distance [14].

4. PERFORMANCE EVALUATION

In this section, designing radio maps of RSS in the studied indoor environment and positioning the difficulty as an optimization problem is presented which is the main idea behind the localization algorithm: providing RSS value measurements of a new position, the function is reversed to determine the evaluated position. Localization fingerprinting methods utilized two main phases (online positioning phase and offline training phase). Afterward, the fingerprint is collected and builds our dataset. The dataset consists of the true location of pre-selected positions and equivalent RSS of nearby AP [16]. The approach that we proposed illustrates the decision tree, Naïve Bayes, and SVN model and compares those models, normalization was applied to fingerprint landmarks.

4.1. Dataset and Simulation Setup

Our RSS data are collected from KIOS Research Center which is a 560 m^2 office environment. This center has many open cubicle-style and special offices, labs, and conference rooms. Wireless LAN standard has been used to install nine local Apps and offer full coverage all over the floor. We utilize five diverse mobiles to collect our data including, HP iPAQ hw6915 personal digital assistant with Windows Mobile, an Asus EeePC T101MT laptop running Windows 7, and HTC Flyer Android tablet and two other Android smartphones (HTC Desire, Samsung Nexus S). We use fingerprinting for our training data, documenting these fingerprints have RSS measurement from entire existing APs, at 105 separate reference positions by carrying all five devices concurrently. We utilize each device to collect fingerprints, 2100 training fingerprints are available, equivalent to 20 fingerprints per reference position. For building device-specific radio maps, these data are utilized by computing the measure of mean values RSS that analogous to each reference position. We indicate that the device-specific radio maps are only required to estimate purposes. After 2 weeks, we utilized a predefined router to gather more test data by walking forward to the router. The router contains two segments and 96 positions; most of them do not concur with the mention position. Each router is tested 10 times using all devices concurrently, while one fingerprint was documented at all test positions [17].

MATLAB toolbox is used to estimate the performance of models. In the first scenario, we tested the DT, Naïve Bayes, and SVM as matching algorithms without applying the RSS normalization (mean and standard deviation) parameters while the device is tested, the training dataset of the device was excluding. The second scenario was using the machine algorithms by applying the RSS normalization, like the previous scenario, the testing device, the training dataset was excluding.

4.2. Simulation Results

We used root mean square error between the estimated and the true locations to evaluate the localization accuracy DT, Naïve Bayes, and SVM algorithms, while there are many methods to evaluate the accuracy.

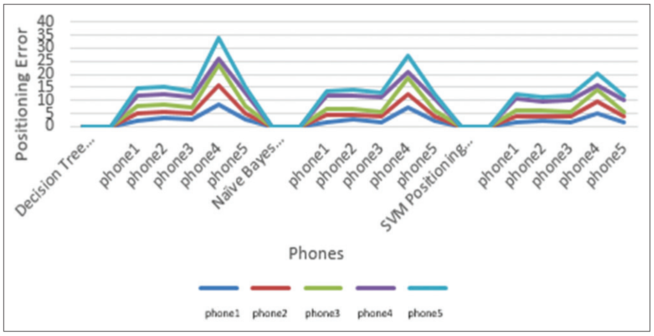


Fig. 2. Decision tree, Naïve Bayes, support vector machine positioning accuracy when normalization not applied.

TABLE 1: Positioning accuracy of DT, Naïve Bayes, and SVM algorithms when normalization not applied

Decision tree positioning accuracy when normalization not applied					
	Phone 1	Phone 2	Phone 3	Phone 4	Phone 5
Phone 1	2.203	2.8795	2.5406	4.4574	2.4796
Phone 2	3.4706	2.1509	2.7549	3.7918	2.9944
Phone 3	2.8008	2.5236	1.9669	3.9633	2.3818
Phone 4	8.4164	7.2392	7.9662	2.4191	8.0881
Phone 5	2.5555	2.6016	2.479	5.1455	2.6042
Naïve Bayes positioning accuracy when normalization not applied					
	Phone 1	Phone 2	Phone 3	Phone 4	Phone 5
Phone 1	1.8598	2.6695	2.0898	4.9985	1.8697
Phone 2	2.7418	1.8852	2.1934	5.0275	2.4532
Phone 3	1.7627	2.2526	1.5297	5.6689	1.8885
Phone 4	7.4514	5.1601	6.2018	1.8415	6.4267
Phone 5	2.0015	2.1527	1.9874	4.3407	1.7289
SVM positioning accuracy when normalization not applied					
Phone	1.748	2.3896	1.9668	4.4945	2.0451
Phone	2.3941	1.7739	1.9912	3.2393	2.083
Phone	1.7156	2.3555	1.4941	4.344	1.9079
Phone	5.0985	4.2347	4.6483	1.8507	4.7268
Phone	1.791	2.0599	1.9353	4.1604	1.6278

DT: Decision tree, SVM: Support vector machine

The average positioning accuracy of the first scenario shown in Table 1 which contains all the devices tested. The DT is generally represented by Boolean function and gives a dataset low prediction accuracy compared to other machine learning algorithms. The SVM has better positioning of the accuracy than other algorithms Fig. 2.

The positioning accuracy for all algorithms is higher in the second scenario after normalization exercised on the dataset than before normalization, caused by fluctuation and heterogeneous distribution of the RSS values. The findings are listed in Table 2. Compared to the SVM with DT and Naïve Bayes, we can see that the SVM exhibits more accuracy in positioning. SVM has the most elegant maths behind them and uses the Kernel trick in the dual problem. The results of Naïve Bayes have a second degree in positioning accuracy at each scenario Fig. 3.

Phone 4's positioning accuracy is weaker than most phones, its return to phone 4's RSS values, Phone 4 read the signal strength from −11 to −90 dB, and most are between −11 and −40 dB, unlike other phones. There is a big gap in RSS values where we have categorized phone4 and others, so the positioning accuracy is worse than some, so the error is large.

TABLE 2: Positioning accuracy of DT, Naïve Bayes, and SVM algorithms when normalization applied

Decision tree positioning accuracy when normalization applied					
	Phone 1	Phone 2	Phone 3	Phone 4	Phone 5
Phone 1	2.2143	2.6063	2.3258	2.5884	2.2332
Phone 2	2.0923	2.1043	2.1062	2.3519	2.2207
Phone 3	2.1431	2.2638	1.9752	2.4954	2.213
Phone 4	2.2618	2.5557	2.4014	2.5941	2.4816
Phone 5	2.2688	2.5763	2.3769	2.6545	2.5061
Naïve Bayes positioning accuracy when normalization applied					
	Phone 1	Phone 2	Phone 3	Phone 4	Phone 5
Phone 1	1.9256	2.1813	1.787	2.3594	1.8646
Phone 2	1.9088	1.8985	1.834	2.2155	1.9758
Phone 3	1.7655	1.9025	1.5323	2.0569	1.7883
Phone 4	1.8462	2.009	2.0197	1.8725	1.7213
Phone 5	1.7523	1.9018	1.9706	2.0426	1.7213
SVM positioning accuracy when normalization applied					
	Phone 1	Phone 2	Phone 3	Phone 4	Phone 5
Phone	1.7554	1.982	1.8261	2.2321	1.9163
Phone	1.7485	1.6998	1.6029	1.8001	1.7596
Phone	1.4364	1.853	1.4621	2.2092	1.7774
Phone	1.8266	1.9311	1.8866	1.8432	1.6975
Phone	1.6622	1.963	2.0058	2.0025	1.6215

DT: Decision tree, SVM: Support vector machine

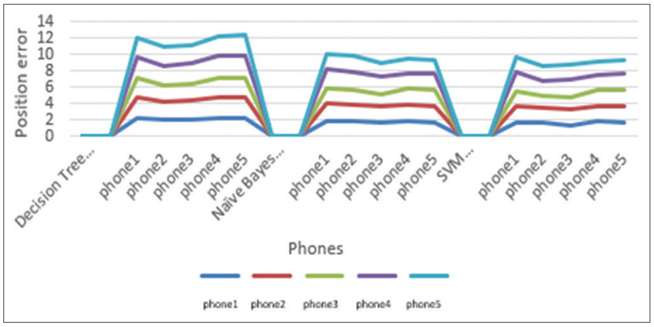


Fig. 3. Decision tree, Naïve Bayes, support vector machine positioning accuracy when normalization applied.

Each of the algorithms has strengthens and weaknesses, if we compare DT to other algorithms, it has less requirement for data pre-processing and not affected by missing values in data set, but any changes in data set impact the structure of it which is lead to instability. NB needs a smaller amount of training data to evaluate the test data, and implementation is easy. However, the main problem of NB is the assumption of independence. The SVM algorithm has more effective when the number of dimensions is greater than the sample number, and it comparatively memory efficient. SVM has a disadvantage like it is not the best option for the large data set and does not execute well with a data set that has more noises. In general, our proposed approach has many advantages; it does not need extra hardware to be installed, and high performance was achieved. Our proposed concept's disadvantages, require more computational work (especially SVM) compared to others by the system. To compare our results with other works, this system has more positioning accuracy. This is a very different outcome than conventional protocol.

5. CONCLUSION

In this research article, RSS fingerprint-based Wi-Fi localization was assessed in regards to the in-operation infrastructure of an indoor environment. We review the modern resolutions for very accurate localization in indoor schemes. Next, we outline the rise in positioning error when dissimilar platform-devices are used in the fingerprinting technique for training and testing the dataset. In addition, RSS measurements produce different values for the same position and time when dissimilar platform-devices are used. We implement the most popular and reliable machine learning algorithms, namely, DTs, Naïve Bayes, and SVM learning algorithms. Examine ensemble estimators that apply multiple algorithms to estimate the position and then we choose a combination the leads to the most efficient performance. SVM error positioning shows that 1–1.5 m positioning accuracy for indoor environments can be accomplished by the presented technique which is an obvious improvement compared to existing approaches. Thus, fingerprinting localizations can utilize RSS data to minimize the notable amount of time and energy.

REFERENCES

[1] J. Xiao, Z. Zhou, Y. Yi and L. M. Ni. "A survey on wireless indoor localization from the device perspective," *ACM Computing Surveys*, vol. 49, no. 2, p. 2933232, 2016.

[2] A. S. Paul and E. A. Wan. "RSSI-Based indoor localization and tracking using sigma-point kalman smoothers," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 5, pp. 860-873, 2009.

[3] N. Alikhani, S. Amirinanloo, V. Moghtadaiee and S. A. Ghorashi. "Fast Fingerprinting Based Indoor Localization by Wi-Fi Signals," *2017 7th International Conference on Computer and Knowledge Engineering*, vol. 2017 Janua, pp. 241-246, 2017.

[4] S. Dai, L. He and X. Zhang. "Autonomous WiFi Fingerprinting for Indoor Localization". In: 2020 *ACM/IEEE 11th International Conference on Cyber-Physical Systems* (*ICCPS*), pp. 141-150, 2020.

[5] F. Li, M. Liu, Y. Zhang and W. Shen."A two-level wifi fingerprint-based indoor localization method for dangerous area monitoring," *Sensors* (*Basel*), vol. 19, no. 19, p. 4243, 2019.

[6] J. W. Jang and S. N. Hong. "Indoor Localization with WiFi Fingerprinting Using Convolutional Neural Network," *International Conference on Ubiquitous and Future Networks*, vol. 2018, pp. 753-758, 2018.

[7] M. Alfakih and M. Keche. "An enhanced indoor positioning method based on Wi-fi RSS fingerprinting," *The Journal of Communications Software and Systems*, vol. 15, no. 1, pp. 18-25, 2019.

[8] C. Chen, Y. Chen, Y. Han, H. Q. Lai and K. J. R. Liu, "Achieving centimeter-accuracy indoor localization on wifi platforms: A frequency hopping approach," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 111-121, 2017.

[9] A. Sahar and D. Han. "An LSTM-based Indoor Positioning Method Using Wi-Fi Signals," *ACM's International Conference Proceedings*, 2018.

[10] M. Brunato and R. Battiti. "Statistical learning theory for location fingerprinting in wireless LANs," *Computer Networks*, vol. 47, no. 6, pp. 825-845, 2005.

[11] Y. Li. "Predicting materials properties and behavior using classification and regression trees," *Materials Science and Engineering A*, vol. 433, no. 1-2, pp. 261-268, 2006.

[12] N. Gutierrez, C. Belmonte, J. Hanvey, R. Espejo and Z. Dong. "Indoor Localization for Mobile Devices," *Proceeding. 11th IEEE International Conference on Sensing Control*, pp. 173-178, 2014.

[13] Z. Wu, Q. Xu, J. Li, C. Fu, Q. Xuan and Y. Xiang. "Passive indoor localization based on CSI and naive bayes classification," *IEEE Transactions on Systems, Man, and Cybernetics Systems*, vol. 48, no. 9, pp. 1566-1577, 2018.

[14] B. Schölkopf. "Slides learning with kernels," *Journal of the Electrochemical Society*, vol. 129, p. 2865, 2002.

[15] T. Joachims. "Transductive Inference for Text Classification Using Support Vector Machines," *Proceeding 20th International Conference on Machine Learning*, 2000.

[16] Z. Zhong, Z. Tang, X. Li, T. Yuan, Y. Yang, M. Wei, Y. Zhang, R. Sheng and N. Grant. "XJTLUIndoorLoc: A New Fingerprinting Database for Indoor Localization and Trajectory Estimation Based on Wi-Fi RSS and Geomagnetic Field," *Proceeding 2018 6th Internationl Symposium Computer Netwwork*, pp. 228-234, 2018.

[17] A. H. Salamah, M. Tamazin, M. A. Sharkas and M. Khedr. "An Enhanced WiFi Indoor Localization System Based on Machine Learning," *2016 International Conference Indoor Position Indoor Navigation*, pp. 4-7, 2016.

REVIEW ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Review Study for Electrocardiogram Signal Classification

Lana Abdulrazaq Abdullah[1,2], Muzhit Shaban Al-Ani[3]

[1]Department of Computer Science, College of Science and Technology, University of Human Development, Sulaymaniyah, KRG, Iraq, [2]Department of Computer, College of Science, University of Sulaimani, Sulaymaniyah, KRG, Iraq, [3]Department of Information Technology, College of Science and Technology, University of Human Development, Sulaymaniyah, KRG, Iraq

ABSTRACT

An electrocardiogram (ECG) signal is a recording of the electrical activity generated by the heart. The analysis of the ECG signal has been interested in more than a decade to build a model to make automatic ECG classification. The main goal of this work is to study and review an overview of utilizing the classification methods that have been recently used such as Artificial Neural Network, Convolution Neural Network (CNN), discrete wavelet transform, Support Vector Machine (SVM), and K-Nearest Neighbor. Efficient comparisons are shown in the result in terms of classification methods, features extraction technique, dataset, contribution, and some other aspects. The result also shows that the CNN has been most widely used for ECG classification as it can obtain a higher success rate than the rest of the classification approaches.

Index Terms: Artificial neural network, Convolution neural network, Discrete wavelet transform, Support vector machine, K-nearest neighbor

1. INTRODUCTION

An electrocardiogram (ECG) is simply a recording of the electrical activity generated by the heart [1]. The heart produces the electrical activity that measures by a medical test called an ECG, which identifies the cardiac abnormality [2]. A heart produces tiny electrical impulses that spread through the heart muscle [3]. An ECG all data about the electrical activity of the heart records and shows on a paper by an ECG machine [4]. Then, a medical practitioner interprets this data; ECG leads to find the cause of symptoms of chest pain and also leads to detect abnormal heart rhythm [5].

Access this article online

DOI:10.21928/uhdjst.v4n1y2020.pp103-117	E-ISSN: 2521-4217 P-ISSN: 2521-4209

An ECG signal has a total of five primary turns, counting P, Q, R, S, and T waves, plus the depolarization of the atria causes a small turn before atria contraction as the activation (depolarization) wave-front propagates from the Sino atria node through the atria [6]. The Q wave is a downward deflection after the P wave [7]. The R wave follows as an upward deflection, and the S wave is a downward deflection following the R wave [8]. Q, R, and S waves together indicate a single event [9]. Hence, they are usually considered to be QRS complex, as shown in Fig. 1 [10], [11].

The features based on the QRS complex are among the most powerful features for ECG analysis [13]. The QRS-complex is caused by currents that are generated when the ventricles depolarize before their contraction [14]. Although atrial depolarization occurs before ventricular depolarization, the latter waveform (i.e., the QRS-complex) has much higher amplitude, and atria depolarization is, therefore, not seen on an ECG. The T wave, which follows the S wave, is ventricular depolarization, where the heart muscle prepares for the next

Corresponding author's e-mail: Lana Abdulrazaq Abdullah, Department of Computer Science, College of Science and Technology, University of Human Development, Sulaymaniyah, KRG, Iraq, Department of Computer, College of Science, University of Sulaimani, Sulaymaniyah, KRG, Iraq. E-mail: lana.abdulla@uhd.edu.iq

Received: 05-02-2020 Accepted: 12-06-2020 Published: 29-06-2020

ECG cycle [15]. Finally, the U wave is a small deflection that immediately follows the T wave. The U wave is usually in the same direction as the T wave [16].

There are different kinds of arrhythmias, and each kind is associated with a pattern, and as such, it is possible to recognize and classify it [17]. The arrhythmias can be categorized into two major classes; the first class consists of arrhythmias formed by a single irregular ECG signal, herein called morphological arrhythmia, the other type consists of arrhythmias formed by a set of irregular heartbeats, herein called rhythmic arrhythmias [18].

The main problem in the process of identifying and classifying arrhythmias ECGs is that an ECG signal can vary for each person, and sometimes different patients have separate ECG morphologies for the same disease [19]. Moreover, two various diseases could have approximately the same properties on an ECG signal [20]. These problems cause some difficulties in the issue of heart disease diagnosis [21].

Furthermore, the ECG records analysis is complicated for a human due to fatigue; an alternative way for automatic classification is computerization techniques [22]. For arrhythmia classification from the signal received by ECG device needed an automated system that can be divided into three main steps, as follows first: Pre-processing, next: Feature extraction and finally: Classification, as shown in Fig. 2 [23].

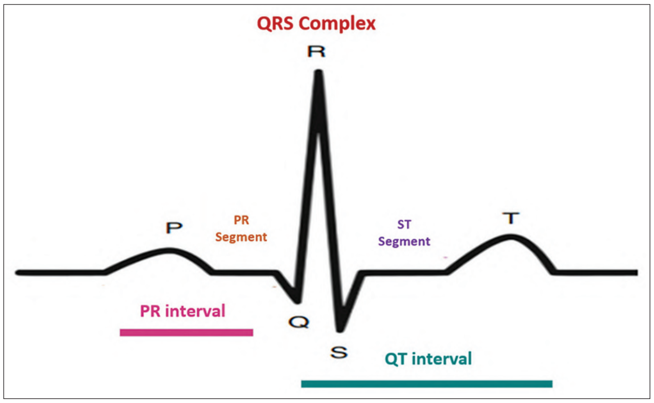


Fig. 1. A typical electrocardiogram signal [12].

ECG signals may contain several kinds of noises, which can affect the extraction of features used for classification; therefore, the pre-processing step is necessary for removing the noises [24]. Researchers have applied different pre-processing techniques for ECG classification. For noise removal, techniques such as low pass linear phase filter and linear phase high pass filters, etc., are used [25]. Some methods, such as median filter, linear phase high pass filter, and mean median filter are used baseline adjustment [26].

After the pre-processing step, extracting different ECG features then used as inputs to the classification model [27]. Feature extraction techniques used by researchers are discrete wavelet transform (DWT), continuous wavelet transform, discrete cosine transform (DCT), discrete Fourier transform, principal component analysis (PCA), Pan-Tompkins algorithm, and independent component analysis (ICA) [28].

When the set of features has been defined from the heartbeats, models can be built from these data using artificial intelligence algorithms from machine learning and data mining domains for arrhythmia heartbeat classification. The most popular techniques employed for this task and found in the literature are artificial neural networks (ANN), convolution neural network (CNN), DWT, support vector machines (SVM), decision tree (DT), Bayesian, Fuzzy, linear discriminate analysis (LDA), and k-nearest neighbors (KNN) [29].

Many surveys on ECG analysis and classification have been published. In Karpagachelvi [30] surveyed the most effective features for ECG analysis and classification as ECG. Features play a significant role in diagnosing most of the cardiac diseases. Nasehi and Pourghassem [31] provided a survey of variance types of seizure detection algorithms and their potential role in diagnostic. Various machine-learning approaches for ECG analysis and classification were reviewed in Roopa and Harish [23]. A comprehensive review was published in 2018, which includes a literature on ECG analysis mostly from the past decade, and most of the major aspects of ECG analysis were addressed such as preprocessing, denoising, feature extraction, and classification methods [16] (Previous works on ECG survey paper, Reviewer 2).

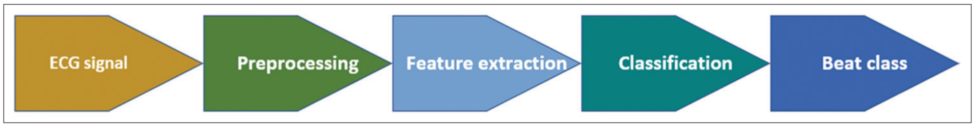


Fig. 2. General diagram of electrocardiogram classification.

The main purpose of this work is to review most of the common techniques that have been used mostly from the past 5 years. Moreover, the paper can be useful for the other researchers in identifying any issue in ECG classification and analyzing the research area as many aspects of the methods are addressed. (This section is the main purpose of the paper (reviewer 3)).

The section of this paper is ordered as follows: Section 2 contains Classification Techniques, and then Section 3 provides of Discussion, and finally, Section 4 presents the Conclusion.

2. CLASSIFICATION

A lot of pathological information about a patient's heart processes can be obtained by studying the ECG signal [32]. There are many approaches have been developed to classify heartbeats as it is essential for the detection of an arrhythmia [33]. Arrhythmias can be divided into two parts, which are life-threatening and non-life-threatening arrhythmias, a long-term ECG classification is required for the diagnosis of non-life-threatening arrhythmias that could be time-consuming and impractical, automatic algorithms exhibit a great aid. Consequently, automatic ECG classification of arrhythmias is one of the most worth studying in the world [34].

There are various classifiers that have been used for ECG classification task. In this paper, most common ECG classification methods are reviewed that were proposed since 2016–2020, these classification methods can be mainly clustered based on the classifiers into several categories such as ANNs, CNN, kNN, SVM, and DWT. All of the reviewed papers were accessed by three well-known publishers, which are IEEE, ScienceDirect, and Springer. (This section was wrote about why and how the authors select the papers for this state (Reviewer 2 and reviewer 3).

Different types of classification techniques are studied to classify ECG data under the variance features, as there are plenty of features in the ECG signal that can be extracted. Some of the classification methods are addressed below.

2.1. ANN

The ANN is an adaptive system with exciting features such as the ability to adapt, learn, and summarize; because ANN's parallel processing, self-organizing, fault-tolerant, and adaptive capabilities make it capable of solving many complex problems, ANN is also very accurate in the classification and prediction of outputs [35]. The neural network (NN) consists of the number of layers; the initial layer has an association as of the system input, and the end layer gives the output of the network [36]. NN s having hidden layers and sufficient neurons can be applied to any limited input-output mapping trouble [37]. The NN model consists of an input layer, the hidden layer, and output layer, as shown in Fig. 3 [38].

Many kinds of literature are published related to the ECG classification based on ANN. Below some of these new approach:

Chen *et al.* (2016) proposed a wavelet-based ANN (W-ANN) method that was based on the wavelet transform. The result illustrated that the W-ANN can provide lower computing time such that reduction time was 49% and cleaner ECG input signal. The method was implemented on the data MIT-BIH arrhythmia database and real ECG signal measurement [39].

Boussaa *et al.* (2016) presented the design of a cardiac pathologies detection system with high precision of calculation and decision, which consists of the mel-frequency coefficient cepstrum algorithms such as fingerprint extractor (or features) of the cardiac signal and the algorithms of ANN multilayer perceptron (MLP) type MLP classifier as fingerprints extracted into two classes: Normal or abnormal. The design and testing of the proposed system are performed on two types of data extracted from the MIT-BIH database: A learning base containing labeled data (ECG normal and abnormal) and another test base containing no-labeled data. The experimental results were shown that the proposed system combines the respective advantages of the descriptor mel-frequency cepstrum coefficient and the MLP classifier [40].

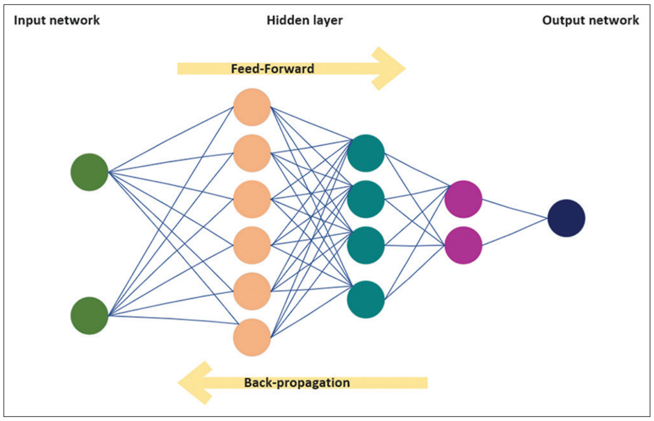


Fig. 3. Artificial neural network.

Savalia *et al*. (2017) distinguished between normal and abnormal ECG data using signal processing and NNs toolboxes in Matlab. Data, which were downloaded from an ECG database, PhysioBank, were used for learning the NN. The feature extraction method was also used to identify variance heart diseases such as bradycardia, tachycardia, first-degree atrioventricular (AV), and second-degree AV. Since ECG signals were very noisy, signal processing techniques were applied to remove the noise contamination. The heart rate of each signal was calculated by finding the distance between R-R intervals of the signal. The QRS complex was used to detect AV blocks. The result showed that the algorithm strongly distinguished between normal and abnormal data as well as identifying the type of disease [41].

Wess *et al*. (2017) presented field-programmable gate array (FPGA)-based ECG arrhythmia detection using an ANN. The objective was to implement a NN-based machine-learning algorithm on FPGA to detect anomalies in ECG signals, with better performance and accuracy (ACC), compared to statistical methods. An implementation with PCA for feature reduction and a MLP for classification, proved superior to other algorithms. For implementation on FPGA, the effects of several parameters and simplification on performance, ACC, and power consumption were studied. Piecewise linear approximation for activation functions and fixed-point implementation was effective methods to reduce the number of needed resources. The resulting NN with 12 inputs and six neurons in the hidden layer, achieved, in spite of the simplifications, and the same overall ACC as simulations with floating-point number representation. An ACC of 99.82% was achieved on average for the MIT-BIH database [42].

Pandey *et al*. (2018) compared three different ANN models for classification normal and abnormal signals and using University of California, Irvine ECG 12 lead signal data. This work had used methods, namely, back propagation (BP) network, radial basis function (RBF) networks, and recurrent neural network (RNN). RNN models have shown better analysis results. ACC for testing classification was 83.1%. This result was better than some work, using the same database [43].

Sannino and Pietro (2018) proposed an approach based on a deep neural network (DNN) for the automatic classification of abnormal ECG beats, differentiated from normal ones. DNN was developed using the Tensor Flow framework, and it was composed of only seven hidden layers, with 5, 10, 30, 50, 30, 10, and 5 neurons, respectively. Comparisons were made among the proposed model with 11 other well-known classifiers. The numerical results showed the effectiveness of the approach, especially in terms of ACC [44].

Debnath *et al*. (2019) proposed two schemes; at first, the QRS components have been extracted from the noisy ECG signal by rejecting the background noise. This was done using the Pan-Tompkins algorithm. The second task involved the calculation of heart rate and detection of tachycardia, bradycardia, asystole, and second-degree AV block from detected QRS peaks using MATLAB. The results showed that from detected QRS peaks, and arrhythmias, which are based on an increase or decrease in the number of QRS peaks, the absence of a QRS peak, could be diagnosed. The final task is to classify the heart abnormalities according to previously extracted features. The BP trained feed-forward NN has been selected for this research. Here, data used for the analysis of ECG signals are from the MIT database [45].

Abdalla *et al*. (2019) presented that approach was developed based on the non-linearity and nonstationary decomposition methods due to the nature of the ECG signal. Complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) was used to obtain intrinsic mode functions (IMFs). Established on those IMFs, four parameters have been computed to construct the feature vector. Average power, coefficient of dispersion, sample entropy, and singular values were calculated as parameters from the first six IMFs. Then, ANN was adopted to apply the feature vector using them and classify five different arrhythmia heartbeats downloaded from PhysioNet in the MIT–BIH database. The performance of the CEEMDAN and ANN was better than all existing methods, where the sensitivity (SEN) is 99.7%, specificity (SPE) is 99.9%, ACC is 99.9%, and receiver operating characteristic (ROC) is 01.0% [46].

2.2. Convolutional Neural Network (CNN)

The CNN is the most common technique to classify ECG, CNN is mainly composed of two parts, feature extraction and classification [47]. The section of feature extraction is responsible for extracting effective features from the ECG signals automatically, while the part of classification is in charge of classifying signals accurately by making use of the extracted features, as shown in Fig. 4 [48].

Many approaches are published the ECG classification based on CNN. Below some of these update works:

Zubair *et al*. (2016) proposed a model which was integrated into two main parts, feature extraction, and classification.

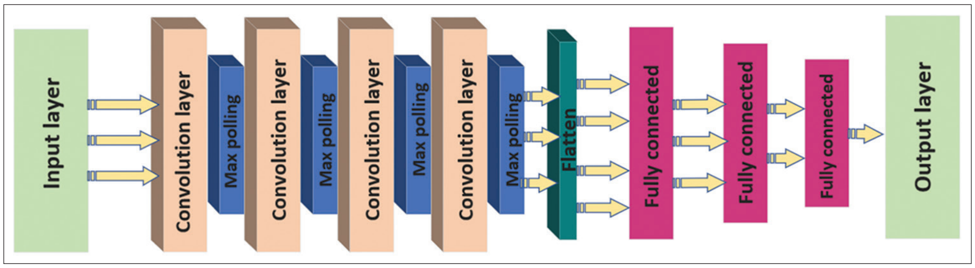


Fig. 4. Typical convolution neural network structure.

The model automatically remembers a suitable feature representation from raw ECG data and thus negates the need for hand-crafted features. Using small and patient-specific training data, the proposed classification system efficiently classified ECG beats into five different classes. ECG signal from 44 recordings of the MIT-BIH database is used to assess the classification performance, and the results demonstrate that the proposed approach achieves a significant classification ACC and superior computational efficiency than most of the state-of-the-art methods for ECG signal classification [49].

Yin *et al*. (2016) proposed a system that applies the impulse radio ultra-wideband radar data as additional information to assist the arrhythmia classification of ECG recordings in the slight motion state. Besides, this proposed system employs a cascaded CNN to achieve an integrated analysis of ECG recordings and radar data. The experiments are implemented in the Caffe platform, and the result reaches an ACC of 88.89% in the slight motion state. It turns out that this proposed system keeps a stable ACC of classification for normal and abnormal heartbeats in the slight motion state [50].

Oh *et al*. (2017) designed a nine-layer deep CNN DCNN to identify five different categories of heartbeats in ECG signals automatically. The test was applied in original and ECG signals that were derived from the available database. The set was artificially augmented for removing high-frequency noise. The CNN model was trained to utilize the augmented data and obtained an ACC of 93.47% and 94.03% in the identification of heartbeats in noise-free and original ECGs [51].

Zhai and Tin (2018) proposed an approach based on the CNN model with a different structure. The model was improved SEN, and positive predictive rate for S beats by more than 12.2% and 11.9%, respectively. The system provided a fully automatic tool and reliable to detect the arrhythmia heartbeat without any manual feature extraction or any expert assistant [52].

Zhang *et al*. (2019) introduced a new pattern recognition method in ECG data using DCNN. Different from past methods that utilized learn features or hand-crafted features from the raw signal domain, the proposed method was learned the features and classifiers from the time-frequency domain. First, the ECG wave signal was transformed into the time-frequency domain using the Short-Time Fourier Transform. Then, several scale-specific DCNN models were trained on ECG samples of a specific length. Eventually, an online decision fusion method was proposed to fuse decisions at different scales into a more accurate and stable one [53].

Wang (2020) proposed a novel approach for the automated atria fibrillation (AF) detection based DNN, which was built 11-layers. The network structure was combined using a modified Elman neural network (MENN) and CNN. Ten-Fold cross-validation was conducted to evaluate the classification performance of the model on the MIT-BIH AF database. The result confirmed that the model yielded excellent classification performance with the ACC, SEN, and SPE of 97.4%, 97.9%, and 97.1%, respectively [54].

Yao *et al*. (2020) designed model attention based on time-incremental CNN (ATI-CNN); a DNN model could obtain both spatial and temporal fusion of information from ECG signals using integrating CNN. The features were flexible input length, halved parameter amount as well as more than 90% computation reduction in real-time processing. The experiment result showed that ATI-CNN achieved an overall classification rate of 81.2% compared to VGGNET that is a classical 16-layer CNN, ATI-CNN achieved ACC increases of 7.7% in average, and up to 26.8% in detecting paroxysmal arrhythmias [55].

2.3. DWT

The DWT is used to recognize and diagnose the ECG signals and widely used in signal processing [56]. A perfect time resolution is the main advantage of DWT [57]. It provides good frequency resolution at low frequency and good resolution at high frequency [58]. The DWT can reveal the local characteristics of the input signal because of this great time and frequency localization ability [59].

Many kinds of literature are published related to the ECG classification based on DWT. Below some of these new approach:

Desai *et al*. (2015) described a machine learning-based approach for detecting five classes of ECG arrhythmia beats based on DWT features. Moreover, ICA was used to comprise dimensionality reduction. ANOVA approach was used to select significant features, and ten-fold cross-validation was used to perform SVM. The experiment was conducted on MIT–BIH arrhythmia, which is grouped into five classes of arrhythmia beats, namely, non-ectopic (N), ventricular ectopic (V), supraventricular ectopic (S), fusion (F), and unknown (U). Using SVM quadratic kernel classified ECG features with an overall average ACC of 98.49% [60].

Saraswat (2016) explored diverse possibilities of the decomposition using the DWT method to classify Wolff Parkinson White Syndrome ECG signals. In this work, ECG signals are discretely sampled till the 5th resolution level of the decomposition tree using DWT with Daubechies wavelet of order 4 (db4), which helps in smoothing the feature more appropriate for detecting changes in signals. The MIT-BIH database was used for some experimental results [61].

Alickovic and Subasi (2016) noted that RF classifiers achieved superior performances compared to DT methods using ten-fold cross-validation for the ECG datasets. The results suggested that further significant developments in words of classification ACC could be accomplished by the proposed classification system. Accurate ECG signal classification was the major requirement for the detection of all arrhythmia types. Performances of the proposed system were evaluated on two different databases, namely, MIT-BIH database and St. Petersburg Institute of Cardiological Techniques 12-lead Arrhythmia Database. For the MIT-BIH database, the RF classifier generated an overall ACC of 99.33 % against 98.44 and 98.67 %, respectively. For St. Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database, RF classifier yielded a general ACC for the C4.5 and CART classifiers of 99.95% against 99.80% for both C4.5 and CART classifiers, respectively. The merged model with multiscale PCA de-noising, DWT, and RF classifier also achieves good performance for MIT-BIH database with the area under the ROC curve (area under the curve [AUC]) and F-measure equal to 0.999 and 0.993 and 1 and 0.999 for and St. Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database, respectively. The results demonstrated that the proposed system was able for reliable classification of ECG signals and to help the clinicians to make an accurate diagnosis of cardiovascular disorders (CVDs) [62].

Pan *et al*. (2017) proposed a comprehensive approach based on random forest techniques and discrete wavelet for arrhythmia diagnosis. Specifically, DWT was used to remove high-frequency noise and baseline drift, while DWT, autocorrelation, PCA, variances, and other mathematical methods are used to extract frequency-domain features, time-domain features, and morphology features. Moreover, an arrhythmia classification system was developed, and its availability was verified that the proposed scheme could significantly be used for guidance and reference in clinical arrhythmia automatic classification [63].

Sahoo (2017) proposed an improved algorithm to find QRS complex features based on the wavelet transform to classify four kids of ECG beats: Normal (N), left bundle branch block (LBBB), right bundle branch block (RBBB), and Paced beats (P); using NN and SVM classifier. Model performance was evaluated in terms of SEN, SPE, and ACC for 48-recorded ECG signals obtained from the MIT–BIH arrhythmia database. The proposed procedure achieved high detection efficiency with a low error rate of 0.42% when detecting the QRS compound. The classifier fixed its superiority with an average ACC of 96.67% and 98.39% in SVM and NN, respectively. The classification ACC of the SVM approach proves superior for the proposed method to that of the NN classifier with extracted parameters in detecting ECG arrhythmia beats [64].

Ceylan (2018) studied a model based on spared coefficients of the signals that were achieved by employing sparse representation algorithms and dictionary learning. The obtained coefficients were utilized in the weight update process of three different classification approaches, which were created using SVM, AdaBoost, and LDA algorithms. In the first step, the proposed Dictionary Learning (DL) based AdaBoost classifiers isolated the ECG signals. Then, the selected feature was applied to ECG signals, and six different feature subsets were obtained by DWT, T-test, Bhattacharyya, First Order Statistics (FOS), Wilcoxon test, and Entropy methods. The subscription of objects was used as a new dataset. The classification process is performed according to the proposed method, and satisfactory results are obtained. The best classification ACC was received at 99.75% using the proposed commercial-based terminology method called DL-AdaBoost-SVM for the subset of attributes obtained using the DWT and Wilcoxon test methods [65].

Tea and Vladan (2018) proposed a novel framework that combined the theory of compressive sensing and random forests to achieve reliable automatic cardiac arrhythmia detection. Moreover, it evaluated the characterization power of DCT, DWT, and FFT data transformations to extract significant features that can bring an additional boost to the classification performance. The experiments conducted on the MIT-BIH benchmark arrhythmia database, the result demonstrated that DWT based features exhibit better returns compared to the feature extraction technique for a relatively small number of random projected coefficients. Furthermore, due to its low-complexity, the proposed model could be implemented for practical applications of real-time ECG monitoring [66].

Zhang *et al*. (2019) proposed a lightweight approach to classify five types of cardiac arrhythmia; namely, normal beat (N), premature ventricular contraction (PVC) (V), atria premature contraction (APC) (A), RBBB beat (R), and LBBB beat (L). The mixed method of frequency analysis and Shannon entropy was applied to extract appropriate statistical features. The information gain criterion was manipulated for selecting features. The selected features were then fed to the input of Random Forest, KNN, and J48 for classification. To evaluate classification performance, ten-fold cross-validation was used to verify the effectiveness of our method. Experimental results showed that the Random Forest classifier demonstrates significant performance with the SPE of 99.5%, the highest SEN of 98.1%, and the ACC of 98.08%, outperforming other representative approaches for automated cardiac arrhythmia classification [67].

Kora *et al*. (2019) showed that an algorithm to detect atrial fibrillation (AF) in the ECG signal is developed. For correct detection of AF, pre-processing and feature extraction of the ECG signal shall be performed before it detects AF. After considering the ECG signal from the database, in the pre-processing stage, denoising of the ECG signal is carried out to obtain a clean ECG signal. After pre-processing, before feature extraction, R peak detection is carried out for the signal. Since R peak has the highest amplitude, and therefore, it is detected in the first round, and subsequently location of other peaks of the ECG signals is performed. After completing, pre-processing and feature extraction using DWT applied based on inverted T wave logic and ST-segment elevation. Our classification algorithm was demonstrated to successfully acquire, analyze, and interpret ECGs for the presence of AF, indicating its potential to support m-Health diagnosis, monitoring, and management of therapy in AF patients [68].

2.4. SVM

SVM is a learning algorithm that has many good properties. It is associated with data analysis and recognizes the pattern. SVM uses a linear discriminate function for classification; however, non-linear classification can also be done if a non-linear kernel is used [69]. SVM performs well in real-time situations, robust, easy to understand. While compared to other classifiers [30]. A classification task typically requires the knowledge about the data to be classified; hence, the classifier must be trained before classifying any data [70]. One of the main advantages of the SVM classifier is that it automatically finds the support vectors for better classification [71]. Majorly, in every case the performance of SVM depends on the affected kernel function selection [72].

Many types of research are published in the ECG classification based on the SVM. Below some of these recent studies:

Elhaj *et al*. (2016) investigated a combination of linear and non-linear features to improve the classification of ECG data. In the study, five types of beat classes of arrhythmia as recommended by the Association for Advancement of Medical Instrumentation are analyzed: Non-ectopic beats (N), supra-ventricular ectopic beats (S), ventricular ectopic beats (V), fusion beats (F), and unclassifiable and paced beats (U). The characterization ability of non-linear features such as high order statistics and cumulants and non-linear feature reduction methods such as ICA is combined with linear features, namely, the PCA of DWT coefficients. The features are tested for their ability to differentiate different classes of data using different classifiers, namely, the SVM and NN methods, with tenfold cross-validation. This method can classify the N, S, V, F, and U arrhythmia classes with high ACC (98.91%) using a combined SVM and RBF method [73].

Arjunan (2016) reported that statistics features could be useful for categorizing the ECG signals. Like the first, the signal has been passed from the de-noising process as a pre-processing. Then, the following statistics features such that mean, variance, standard deviation, and skewness are extracted from the signal. SVM was implemented to classify the ECG signal into two categories; normal or abnormal. The results show that the system classifies the given ECG signal with 90% SEN and SPE [74].

Smíšek *et al*. (2017) proposed method for automatic ECG classification to four classes (normal rhythm [N], AF [A], another rhythm [O], and noisy records [P]). The SVM approach was involved in the two different stages in the model. In the first stage, SVM was used to extract the global features from the entire ECG signal. In the second stage, the features from the previous step were used to train the second SVM classifier. The cross-validation technique was used to evaluate both classifiers. The result showed that in Phase II of challenge, the total F1 score of the method was 0.81 and 0.84 within the hidden challenge dataset and training set, respectively [75].

Wu *et al*. (2017) developed a system for identifying excessive alcohol consumption. Three sensors were used to acquire signals regarding (ECG), intoxilyzers, and photoplethysmograph (PPG). Intoxilyzers were used to know alcohol consumption levels of participants before and after drinking. The signals were pre-processed, segmented, and subjected to feature extraction using specific algorithms to produce ECG and PPG training and test data. Using the ECG, PPG, and alcohol consumption data, the developed model was fast and accurate for the identification scheme using the SVM algorithm. Using the training data for training and the test data were applied to comfort the recognition performance of the trained SVMs. The identification performance of the proposed classifiers achieved 95% on average. In the approach, different feature combinations were tested to select the optimum technological configuration. Because the PPG and ECG features produce identical classification performance and the PPG features were more convenient to acquire, the technical setting based on PPG is preferable for developing smart and wearable devices for the identification of driving under the influence [76].

Venkatesan *et al*. (2018), ECG signal pre-processing and SVM -based arrhythmic beat classification is performed to categorize into normal and abnormal subjects. In ECG signal pre-processing, a delayed error normalized LMS adaptive filter is used to achieve high speed and low latency design with less computational elements. Since the signal processing technique is developed for distant healthcare systems, white noise removal is mainly focused. DWT is applied to the pre-processed signal for HRV feature extraction, and machine-learning techniques are used for performing arrhythmic beat classification. In this paper, the SVM classifier and other popular classifiers have been used on noise removed feature extracted signal for beat classification. The results show that the SVM classifier performs better than additional machine learning-based classifiers [77].

Liu *et al*. (2019) proposed an ECG arrhythmia classification algorithm based on CNN. They compared the CNN models with combining linear discriminant analysis (LDA) and SVM. All cardiac arrhythmia beats are derived from the MIT-BIH Arrhythmia Database, which was classed into five groups according to the standard developed by the Association for the Advancement of Medical Instrumentation (AAMI). The training set and the testing set come from different people, and the correction of classification is >90% [78].

2.5. KNN

The KNN algorithm is a simple machine-learning algorithm compared to similar machine learning approaches [79]. Most of the machine-learning algorithms work on the KNN algorithm [80]. KNN classifier is an instance-based learning method, which stores all training sample vectors [81]. It is a very simple and effective method, especially for high-dimensional problems [82]. It classifies the new unknown test samples based on similar training samples [83]. The similarity measure is usually the Euclidean distance [84]. K-NN classifier was based on grouping of closest training points of data in the considered feature space. The majority of voters do the cluster to the nearest neighbor points [85].

Many approaches are published the ECG classification based on KNN. Below some of these new works:

Faziludeen and Sankaran (2016) presented a method for automatic ECG classification into two classes: Normal and PVC. The Evidential K-Nearest Neighbors (EKNN) was based on the Dempster Shafer Theory for classifying the ECG beats. RR interval features were used. The analysis was performed on the MIT-BIH database. The performance of EKNN was compared with the traditional KNN (maximum voting) approach. The effect of training data size was assessed using training sets of varying sizes. The EKNN based classification system was shown to out perform the KNN based classification system consistently [86].

Bouaziz *et al*. (2018) implemented an automatic ECG heartbeats classifier based on KNN. The segmentation of ECG signals has been performed by DWT. The considered categories of beats are normal (N), PVC, APC, RBBB, and LBBB. The validation of the presented KNN based classifier has been achieved using ECG data from MIT-BIH arrhythmia database. They have obtained the excellent classification performances, in terms of the calculated values of the SPE and the SEN of the classifier for several pathological heartbeats and the global classification rate, which is equal to 98, 71% [87].

Khatibi and Rabinezhadsadatmahaleh (2019), a novel feature engineering method, was proposed based on deep learning and K-NNs. The features extracted were classified with different classifiers such as DTs, SVMs with different kernels, and random forests. This method has good performance for beat classification and achieves the average ACC of 99.77%, AUC of 99.99%, precision of 99.75%, and recall of 99.30% using fivefold Cross-Validation strategy. The main advantage of the proposed method was its low computational time compared to training deep learning models from scratch and its high ACC compared to the traditional machine learning models. The strength and suitability of the proposed method for feature extraction are shown by the high balance between SEN and SPE [88].

3. DISCUSSION

The ECG classification, which shows the status of the heart and the cardiovascular condition, is essential to improve the patient's living quality. The main purpose of this work is to review the main techniques of ECG signal classification. In general, any structure of ECG classification can be divided into four stages. The first one is a preprocessing step, which is a crucial step in the ECG signal classification. For that reason, most well-known techniques are reviewed in this paper. The idea of using the preprocessing step and the combination of preprocessing techniques is to improve the performances of the model. The second step is extracting the most relevant information from the ECG signal, which represents the heart status. The step is called a feature extraction step. There is a vital challenge to extract efficient information that can be discriminated based on the variance status of the ECG signal. The success rate of the model can evaluate whether the feature contains valuable knowledge of the signal or not. The third step is named as the feature selection step. Time execution of the model is a crucial part and can be reduced using optimal features among the feature spaces. Many techniques have been adopted for reducing the dimensionality of the features. Some of the methods have been inspired by nature and the others, working based on the mathematical rules. The primarily focused step is selecting a machine-learning algorithm to classify the ECG features. Plenty of approaches has been used for this purpose. Most of the classifier methods are fed by the features, but CNN is supplied using the raw signal as CNN is a feature-less technique. ANN, CNN, DWT, KNN, and SVM are reviewed. All reviewed articles are downloaded from three trusted sources, IEEE, ScienceDirect, and Springer for 2015–2020. Tables 1-5 show the summarization of all the reviewed articles in term of what kind of machine-learning were used, how the methods were effective to the ECG

TABLE 1: Heartbeat methods classification based on ANN

Artificial neural networks					
Author (year)	**Dataset**	**Purpose**	**Methods**	**Result**	**Remarks**
Chen *et al*. (2016)	MIT-BIH arrhythmia dataset	Reduce the computing time by a simple method	Wavelet Artificial Neural Network (W-ANN)	The average computing time can be reduced by 49%	Use a mobile real-time applications to classify ECG
Boussee *et al*. (2016)	MIT-BIH arrhythmia dataset	Record, proceed, and classify ECG signal	Mel Frequency Coefficient Cepstrum (MFCC)+ANN	Available a robust and quick classification system	Build a system to classify ECG by a combination of signal processing algorithms
Savalia *et al*. (2017)	MIT/BIH Normal Sinus Database and MIT-BIH arrhythmia dataset	To distinguish normal and abnormal ECG	ANN	Accuracy=86%	Abnormal ECG is used to identify specific heart diseases
Wess *et al*. (2017)	MIT-BIH arrhythmia dataset	To present FPGA-based ECG arrhythmia detection	PCA+ANN	Accuracy=99.82%	Increased the number of inputs, hidden layer, and fixed point
Pandey *et al*. (2018)	UCI arrhythmia dataset	Early and right identification of cardiac disease	RNN, RBF and BPA	Accuracy RNN=83.05% RBF=75.25% BPA=74.35%	Accuracy of RNN is better than two ANN models
Sannino and Pietro (2018)	MIT-BIH arrhythmia dataset	The automatic recognition of abnormal	DNN	Accuracy=99.68%	The model is competitive in sensitivity and specificity
Debnath *et al*. (2019)	MIT-BIH arrhythmia dataset	Analyze and Predict heart abnormality	ANN	Accuracy Normal=97.46% Bradycardia=87.20% Tachycardia=99.97% Block=66.72%	Input noisy ECG signals
Abdalla *et al*. (2019)	MIT-BIH arrhythmia dataset	Distinguish between different types of ECG arrhythmia	CEEMDAN+ANN	Accuracy=99.9%	The performance of the CEEMDAN and ANN is better than all existing methods

TABLE 2: Heartbeat methods classification based on CNN

Convolutional neural network					
Author (year)	**Dataset**	**Purpose**	**Methods**	**Result**	**Remarks**
Zubair *et al*. (2016)	MIT-BIH arrhythmia database	Proposed learns features from raw ECG	1D-CNN	Accuracy=92.7%	The model avoids the need for hand-crafted features
Yin *et al*. (2016)	Data is built on ECG sensor chip BMD101 and Bluetooth module	Monitoring and classifying ECG signals and radar signals	Cascade CNN	Accuracy=88.89%	The system can achieve stable performance in the slight motion state
Oh *et al*. (2017)	MIT-BIH arrhythmia database	Identified automatically five different categories of ECG	9-layers CNN	Accuracy With noise=94.03% Without noise=93.47%	Generated synthetic data to overcome imbalance problems
Zhai and Tin (2018)	MIT-BIH arrhythmia database	Implemented model on portable device for long-term monitoring	CNN	Accuracy>97%	The model doesn't need manual feature extraction or expert assistant
Zhang *et al*. (2019)	Synthetic and real-world ECG datasets	Proposed learns features and classifiers from the time-frequency domain	DCNN	Accuracy=99%	The model can integrated into a portable ECG monitor with limited resources
Wang (2020)	MIT-BIH AF dataset	Proposed approach for automated AF detection	CNN+MENN	Accuracy=97.4%	The model has great potential to assist physicians and reduce mortality
Yao *et al*. (2020)	China Physiological Signal Challenge 2018 database	Classify varied-length ECG signals	Attention-based time-incremental (ATI)-CNN	Accuracy=81.2%	The model compares with VGGNet, increases the accuracy

TABLE 3: Heartbeat methods classification based on DWT

Discrete wavelet transforms					
Author (year)	**Dataset**	**Purpose**	**Methods**	**Result**	**Remarks**
Desai *et al*. (2015)	MIT-BIH arrhythmia dataset	Detected five classes of ECG arrhythmia	DWT+ICA+SVM	Accuracy =98.49%	efficient system in healthcare diagnosis
Saraswat *et al*. (2016)	MIT-BIH arrhythmia dataset	Presented a clear difference between normal and abnormal ECG	DWT	Provide min and max values of normal and abnormal ECG.	detecting changes in signals leading to smooth the feature
Alickovic and Subasi (2016)	MIT-BIH arrhythmia and St. -Petersburg Institute of Cardiological Technics Arrhythmia Database	Automated system for the classification of ECG	DWT+C4.5+CART	Accuracy C4.5=99.95% CART=99-80%	Efficient system for cardiac arrhythmia detection
Pan *et al*.(2017)	MIT-BIH arrhythmia dataset	Developed system for clinical arrhythmia classification	DWT+random forest	Accuracy=99.77%	The system improves classification accuracy and speed
Sahoo *et al*. (2017)	MIT-BIH arrhythmia dataset	Improved algorithm to detect QRS complex features to classify four types of ECG	Multiresolution WT +NN+SVM	Accuracy NN=96.67% SVM=98.39%	Extracted features are acceptable for classifying ECG by SVM
Ceylan (2018)	MIT-BIH arrhythmia dataset	system for signal compression, noise elimination, and classification	DWT+AdaBoost+ SVM+LDA+DL	Accuracy > 99%	The best classification accuracy was obtained by (DL-AdaBoost – SVM)
Tea and Vladan (2018)	MIT-BIH benchmark arrhythmia dataset	Monitor ECG in real –time	FFT+DCT+DWT +random forests	Accuracy=97.33%	DWT provides the best performance in comparison with FFT and DCT
Zhang *et al*. (2019)	MIT-BIH arrhythmia dataset	Diagnosis of lowecost wearable ECG device	DWT+RF+KNN+J48	Accuracy=98.08%	Reduce the computational cost and improves the classification efficiency.
Kora *et al*. (2019)	MIT-BIH arrhythmia dataset	Detect Atrial Fibrillation in the ECG signal	DWT+KNN+SVM	Accuracy DWT+SVM=94.07% DWT+KNN= 99.5%	DWT represent the essential characteristics of the ECG

TABLE 4: Heartbeat methods classification based on SVM

Support vector machines					
Author (year)	**Dataset**	**Purpose**	**Methods**	**Result**	**Remarks**
Elhaj *et al*.(2016)	MIT-BIH arrhythmia database	Classifying ECG signal with high accuracy	PCA+DWT+ICA+ HOS+NN+SVM-RBF	Accuracy SVM-RBF=98.91% NN=98.90%	Both classifiers provide equal average accuracy, sensitivity, and specificity
Arjunan (2016)	MIT-BIH arrhythmia database	Categorize ECG by an automated system	SVM	Accuracy=90%	Mean, variance, standard deviation, and skewness are used for feature extraction
Smíšek *et al*. (2017)	Hidden dataset of 2017 PhysioNet/ CinC Challenge	An advanced method for automatic classification ECG	SVM-RBF	F1-measure=0.81	Quite high performance was achieved even for low number in training set
Wu *et al*. (2017)	Collect data by sensors.	Recognize drunk driving by ECG and PPG	SVM	Accuracy=95%	The smart and wearable sensing devices offer right solution for drunk driving
Venkatesan *et al*. (2018)	MIT-BIH arrhythmia database.	Classifier with low computational complexity	SVM	Accuracy=96%	SVM is better than various classification techniques
Liu *et al*. (2019)	MIT-BIH arrhythmia database.	Robust and efficient model to achieve a real-time analysis ECG	SVM+CNN+LDA	Accuracy >90%	Sometimes, do not need to extract complex features of ECG

TABLE 5: Heartbeat methods classification based on KNN

K- nearest neighbors					
Author (year)	**Dataset**	**Purpose**	**Methods**	**Result**	**Remarks**
Faziludeen and Sankaran(2016)	MIT-BIH arrhythmia database	Classify ECG beat into two classes	KNN+EKNN	Lower error rates	Increase in training size is shown to lower the error rates
Bouaziz *et al*. (2018)	MIT-BIH arrhythmia database	Implement an automatic ECG heartbeats classifier	KNN	Accuracy=98.71%	KNN an important and significant tool for ECG recognition
Khatibi and Rabinezhadsadatmahaleh (2019)	MIT-BIH arrhythmia database	Classify ECG for arrhythmia detection	CNN+DT+S VM+RF+K-NNs	Accuracy=99.77%	The method is low computational time

classification and which kind of ECG datasets were used. Some important points in the ECG classification are observed and highlighted in the below:

According to the previous works based on the ANN algorithm for heartbeat classification, ANN is trained using the polyspectrum patterns and features extracted from the higher-order spectral analysis of normal and abnormal ECG signal. ANN is used as a classifier to help knowledge management and decision-making system to improve classification ACC. The result shows that ANN with PCA obtains lowest error rate to classify the ECG signal. The performance of the CEEMDAN and ANN is better than all higher than all existing and previous algorithms (Table 1). (The main point are extracted from ANN [Reviewer 1 and 2 and 3]).

CNN is straight forward to apply as the CNN is a features less techniques. Hence, the researcher does not concern about the feature that means any handcraft feature does not require in the CNN model. 1 D and 2 D of CNN have been adopted, According to the observed result, 1 D CNN outperformed of the 2D CNN. Moreover, the 1D CNN is less complex compare to the 2 D CNN in term of computational steps. CNN also can be integrated with MENN to improve the classification ACC (Table 2). (Roles CNN in ECG classification [Reviewer 1 and 2 and 3]).

DWT is applied on each heartbeat to obtain the morphological features. It provides better time and frequency resolution of ECG signal. DWT shows the powerful tool for ECG classification and it is straight forward tool to implantation. Moreover, DWT is an assisting the clinicians for making an accurate diagnosis of CVDs. Based on the summarization of some works on DWT, the integration DWT model with random forest can achieve 99.77% ACC (Table 3) (The main notes about DWT [Reviewer 1 and 2 and 3]).

SVM (SVM) is widely used for pattern recognition. SVM model with a weighted kernel function method significantly recognizes the Q wave, R wave, and S wave in the input ECG signal to categorize the heartbeat. SVM is also the powerful tool to ECG classification; however, the performance CNN has outperformed of the SVM. Moreover, the time consumption of implementing SVM is higher than KNN model and smaller than the CNN model. SVM-RBF classifier classifies 95% of the given ECG signal correctly with simple statistical features Table4. (The contributions of SVM [Reviewer 1 and 2 and 3]).

The lowest computational rate for diagnosing arrhythmia can be achieved by applying KNN as the KNN algorithm does not require the training stage. The role of the handcraft features is a vital subject to the KNN model as long as the dimensional of the obtained features is low because the KNN model works based on the distance. Time domain and frequency domain features are applied to KNN classifier for ECG classification which is simpler than other machine-learning approaches (Table 5). (The main roles of kNN in ECG classification [Reviewer 1 and 2 and 3]).

4. CONCLUSION

Classification of ECG signals is acting an important role in recognizing normal and abnormal heartbeat. Increasing the ACC of ECG classification is a challenging problem. It has been interested in more than a decade; for this reason, many approaches have been developed. In this paper, most recent approaches are reviewed in terms of some aspects such as method, dataset, contribution, and success rate. The table (CNN) summarizes variance approaches in ECG signal analysis. We suggest using a hybrid model based on CNN with long- and short-term memory (LSTM). The CNN part can extract the features from the raw signal which can be a temporal features based on how many convolution layers we will used, and LSTM can learn the pattern in the temporal feature as the LSTM is more suitable to time series features. Then, the model can predict unknown ECG signals. We will tune filters in the CNN model and layers in the LSTM model to increase the classification rate. (Explain how use CNN+LSTM [Reviewer 3]).

REFERENCES

[1] A. Alberdi, A. Aztiria and A. Basarab. “Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review”. *Journal of Biomedical Informatics*, vol. 59, pp. 49-75, 2016.

[2] M. S. Al-Ani. “Electrocardiogram waveform classification based on P-QRS-T wave recognition”. *UHD Journal of Science and Technology*, vol. 2, no. 2, pp. 7-14, 2018.

[3] M. Al-Ani. “A rule-based expert system for automated ecg diagnosis”. *International Journal of Advances in Engineering and Technology*, vol. 6, no. 4, 1480-1492, 2014.

[4] M. S. Al-Ani and A. A. Rawi. “ECG Beat diagnosis approach for ECG printout based on expert system”. *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 4, pp. 797-807, 2013.

[5] S. H. Jambukia, V. K. Dabhi and H. B. Prajapati. “Classification of ECG Signals Using Machine Learning Techniques: A Survey”. In: *Conference Proceeding 2015 International Conference on Advances in Computer Engineering and Applications*, pp. 714-721, 2015.

[6] J. Li, Y. Si, T. Xu and S. Jiang. “Deep convolutional neural network based ecg classification system using information fusion and one-hot encoding techniques”. *Mathematical Problems Engineering*, vol. 2018, p. 7354081, 2018.

[7] D. Sung, J. Kim, M. Koh and K. Park. “ECG Authentication in post-exercise situation ECG authentication in post-exercise situation”. *Conference Proceeding IEEE Engineering Medical Biology Socirty*, vol. 1, pp. 446-449, 2017.

[8] M. Lakshmi, D. Prasad and D. Prakash. “Survey on EEG signal processing methods”. *International Journal of Advanced Research in Computer Science*, vol. 4, no. 1, pp. 84-91, 2014.

[9] R. Chaturvedi and Y. Yadav. “A survey on compression techniques”. *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 9, pp. 3511-3513, 2013.

[10] H. Y. Lin, S. Y. Liang, Y. L. Ho, Y. H. Lin and H. P. Ma. “Discrete-wavelet-transform-based noise removal and feature extraction for ECG signals”. *IRBM*, vol. 35, no. 6, pp. 351-361, 2014.

[11] M. Hammad, S. Zhang and K. Wang. “A novel two-dimensional ECG feature extraction and classification algorithm based on convolution neural network for human authentication”. *Future Generation Computer Systems*, vol. 101, pp. 180-196, 2019.

[12] J. Juang. “Proceedings of the 3rd International Conference on Intelligent Technologies and Engineering Systems (ICITES2014)”. Vol. 345. In: *Lecture Notes in Electrical Engineering,* pp. 545-555, 2016.

[13] A. Giorgio, M. Rizzi and C. Guaragnella. “Efficient detection of ventricular late potentials on ECG signals based on wavelet denoising and SVM classification”. *Information*, vol. 10, no. 11, p. 328, 2019.

[14] F. A. R. Sánchez and J. A. G. Cervera. “ECG classification using artificial neural networks”. *Journal of Physics: Conference Series*, vol. 1221, no. 1, pp. 1-6, 2019.

[15] S. V. Deshmukh and O. Dehzangi. “ECG-Based Driver Distraction Identification Using Wavelet Packet Transform and Discriminative Kernel-Based Features”. *2017 IEEE International Conference on Smart Computing*, 2017.

[16] S. K. Berkaya, A. K. Uysal, E. S. Gunal, S. Ergin, S. Gunal and M. B. Gulmezoglu. “A survey on ECG analysis”. *Biomedical Signal Processing and Control*, vol. 43, pp. 216-235, 2018.

[17] N. A. Polytechnic. “Automated identification of shockable and non-shockable life-threatening ventricular arrhythmias using convolutional neural network”. *Future Generation Computer Systems the International Journal of Escience*, vol. 79, p. 952, 2017.

[18] F. A. Elhaj, N. Salim, T. Ahmed, A. R. Harris and T. T. Swee. “Hybrid Classification of Bayesian and Extreme Learning Machine for Heartbeat Classification of Arrhythmia Detection”. In: *6th ICT*

International Student Project Conference, pp. 1-4, 2017.

[19] P. Li, K. L. Chan, S. Fu and S. M. Krishnan. "An abnormal ECG beat detection approach for long-term monitoring of heart patients based on hybrid kernel machine ensemble". *Lecture Notes in Computer Science*, Vol. 354. 1Springer, Berlin, pp. 346-355, 2005.

[20] S. Shadmand and B. Mashoufi. "A new personalized ECG signal classification algorithm using block-based neural network and particle swarm optimization". *Biomedical Signal Processing and Control*, vol. 25, pp. 12-23, 2016.

[21] J. Mateo, A. M. Torres, A. Aparicio and J. L. Santos. "An efficient method for ECG beat classification and correction of ectopic beats". *Computers and Electrical Engineering*, vol. 53, pp. 219-229, 2016.

[22] C. Kamalakannan, L. P. Suresh, S. S. Dash and B. K. Panigrahi. "*Power Electronics and Renewable Energy Systems: Proceedings of ICPERES 2014*. Vol. 326. Lecture Notes in Electrical Engineering, pp. 1537-1544, 2014.

[23] C. K. and B. S. "A survey on various machine learning approaches for ECG analysis". *International Journal of Computer Applications*, vol. 163, no. 9, pp. 25-33, 2017.

[24] S. Z. Islam, S. Z. Islam, R. Jidin and M. A. M. Ali. "Performance Study of Adaptive Filtering Algorithms for Noise Cancellation of ECG Signal". Vol. 4. In: *ICICS 2009 Conference Proceeding 7th International Conference Information, Communication Signal Process*, 2009.

[25] M. A. Rahman, M. M. Milu, A. Anjum, A. B. Siddik, M. H. Sifat, M. R. Chowdhury, F. Khanam, M. Ahmad. "A statistical designing approach to MATLAB based functions for the ECG signal preprocessing". *The Iran Journal of Computer Science*, vol. 2, no. 3, pp. 167-178, 2019.

[26] M. T. Almalchy, V. Ciobanu and N. Popescu. "Noise removal from ECG signal based on filtering techniques". *Proceeding 2019 22nd International Conference Control Systems Computer Science*, pp. 176-181, 2019.

[27] G. H. Choi, E. S. Bak and S. B. Pan. "User identification system using 2D resized spectrogram features of ECG". *IEEE Access*, vol. 7, pp. 34862-34873, 2019.

[28] A. Lay-Ekuakille, M. A. Ugwiri, C. Liguori and P. K. Mvemba. "Enhanced methods for extracting characteristic features from ECG". *IEEE International Symposium on Medical Measurements and Applications*, pp. 1-5, 2019.

[29] J. Oster, J. Behar, O. Sayadi, S. Nemati, A. E. W. Johnson and G. D. Clifford. "Semisupervised ECG ventricular beat classification with novelty detection based on switching kalman filters". *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 9, pp. 2125-2134, 2015.

[30] S. Karpagachelvi. "ECG feature extraction techniques a survey approach". *International Journal of Computer Science and Information Security*, vol. 8, no. 1, pp. 76-80, 2010.

[31] S. Nasehi and H. Pourghassem. "Seizure detection algorithms based on analysis of EEG and ECG signals: A survey". *Neurophysiology*, vol. 44, no. 2, pp. 174-186, 2012.

[32] S. M. J. Jalali, M. Karimi, A. Khosravi and S. Nahavandi. "An efficient neuroevolution approach for heart disease detection". *Conference Proceeding IEEE International Conference System Man Cybernetics*, pp. 3771-3776, 2019.

[33] D. Carrera, B. Rossi, P. Fragneto and G. Boracchi. "Online anomaly detection for long-term ECG monitoring using wearable devices". *Pattern Recognition*, vol. 88, pp. 482-492, 2019.

[34] E. K. Wang, X. Zhang and L. Pan. "Automatic classification of CAD ECG signals with SDAE and bidirectional long short-term network". *IEEE Access*, vol. 7, pp. 182873-182880, 2019.

[35] N. Omer, Y. Granot, M. Kähönen, R. Lehtinen, T. Nieminen, K. Nikus. "Blinded analysis of an exercise ECG database using high frequency QRS analysis". Vol. 44. In: *2017 Computing in Cardiology*, pp. 1-4, 2017.

[36] I. Karagoz. "Cmbebih 2019". Vol. 73. In: *IFMBE Proceeding C*, pp. 159-163, 2019.

[37] S. Lata and R. Kumar. "Disease classification using ECG signals based on R-peak analysis with ABC and ANN". *The International Journal of Electronics, Communications, and Measurement Engineering*, vol. 8, no. 2, pp. 67-86, 2019.

[38] A. Delrieu, M. Hoël, C. T. Phua and G. Lissorgues. "Multi physiological signs model to enhance accuracy of ECG peaks detection". *IFMBE Proceeding*, vol. 61, pp. 58-61, 2017.

[39] K. C. J. Chen, Y. S. Ni and J. Y. Wang. "Electrocardiogram Diagnosis Using Wavelet-Based Artificial Neural Network". In: *2016 IEEE 5th Globel Conference Consumer Electronics GCCE 2016*, pp. 5-6, 2016.

[40] M. Boussaa, I. Atouf, M. Atibi and A. Bennis. "ECG Signals Classification Using MFCC Coefficients and ANN Classifier". *Proceeding 2016 International Conference Electronics Information Technology*, pp. 480-484, 2016.

[41] S. Savalia, E. Acosta and V. Emamian. "Classification of cardiovascular disease using feature extraction and artificial neural networks". *Journal of Biosciences and Medicines*, vol. 5, no. 11, pp. 64-79, 2017.

[42] M. Wess, P. D. S. Manoj and A. Jantsch. "Neural Network Based ECG Anomaly Detection on FPGA and Trade-off Analysis. In: *Proceedings IEEE International Symposium on Circuits and Systems*, 2017.

[43] S. Pandey and R. R. Janghel. "Classification of ECG arrhythmia using recurrent neural networks ECG arrhythmia classification using artificial neural networks". *Procedia Computer Science*, vol. 8, pp. 1290-1297, 2018.

[44] G. Sannino and G. De Pietro. "A deep learning approach for ECG-based heartbeat classification for arrhythmia detection". *Future Generation Computer Systems*, vol. 86, pp. 446-455, 2018.

[45] T. Debnath, M. Hasan and T. Biswas. "Analysis of ECG Signal and Classification of Heart Abnormalities Using Artificial Neural Network". In: *Proceeding 9th International Conference Electrical and Computer Engineering*, pp. 353-356, 2017.

[46] F. Y. O. Abdalla, L. Wu, H. Ullah, G. Ren, A. Noor and Y. Zhao. "ECG arrhythmia classification using artificial intelligence and nonlinear and nonstationary decomposition". *Signal, Image Video Process*, vol. 13, no. 7, pp. 1283-1291, 2019.

[47] Z. K. Abdul. "Kurdish speaker identification based on one dimensional convolu- tional neural network". *Computational Methods for Differential Equations*, vol. 7, no. 4, pp. 566-572, 2019.

[48] D. Li, J. Zhang, Q. Zhang and X. Wei. "Classification of ECG signals based on 1D convolution neural network". In: *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services, Healthcom*, pp. 1-6, 2017.

[49] M. Zubair, J. Kim and C. Yoon. "An automated ECG beat classification system using convolutional neural networks". In: *2016 6th International Conference on IT Convergence and Security*, 2016.

[50] W. Yin, X. Yang, L. Zhang and E. Oki. "ECG monitoring system integrated with IR-UWB radar based on CNN". *IEEE Access*, vol.

4, pp. 6344-6351, 2016.

[51] S. L. Oh, N. A. Polytechnic, N. A. Polytechnic, Y. Hagiwara and J. H. Tan. “A deep convolutional neural network model to classify heartbeats”. *Computers in Biology and Medicine*, vol. 89, pp. 389-396, 2017.

[52] X. Zhai and C. Tin. “Automated ECG classification using dual heartbeat coupling based on convolutional neural network”. *IEEE Access*, vol. 6, pp. 27465-27472, 2018.

[53] J. Zhang, J. Tian, Y. Cao, Y. Yang and X. Xu. “Deep time frequency representation and progressive decision fusion for ECG classification”. *Knowledge-Based Systems*, vol. 190, p. 105402, 2020.

[54] J. Wang. “A deep learning approach for atrial fibrillation signals classification based on convolutional and modified Elman neural network”. *Future Generation Computer Systems*, vol. 102, pp. 670-679, 2020.

[55] Q. Yao, R. Wang, X. Fan, J. Liu and Y. Li. “Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network”. *Information Fusion*, vol. 53, no. 1, pp. 174-182, 2020.

[56] H. Limaye and V. V. Deshmukh. “ECG noise sources and various noise removal techniques: A survey”. *International Journal of Application or Innovation in Engineering and Management*, vol. 5, no. 2, pp. 86-92, 2016.

[57] S. L. Joshi. “*A Survey on ECG Signal Denoising Techniques 2013 International Conference on Communication Systems and Network Technologies A Survey on ECG Signal DenoisingTechniques*”, 2013.

[58] H. El-Saadawy, M. Tantawi, H. A. Shedeed and M. F. Tolba. “Electrocardiogram (ECG) classification based on dynamic beats segmentation”. *The ACM International Conference Proceeding Series*, pp. 75-80, 2016.

[59] T. R. Naveen, K. V. Reddy, A. Ranjan and S. Baskaran. “Detection of abnormal ECG signal using DWT feature extraction and CNN”. *International Research Journal of Engineering and Technology*, vol. 6, no. 3, pp. 5175-5180, 2019.

[60] U. Desai, R. J. Martis, C. G. Nayak, K. Sarika and G. Seshikala. “Machine Intelligent Diagnosis of ECG for Arrhythmia Classification Using DWT, ICA and SVM Techniques. *12th IEEE International Conference Electronic Energy, Environmental Research Communications*, pp. 2-5, 2016.

[61] S. Saraswat, G. Srivastava and S. Shukla. “Decomposition of ECG Signals using Discrete Wavelet Transform for Wolff Parkinson White Syndrome Patients”. In: *Proceedings 2016 International Conference on Micro-Electronics and Telecommunication Engineering*, pp. 361-365, 2016.

[62] E. Alickovic and A. Subasi. “Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier”. *The Journal of Medical Systems*, vol. 40, no. 4, pp. 1-12, 2016.

[63] G. Pan, Z. Xin, S. Shi and D. Jin. “Arrhythmia classification based on wavelet transformation and random forests”. *Multimedia Tools and Applications Journal*, vol. 77, no. 17, pp. 21905-21922, 2018.

[64] S. Sahoo, B. Kanungo, S. Behera and S. Sabut. “Multiresolution wavelet transform based feature extraction and ECG classification to detect cardiac abnormalities Multiresolution wavelet transform based feature extraction and ECG classification to detect cardiac abnormalities”. *Measurement*, vol. 17, no. 1, pp. 55-66, 2017.

[65] M. Barstuğan and R. Ceylan. “The effect of dictionary learning on weight update of AdaBoost and ECG classification”. *Journal of King Saud University*, vol. 30, pp.1-9, 2018.

[66] T. Marasović and V. Papić. “A comparative study of FFT, DCT, and DWT for efficient arrhytmia classification in RP-RF framework”. *International Journal of E-Health and Medical Communications*, vol. 9, no. 1, pp. 35–49, 2018.

[67] Y. Zhang, Y. Zhang, B. Lo and W. Xu. “Wearable ECG signal processing for automated cardiac arrhythmia classification using CFASE-based feature selection”. *Expert System*, vol. 37, no. 1, pp. 1-13, 2020.

[68] P. Kora, C. U. Kumari, K. Swaraja and K. Meenakshi. “Atrial Fibrillation detection using Discrete Wavelet Transform. In: *Proceedings of 2019 3rd IEEE International Conference on Electrical, Computer and Communication Technologies*, pp. 1-3, 2019.

[69] S. Raj and K. C. Ray. “ECG signal analysis using DCT-Based DOST and PSO Optimized SVM”. *IEEE Transactions on Automatic Control*, vol. 66, no. 3, pp. 470-478, 2017.

[70] R. Banerjee, A. Ghose and S. Khandelwal. “A Novel Recurrent Neural Network Architecture for Classification of Atrial Fibrillation Using Single-lead ECG. In: *European Signal Processing Conference*, pp. 1-5, 2019.

[71] H. Khorrami and M. Moavenian. “A comparative study of DWT, CWT and DCT transformations in ECG arrhythmias classification”. *Expert Systems With Applications*, vol. 37, no. 8, pp. 5751-5757, 2010.

[72] V. Mygdalis, A. Tefas and I. Pitas. “Exploiting multiplex data relationships in support vector machines”. *Pattern Recognition*, vol. 85, pp. 70-77, 2019.

[73] F. A. Elhaj, N. Salim, A. R. Harris, T. T. Swee and T. Ahmed. “Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals”. *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 52-63, 2016.

[74] V. R. Arjunan. “ECG signal classification based on statistical features with SVM classification”. *International Journal of Advances in Signal and Image Sciences*, vol. 2, no. 1, p. 5, 2016.

[75] R. Smíšek, J. Hejč, M. Ronzhina, A. Němcová, L. Maršánová, J. Chmelík, K. Jana. SVM Based ECG classification using rhythm and morphology features, cluster analysis and multilevel noise estimation”. *Computing in Cardiolology*, vol. 44, pp. 1-4, 2017.

[76] W. F. Wang, C. Y. Yang and Y. F. Wu. “SVM-based classification method to identify alcohol consumption using ECG and PPG monitoring”. *Personal and Ubiquitous Computing*, vol. 22, no. 2, pp. 275-287, 2018.

[77] C. Venkatesan, P. Karthigaikumar, A. Paul, S. Satheeskumaran and R. Kumar. “ECG signal preprocessing and SVM classifier-based abnormality detection in remote healthcare applications”. *IEEE Access*, vol. 6, pp. 9767-9773, 2018.

[78] J. Liu, S. Song, G. Sun and Y. Fu. “Classification of ECG arrhythmia Using CNN, SVM and LDA”. Vol. 11633. In: *International Conference on Artificial Intelligence and Security*, pp. 191-201, 2019.

[79] J. Zhai and A. Barreto. “Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables”. In: *Proceedings of the 28th IEEE EMBS Annual International Conference*, pp. 1355-1358, 2007.

[80] V. Gupta and M. Mittal. “KNN and PCA classifier with Autoregressive modelling during different ECG signal interpretation”. *Procedia Computer Science*, vol. 125, pp. 18-24, 2018.

[81] N. Flores, R. L. Avitia, M. A. Reyna and C. García. “Readily available ECG databases”. *Journal of Electrocardiology*, vol. 51,

no. 6, pp. 1095-1097, 2018.

[82] R. P. Narwaria, S. Verma and P. K. Singhal. “Removal of baseline wander and power line interference from ECG signal a survey approach”. *International Journal of Information and Electronics Engineering*, vol. 3, no. 1, pp. 107-111, 2011.

[83] N. K. Dewangan and S. P. Shukla. “A survey on ECG signal feature extraction and analysis techniques”. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, vol. 3, no. 6, pp. 12-19, 2015.

[84] I. Saini. “Analysis ECG data compression techniques a survey approach”. *The International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 2, pp. 544-548, 2013.

[85] M. M. Baig, H. Gholamhosseini and M. J. Connolly. “A comprehensive survey of wearable and wireless ECG monitoring systems for older adults”. *Medical and Biological Engineering and Computing*, vol. 51, no. 5, pp. 485-495, 2013.

[86] S. Faziludeen and P. Sankaran. “ECG beat classification using evidential K-nearest neighbours”. *Procedia Computer Science*, vol. 89, pp. 499-505, 2016.

[87] F. Bouaziz, D. Boutana and H. Oulhadj. “Diagnostic of ECG Arrhythmia Using Wavelet Analysis and K-Nearest Neighbor Algorithm”. In: *Proceedings of the 2018 International Conference on Applied Smart Systems*, pp. 1-6, 2019.

[88] T. Khatibi and N. Rabinezhadsadatmahaleh. “Proposing feature engineering method based on deep learning and K-NNs for ECG beat classification and arrhythmia detection”. *Physical and Engineering Sciences in Medicine*, vol. 43, pp. 1-20, 2019.

p-ISSN 2521-4209
e-ISSN 2521-4217

زانكۆی گەشەپێدانی مرۆیی

اقرأ

جامعة التنمية البشرية
UNIVERSITY OF HUMAN DEVELOPMENT

UHD Journal of Science and Technology

A Scientific periodical issued by University of Human Developement

Vol.4 No.(1) June 2020

2020 2720

e.mail:jst@uhd.edu.iq