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.5 No.(1) June 2021

2021 2721

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

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: +964 770 247 3391

Email: jst@uhd.edu.iq

Contents

No.	Author Name	Title	Pages
1	Shalaw Faraj Salih Alan Anwer Abdulla	An Improved Content Based Image Retrieval Technique by Exploiting Bi-layer Concept	1-12
2	Haveen Muhammed Rashid	Modeling Groundwater Potential Zones across Sulaimani Governorate Using Geographic Information System and Multi-influencing Factor Techniques	13-20
3	Twana Latif Mohammed Ahmed Abdullah Ahmed	Offline Writer Recognition for Kurdish Handwritten Text Document Based on Proposed Codebook	21-27
4	Fawzi Abdul Azeez Salih Alan Anwer Abdulla	An Efficient Two-layer based Technique for Content-based Image Retrieval	28-40
5	Ari Mohammed ali Ahmed Aree Ali Mohammed	A State-of-the-Art Review on Machine Learning-based Methods for Prostate Cancer Diagnosis	41-47
6	Kani Namiq Gharib Nawbahar Faraj Mustafa Haveen Muhammed Rashid	Urban Rainwater Harvesting Assessment in Sulaimani Heights District, Sulaimani City, KRG, Iraq	48-55
7	Tofiq Ahmed Tofiq Jamal Ali Hussein	Kurdish Text Segmentation using Projection-Based Approaches	56-65
8	Brzu T. Muhammed Ardalan Husin Awlla Sherko H. Murad Sabah N. Ahmad	Prediction of CoVid-19 mortality in Iraq-Kurdistan by using Machine learning	66-70
9	Sawza Saadi Saeed, Raghad Zuhair Yousif	A Slantlet based Statistical Features Extraction for Classification of Normal, Arrhythmia, and Congestive Heart Failure in Electrocardiogram	71-81

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

An Improved Content Based Image Retrieval Technique by Exploiting Bi-layer Concept

Shalaw Faraj Salih[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

Applications for retrieving similar images from a large collection of images have increased significantly in various fields with the rapid advancement of digital communication technologies and exponential evolution in the usage of the Internet. Content-based image retrieval (CBIR) is a technique to find similar images on the basis of extracting the visual features such as color, texture, and/or shape from the images themselves. During the retrieval process, features and descriptors of the query image are compared to those of the images in the database to rank each indexed image accordingly to its distance to the query image. This paper has developed a new CBIR technique which entails two layers, called bi-layers. In the first layer, all images in the database are compared to the query image based on the bag of features (BoF) technique, and hence, the M most similar images to the query image are retrieved. In the second layer, the M images obtained from the first layer are compared to the query image based on the color, texture, and shape features to retrieve the N number of the most similar images to the query image. The proposed technique has been evaluated using a well-known dataset of images called Corel-1K. The obtained results revealed the impact of exploring the idea of bi-layers in improving the precision rate in comparison to the current state-of-the-art techniques in which achieved precision rate of 82.27% and 76.13% for top-10 and top-20, respectively.

Index Terms: BoF, CBIR, Gabor, HSV, Zernike

1. INTRODUCTION

Digital image processing plays a significant role in various areas such as medical image processing [1], image inpainting [2], pattern recognition, biometrics, content-based image retrieval (CBIR), image compression, information hiding [3], and multimedia security [4]. The retrieval of similar images from a large range of images is becoming a serious challenge with the advent of digital communication technology and the growing use of the Internet. Several penetrating and retrieval utilities are essential for end users to retrieve the images efficiently from different domains of the image databases such as medical, education, weather forecasting, criminal investigation, advertising, social media, web, art design, and entertainment. The query information is either text format or image format.

Different techniques for image retrieval have been developed and they are classified into two approaches: Text-based image retrieval (TBIR) and CBIR [5]. TBIR was first introduced in 1970 for searching and retrieving images from image databases [6]. In TBIR, the images are denoted by text, and then the text is used to retrieve or search the images. Such a system is text-based search and is generally referred to as TBIR. The TBIR method relies on the manual text search or keyword matching of the existing image keywords and the result has been dependent on the human labeling of the

Access this article online

DOI: 10.21928/uhdjst.v5n1y2021.pp1-12

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Shalaw Faraj Salih, Department of Information Technology, Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, Iraq. E-mail: shalaw.faraj.s@spu.edu.iq

Received: 01-11-2020 Accepted: 23-12-2020 Published: 05-01-2021

images. TBIR approach requires information such as image keyword, image location, image tags, image name, and other information related to the image. It needs human intervention to enter the data of images in the database and that is the difficulty of the process. TBIR has the following limitations: (1) It leads to inaccurate results when human has been doing datasets annotation process wrongly, (2) single keyword of image information is not efficient to transfer the overall image description, (3) it is based on manual annotation of the images, which is time consuming [5], [7].

To overcome those mentioned limitations of TBIR, a new approach for image retrieval has been invented by researcher which is known as CBIR. CBIR can be considered as a common tool for retrieving, searching, and browsing images of a query information from a large database of digital images. In CBIR, the image information, visual features such as low level features (color, texture, and/or shape), or bag of features (BoF) have been extracted from the images to find similar images in the database [8]. Fig. 1 shows the general block diagram of CBIR approach [7].

In general, CBIR entails two main steps: The feature extraction and feature matching. In the first step, features are extracted from a dataset of images and stored in a feature vector. In the second step, the extracted features from the query image are compared with the extracted features of images in the dataset using certain distance measurement. If the distance between feature vector of the query image and the image in the database is small enough, the corresponding image in the database is considered as a match/similar image to the query image. Consequently, the matched images are then ranked accordingly to a similarity index from the smallest distance value to the largest one. Finally, the retrieved images are specified according to the highest similarity, that is, lowest distance value [9].

The main objective of CBIR techniques is to improve the efficiency of the system by increasing the performance using the combination of features [6]. Image features can be classified into two types: Local features and global features. Local features work locally which are focused on the key point in images whereas global features extract information from the entire image [10]. When the image dataset is quite large, image relevant to the query image are very few. Therefore, it is important to eliminate those irrelevant images. The main contribution of our proposed approach is filtering the images in the dataset to eliminate/minimize the most irrelevant images, then from the remaining images find the most similar/match images. In this paper, a new CBIR approach based on two layers is developed. The first layer aims in filtering the images using (BoF) strategy on the basis of extracting local features, while the second layer aims to retrieve similar images, from the remaining images, to the query image based on extracting global features such as color, shape, and texture.

The rest of the paper is organized as follows: Section 2 presents the literature review, Section 3 gives the background, Section 4 addresses the proposed approach in detail, Section 5 illustrates the experimental results, and finally, Section 6 presents the conclusion.

2. LITERATURE REVIEW

There are several CBIR techniques proposed for image retrieval applications using various feature extraction methods. Each of these techniques competes to improve the precision rate of finding the best similar images to the query image. In general, all the CBIR techniques have two main steps; the first step is feature extraction and the second step is feature matching. This section concerns the review of the most related and important existing CBIR techniques.

The concept of CBIR was first introduced by Kato in 1992 by developing a technique for sketch retrieval, similarity retrieval, and sense retrieval to support visual interaction [11]. Sketch

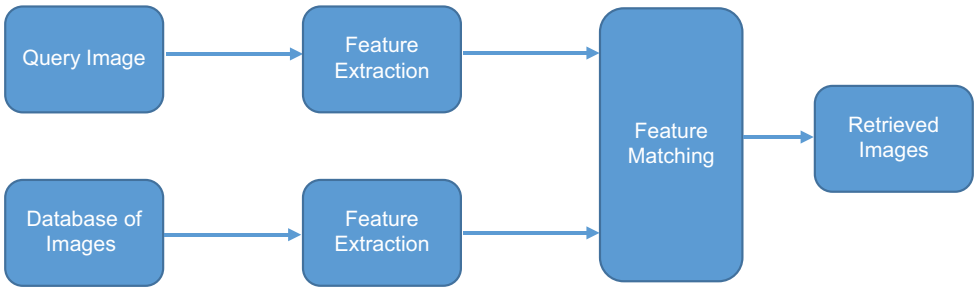

Fig. 1. General block diagram of content-based image retrieval approach.

retrieval accepts the image data of sketches, similarity retrieval evaluates the similarity based on the personal view of each user, and sense retrieval evaluates based on the text data and the image data at content level based on the personal view. In 2009, Lin *et al.* proposed a CBIR technique depending on extracting three types of image features [12]. The first feature, color co-occurrence matrix was extracted as a color feature, while for the second feature, difference between pixels of scan pattern was used for extracting texture feature, and the third feature, color histogram for K-mean was extracted which is based on color distribution. Consequently, feature selection techniques were implemented to select the optimal features not only to maximize the detection rate but also to simplify the computation of image retrieval. In addition, this proposed technique further uses sequential for-ward selection to select features with better discriminability for image retrieval and to overcome the problem of excessive features. Finally, Euclidean distance was used to find the similarity in the feature matching step. The results reported in this work claimed that the proposed technique reached a precision rate of 72.70% for the top-20.

Huang *et al.* proposed a new CBIR technique, in 2010, in which combined/fused the Gabor texture feature and Hue Saturation Value (HSV) color moment feature [13]. Furthermore, the normalized Euclidean distance was used to calculate the similarity between the feature vector of the query image and the feature vector of the images in the dataset. This proposed technique achieved the precision rate of 63.6% for the top-15. In 2012, Singha *et al.* proposed an algorithm for CBIR by extracting features called wavelet based color histogram image retrieval as a color and texture features [14]. The color and texture features are extracted through color histogram as well as wavelet transformation, for the combination of these features is robust to object translation and scaling in an image. This technique was used the histogram intersection distance for feature matching purposes. The results reported in this work claimed that this technique achieved a precision rate of 76.2% for the top-10. Another CBIR technique was proposed by Yu *et al.*, in 2013, that aims to investigate various combinations of mid-level features to build an effective image retrieval system based on the BoF model [15]. Specifically, this work studies two ways of integrating: 1- scale-invariant feature transform (SIFT) with local binary pattern (LBP) descriptors and, 2- histogram of oriented gradients with LBP descriptors. Based on the qualitative and quantitative evaluations on two benchmark datasets, the integrations of these features yield complementary and substantial improvement on image retrieval even with noisy background and ambiguous objects. Consequently, two integration models are proposed, the patch-based integration and the image-based integration. Using a weighted K-means clustering algorithm, the image-based SIFT-LBP integration achieved a precision rate of 65% for the top-20. A new CBIR technique was proposed by Somnugpong *et al.*, in 2016, by combining color correlograms and edge direction histogram (EDH) features to give precedence for spatial information in an image [16]. Color correlogram treats information about spatial color correlation, while EDH provides the geometry information in the case of the same image but different color. Evaluation is performed by simple calculation like Euclidean distance between the query image and the images in the database. Researchers claimed that their proposed technique achieved 65% of precision rate for the top-15. In 2018, Al-Jubouri *et al.* proposed a new CBIR technique that addresses the semantic gap issue by exploiting cluster shapes [17]. The technique first extracts local color using YCbCr color space and texture feature using Discrete Cosine Transform coefficients. The Expectation-Maximization Gaussian Mixture Model clustering algorithm is then applied to the local feature vectors to obtain clusters of different shapes. To compare dissimilarity between two images, the technique uses a dissimilarity measure based on the principle of Kullback-Leibler divergence to compare pair-wise dissimilarity of cluster shapes. This work further investigates two respective scenarios when the number of clusters is fixed and adaptively determined according to cluster quality. The results reported in this work illustrate that the proposed retrieval mechanism based on cluster shapes increases the image discrimination, and when the number of cluster is fixed to a large number, the precision of image retrieval is better than that when the relatively small number of clusters is adaptively determined. Authors claimed that their technique achieved a precision rate of 75% for the top-10.

In 2018, Nazir *et al.* proposed a new CBIR technique in which used color and texture features [18]. The edge histogram descriptor is extracted as a local feature and discrete wavelet transform as well as color histogram features are extracted as global features. Consequently, Manhattan distance measurement was used to measure the similarity between the feature vector of the query image and the feature vector of the images in the dataset. The reported results of the work revealed that this proposed technique achieved a precision rate of 73.5% for the top-20. Pradhan *et al.*, in 2019, developed a new CBIR scheme based on multi-level colored directional motif histogram [7]. The proposed scheme extracts local structural features at three different levels. The performance of this proposed scheme has been

evaluated using different Corel/natural, object, texture, and heterogeneous image datasets. Regarding to the Corel-1k, the precision rate of 64% and 59.6% was obtained for top-10 and top-20, respectively. Qazanfari *et al.*, in 2019, proposed a CBIR technique based on HSV color space [19]. The human visual system is very sensitive to the color and edge orientation, also color histogram and color difference histogram are two kinds of low-level feature extraction which are meaningful representatives of the image color and edge orientation information. This proposed technique was used Canberra distance measurement and this work has been evaluated using three standard databases Corel 5k, Corel 10 and UKBench and achieved 61.82%, 50,67%, and 74.77% of precision rate for the top-12, respectively. In 2019, Rashno *et al.*, proposed a new technique for CBIR in which color and texture features were used. HSV, Red, green, and blue (RGB) and norm of low frequency components were used as color features, while wavelet transformation was used to extract texture features [20]. Consequently, ant colony optimization-based feature selection was used to select the most relevant features, to minimize the number of features, and to maximize F-measure in the proposed CBIR system. Furthermore, Euclidean distance measurement was used to find the similarity between query and database images. The results reported in this work demonstrate that this approach reached the precision rate of 60.79% for the top-20. In 2019, Rana *et al.* proposed a CBIR technique by fusing parametric color and shape features with nonparametric texture feature [21]. The color moments and moment invariants which are parametric feature are extracted to describe color distribution and shapes of an image. The non-parametric ranklet transformation is performed to narrate the texture features. These parametric and non-parametric features were integrated to propose a robust and effective CBIR algorithm. In this proposed work, four similarity measurements are investigated during the experiment, namely, Chi-squared, Manhattan distance or City block distance, Euclidean distance, and Canberra distance. The experimental results demonstrate that Euclidean distance metric yields better precision and recall than other distance measuring criteria. Authors claimed that their technique achieved a precision rate of 67.6% for the top-15 using Euclidean distance. Finally, Sadique *et al.*, in 2019, proposed a CBIR technique in which investigates various global and local feature extraction methods for image retrieval [22]. The proposed work uses a combination of speeded up robust features (SURF) detector and descriptor with color moments as local features, and modified grey level cooccurrence matrices as global features. Both global and local features are used as the only local features are not suitable when the variety of images is large. Finally, fast approximate nearest neighbor search was used for matching the extracted features. Authors claimed that their proposed technique achieved a precision rate of 70.48% for the top-20.

3. BACKGROUND

This section aims to present a reasonable amount of background information about useful techniques such as (SURF) feature descriptor, color-based features, texture-based features, shape-based features, and feature matching techniques.

3.1. SURF Feature Descriptor

The most popular feature descriptor is SURF, which is also the most important one. However, there are other available feature descriptors. SURF can be considered as a local feature. Local features can provide more detailed characters in an image in comparison with global features such as color, texture, and shape. It is a rotation and scale invariant descriptor that performs better with respect to distinctiveness, repeatability, and robustness. It is also photometric deformations, detection errors, geometric, and robust to noise [23]. SURF is used in many applications such as BoF which is used and successful in image analysis and classification [24]. In BoF technique, SURF descriptor is often used to extract local feature first. In the next stage, a quantization algorithm such as K-means is separately applied to the extracted SURF features to reduce high dimensional feature vectors to clusters, which are also known as visual words. Then, K-means clustering is used to initialize the M center point to build M visual words. The K-means clustering algorithm takes feature space as input and reduces it into M cluster as output. Then, the image is represented as a histogram of code word occurrences by mapping the local features to a vocabulary [24]. The methodology of the image representation based on the BoF model is illustrated in Fig. 2.

3.2. Color-based Features Extraction

The color-based features have commonly been used in CBIR systems because of its easy and fast computation [14]. Color-based features can be extracted using a histogram of quantized values of color in Hue (H), Saturation (S), and Value (V) of the HSV color space. HSV color space is more robust to human perception as compared to the RGB color space. Due to the robustness of the HSV color space, first RGB images are converted to HSV color space and then uniform quantization is applied. Feature vectors are generated by considering the values of H=9, S=3, and

V=3 to form the feature vector of size 81 (9 × 3 ×3) bins. Representation of HSV color feature vector of an image is presented in Fig. 3.

3.3. Texture-based Features Extraction

Like color-based features, the texture-based features can be considered as powerful low-level features for image search and retrieval applications. There are certain works that have been done on texture analysis, classification, and segmentation for the last four decades. So far, there is no unique definition for the texture-based features. Texture is an attribute representing the spatial arrangement of the grey levels of the pixels in a region or image. Gabor filter is one of the widely used filters for texture-based feature extraction. It is a Gaussian function modulated by complex sinusoidal of frequencies and orientations. In our proposed approach, texture features of an image are extracted using a Gabor filter for five scales (s) and six orientations (o). The usage of multiple s and o makes the features rotation and scaling invariant on texture feature space. Five scales and six orientations produce thirty magnitudes. Consequently, mean and standard deviation need to be calculated for each magnitude and this leads to producing sixty features as a texture descriptor [14].

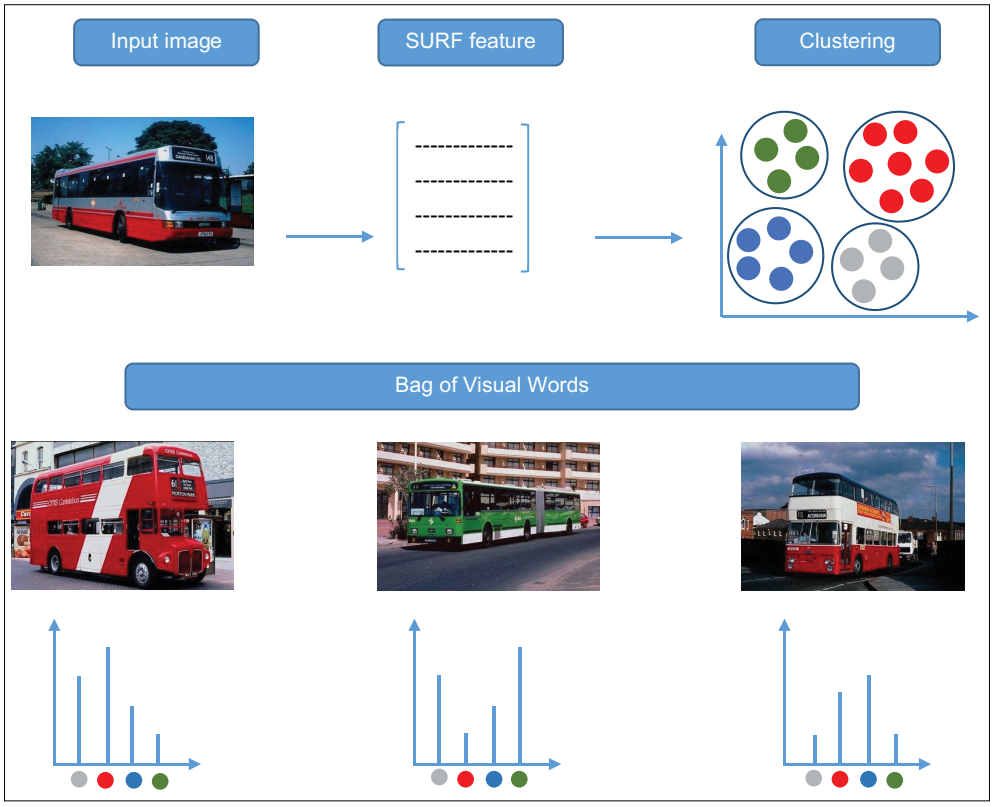

Fig. 2. Methodology of the bag of features based image representation for content-based image retrieval.

$$H=\begin{cases}0\ h\in [1,40]\\ 1\ h\in [41,80]\\ 2\ h\in [81,120]\\ 3\ h\in [121,160]\\ 4\ h\in [161,200]\\ 5\ h\in [201,240]\\ 6\ h\in [241,280]\\ 7\ h\in [281,320]\\ 8\ h\in [321,360]\end{cases} \quad S=\begin{cases}0\ s\in [0.00,0.30]\\ 1\ s\in [0.31,0.70]\\ 2\ s\in [0.71,1.00]\end{cases} \quad V=\begin{cases}0\ v\in [0.00,0.30]\\ 1\ v\in [0.31,0.70]\\ 2\ v\in [0.71,1.00]\end{cases}$$

Fig. 3. Hue saturation value feature vector.

3.4. Shape-based Features Extraction

Shape-based features are also useful to obtain more detailed characters of the images. Shape-based features include turn angle, central angle, distance between two feature points, distance between center of mass and feature point. Zernike Moments (ZMs) are used as a shape-based feature extractor in the proposed approach. ZMs are invariant to rotation, translation, and scaling [25]. Furthermore, ZMs are robust to noise and minor variations in shape and use Zernike polynomials to form feature vectors to represent an image based on shape features [26]. The proposed approach used 21 initial ZMs to represent the images.

3.5. Feature Matching

There are certain similarity measurements that used to compute the similarity between query image and images in the database, in our proposed approach, Manhattan distance is used for the BoF, see equation (1) [27], and Euclidean distance is used for color, texture and, shape features, equation (2) [13].

$$Manhattan\ Distance\left(MD\right)=\sum_{i=1}^{f}\left|Qf-Df\right| \quad (1)$$

$$Euclidean\ Distance\left(ED\right)=\sum_{i=1}^{f}\sqrt{\left(Qf-Df\right)^2} \quad (2)$$

Where $Qf=(Qf_1, Qf_2,\ldots, Qf_{L-1})$ is the feature vector of query image, $Df=(Df_1, Df_2,\ldots, Df_{L-1})$ is the feature vector of the database of images, and L is the dimension of image feature.

Next section will present the proposed approach in detail.

4. PROPOSED APPROACH

This section presents the detailed steps of the proposed bi-layer approach as follows:

1- Let Q be the query image, and $I_{db}=\{I_1, I_2,\ldots, I_n\}$ be the database of n images.
2- First layer entails the following steps:
a) Q_{BoF} and I_{BoF} represent feature vector of Q and I_{db}, respectively, after BoF technique is applied.
b) Manhattan similarity measurement is used to find the similarity between Q_{BoF} and I_{BoF} and as a result, M most similar images to the query image are retrieved.
3- Second layer will implement on the query image Q and the M most similar images M_i that were retrieved/obtained from the first layer. It includes the following steps:
a) Extract the following features from Q and M_i:
- Let $C=\{c_1, c_2,\ldots\ldots, c_{81}\}$ be the extracted 81 color-based features that represent the 81 bins of the quantized HSV color space.
- Let $T=\{t_1, t_2,\ldots\ldots, t_{60}\}$ be the extracted 60 texture-based features using Gabor filter.
- Let $S=\{s_1, s_2,\ldots\ldots, s_{21}\}$ be the extracted 21 shape-based features using ZMs.
- Let $F=C+T+S$ be the feature vector of the fused/combination of all the extracted features above.
- Finally, Q_F *and* M_{Fi} represent the fused feature vector of Q and M_i.

b) Euclidean similarity measurement is used to find the similarity between Q_F and M_{Fi} to retrieve the N most similar match images to the query image.

The block diagram of the proposed bi-layer approach is illustrated in Fig. 4.

5. EXPERIMENTAL RESULTS

In this section, experiments are performed comprehensively to assess the performance of the proposed approach in terms of precision rate, the most common confusion matrix measurement used in the CBIR research area. Furthermore, the proposed approach is compared to the most recent existing works.

5.1. Dataset

The experiments are conducted on the public and well-known dataset called Corel-1K that contains 1000 images in the form of ten categories and each category consists of 100 images with resolution sizes of (256 × 384) or (384 × 256) [28]. The categories are organized as follows: African, people, beaches, buildings, buses, dinosaurs, elephants, flowers, horses, mountains, and foods.

5.2. Evaluation Measurements

Precision confusion matrix measurement is used to assess the performance of the proposed approach. The precision determines the number of correctly retrieved images over the total number of the retrieved images from the tested database of images and it measures the specificity of image retrieval system, as presents in the following equation [21]:

$$Precision=\frac{R_c}{R_t} \quad (3)$$

where R_c represents the total number of correctly retrieved images and R_t represents the total number of retrieved

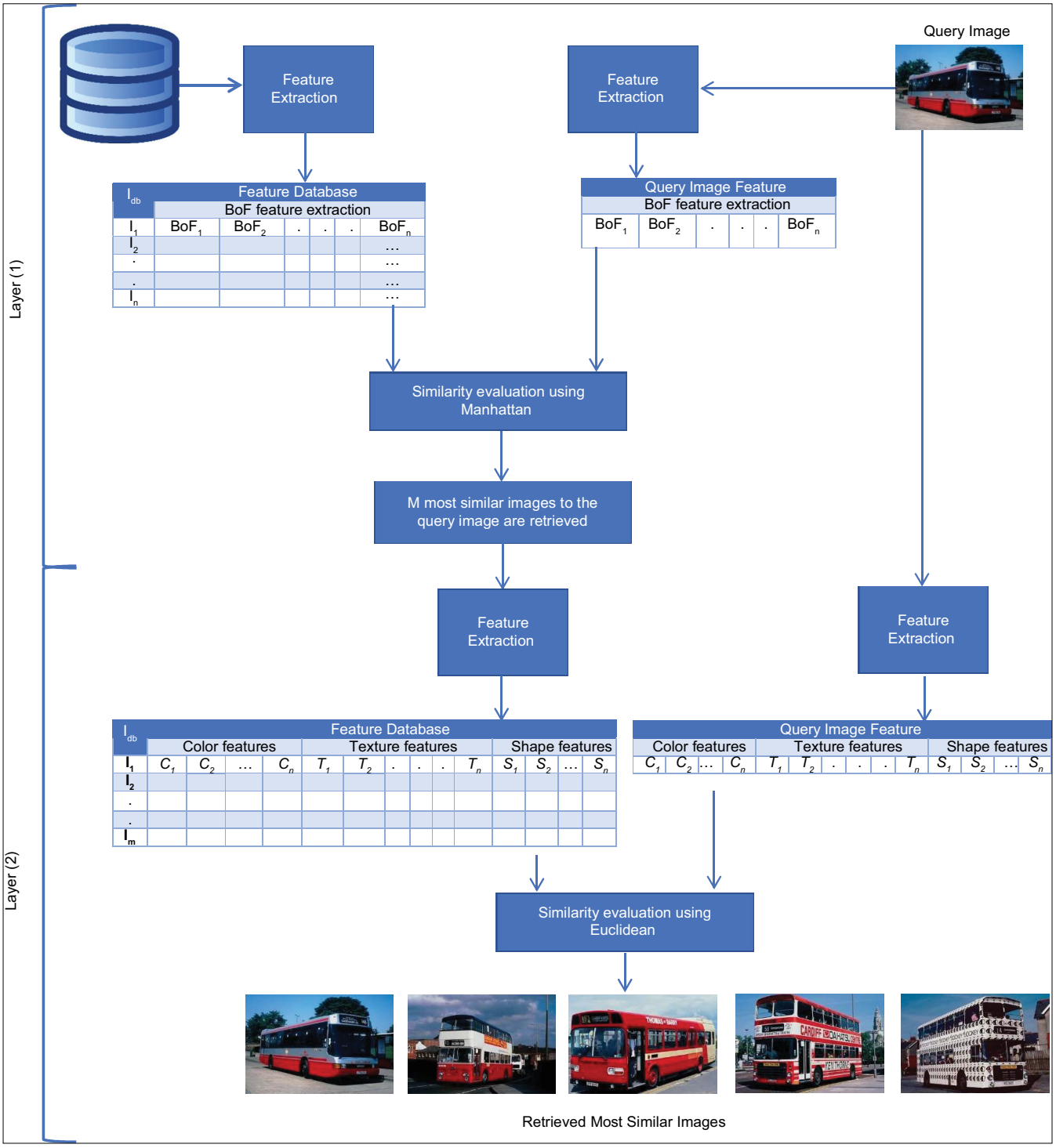

Fig. 4. Block diagram of the proposed bi-layer content-based image retrieval system for top-5.

images. Precision can also be expressed in the following equation.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Where TP represents true positive and FP represents false positive. In this work, top-10 and top-20 have been tested. Top-10 means the total number of retrieved images is 10 images, and top-20 means the total number of retrieved

images is 20 images. Figs. 5 and 6 present examples for the query image based on top-10 and top-20.

5.3. Results

The experiments conducted in this work involve two phases: (a) Single layer CBIR model and (b) Bi-layer CBIR model. In the first phase, the single layer model (i.e., BoF technique) is evaluated alone, and on the other hand, CBIR technique based on extracting shape, texture, and color features is evaluated. In the second phase, the proposed bi-layer model is evaluated. The experiments are detailed in the following steps:

1. BoF-based CBIR technique is tested with different number of clusters, as BoF technique depends on the K-means clustering algorithm to create clusters, which is commonly called visual words. The number of clusters cannot be selected automatically; it needs manual selection. To select the proper number of clusters, (i.e., value of k-means), the different number of clusters have been tested to obtain the best precision result of BoF technique. The precision results of different numbers of clusters are illustrated in the following tables.

From Tables 1 and 2, one can observe that the best result is obtained when k = 500 for both top-10 and top-20.

2. The proposed CBIR technique that relies on extracting shape, color, and texture features has been tested, and the results are presented in Table 3.
3. The proposed bi-layer approach has been tested. It includes two layers: First layer implements BoF technique (for K=500) and M most similar images are retrieved, M is user defined. In the second layer, shape, color, and texture features are extracted from the query image and the M images, as a result, N most similar images are retrieved. The following tables investigate the best value of M. In other words, Tables 4 and 5 show testing different number of M for top-20 and top-10, respectively.

Results in Tables 4 and 5 demonstrate that the best precision results are obtained when M = 200. For this reason, different small numbers of M, in the range of M = 100 to M = 300, are investigated to gain better precision results, Tables 6 and 7.

From Tables 6 and 7, it is quite clear that the best result is obtained when M =225 for both top-20 and top-10.

TABLE 1: Precision rate of BoF technique for different number of clusters for top-20

Categories	Different number of clusters									
	K=100	K=200	K=300	K=400	K=500	K=600	K=700	K=800	K=900	K=1000
Africa	52.05	55.85	55.25	56.25	55.85	56.35	55.8	55.5	54.15	55.5
Beaches	44.35	45.7	45.85	46.35	47.2	45.4	47.15	45.35	46.65	48.15
Buildings	41.3	44.75	46.5	47.55	49.05	50.4	51.25	52.25	52.55	52.15
Buses	83.5	86.15	85.15	86.3	86.75	85.4	84.75	84.4	84.65	83.5
Dinosaur	100	100	100	100	100	99.95	99.95	100	99.95	99.95
Elephant	55.85	59.95	62	61.45	60.5	59.65	58.55	57.85	57.15	58.1
Roses	84.35	84.4	84.95	85.4	85.45	84.55	85.75	86.15	84.1	86.1
Horses	85.9	86.4	87.9	88.6	89.35	87.85	88.9	88.15	88.55	88.7
Mountains	39.5	40.7	41.9	42.5	43.1	45.95	45	46.95	45.6	45.55
Food	39.45	42	41.05	40.35	41	39.3	39.35	38.2	38.45	37.25
Averages	62.625	64.59	65.055	65.475	**65.825**	65.48	65.645	65.48	65.18	65.495

TABLE 2: Precision rate of bof technique for different number of clusters for top-10

Categories	Different number of clusters									
	K=100	K=200	K=300	K=400	K=500	K=600	K=700	K=800	K=900	K=1000
Africa	58.9	60.4	61	62.1	62.6	60.8	62.1	61.4	61.9	59
Beaches	50.8	52	51.6	51.3	52.5	50.6	51.4	51.7	50.5	51.1
Buildings	50.1	54.4	56.1	58.1	58.4	59.7	59.5	58.3	60.6	61
Buses	89.2	89.5	89.7	89.6	89.2	88.4	88.4	88.6	87.9	87.4
Dinosaur	100	100	100	100	100	100	100	100	100	100
Elephant	69.1	71.3	73.2	69.8	71.4	69.6	70.5	69.7	67.3	66.6
Roses	86.6	88.3	88.5	88.4	87.8	88.9	88	87.6	88.4	89.3
Horses	88.8	91.2	93.3	93.2	93.7	93.9	93.7	93.1	94.4	94
Mountains	46	49.1	50.6	50.2	50.4	52.4	50.6	53	52.2	52.5
Food	46.7	49	50.5	49.4	48.6	47.8	48.4	46.6	44.5	45.8
Averages	68.62	70.52	71.45	71.21	71.46	71.21	71.26	71	70.77	70.67

Fig. 5. Query result for top-10.

Fig. 6. Query result for top-20.

TABLE 3: Precision rate for the tested feature extractors for top-20 and top-10

Categories	Top-20	Top-10
Africa	70.55	75.4
Beaches	36.55	43.6
Buildings	42.4	50.9
Buses	72.85	79.8
Dinosaur	92.45	95.9
Elephant	40.5	54.3
Roses	58.9	72.2
Horses	84.9	89.5
Mountains	34.9	40.8
Food	65.7	70.5
Averages	59.97	67.29

The ratio of correctly retrieved images over the total number of images of the semantic class in the image database is known as recall and it measures the sensitivity of the image retrieval system, equation (5):

$$Recall = \frac{R_c}{T_s} \tag{5}$$

Where Rc is the total number of retrieved images and Ts is the total number of images in the semantic class in the database. More experiments have been done to compare the proposed approach with the state-of-the-art

TABLE 4: Precision rate for different number of M for top-20

Categories	Different number of M									
	M=100	M=200	M=300	M=400	M=500	M=600	M=700	M=800	M=900	M=1000
Africa	78.65	77.75	76.75	76.6	76.2	76.5	76.65	76.65	76.3	76.25
Beaches	55.7	56.65	56.55	56.5	55.5	54.6	54	53.85	53.7	53.55
Buildings	56.1	56.05	56.35	56.35	56.15	56.65	57	56.6	56.5	56.55
Buses	94.45	94.25	93.8	93.45	93.1	92.55	92.25	92.15	91.85	91.8
Dinosaur	100	100	100	100	100	100	100	100	100	100
Elephant	64.05	64.25	64.5	64.8	65.15	65.15	65.05	64.65	64.65	64.45
Roses	91.35	91.15	90.7	89.75	89.3	88.75	88.6	88.3	88.05	87.9
Horses	94.65	94.95	94.9	94.9	94.8	94.7	94.6	94.6	94.6	94.6
Mountains	52	52.65	52.15	52	50.9	50.65	50	49.55	49.25	49.3
Food	68.55	72.15	72.6	72.75	72.35	71.55	70.85	70.85	70.55	70.4
Averages	75.55	75.985	75.83	75.71	75.345	75.11	74.9	74.72	74.545	74.48

TABLE 5: Precision results for different number of M for top-10

Categories	Different number of M									
	M=100	M=200	M=300	M=400	M=500	M=600	M=700	M=800	M=900	M=1000
Africa	83.7	82.4	82.8	82.5	82	82	82	81.7	81.7	81.6
Beaches	63.4	64	63	62.6	61.6	61	60.4	60	59.8	59.8
Buildings	68.4	67.5	66.6	66.6	66.3	66.9	67	66.6	66.8	66.8
Buses	97.5	96.8	96.3	96	95.6	95.1	94.6	94.6	94.6	94.6
Dinosaur	100	100	100	100	100	100	100	100	100	100
Elephant	77.8	78.4	79.4	78.9	79.1	79	78.7	78.5	78.5	78.3
Roses	94.4	94.2	94	93.6	93.3	93	92.9	92.6	92.5	92.5
Horses	97.6	97.4	97.4	97.4	97.3	97.3	97.2	97.1	97.1	97.1
Mountains	60.4	62.4	61.3	60.3	59	58.6	58.3	57.6	56.9	57
Food	77.8	78.6	79	78.9	78.6	78.3	77.4	77.5	77.4	77.4
Averages	82.1	82.17	81.98	81.68	81.28	81.12	80.85	80.62	80.53	80.51

TABLE 6: Precision results for different number of M in the range 100–300 for top-20

Categories	Different number of M								
	M=100	M=125	M=150	M=175	M=200	M=225	M=250	M=275	M=300
Africa	78.65	78.4	78.15	77.7	77.75	78	77.4	77.2	76.75
Beaches	55.7	56.35	56.75	57	56.65	57.1	56.3	56.3	56.55
Buildings	56.1	56.45	56.55	56.5	56.05	56.4	56.45	56.05	56.35
Buses	94.45	94.8	95.05	94.6	94.25	94.55	94.2	94.1	93.8
Dinosaur	100	100	100	100	100	100	100	100	100
Elephant	64.05	64.5	63.9	64.1	64.25	64.3	64.2	64.4	64.5
Roses	91.35	91.4	91.3	91.5	91.15	90.95	90.9	90.6	90.7
Horses	94.65	94.9	95.05	94.95	94.95	94.95	94.85	94.95	94.9
Mountains	52	52.8	52.6	52.3	52.65	52.65	52.55	52.4	52.15
Food	68.55	69.4	71.3	71.45	72.15	72.4	72.6	72.7	72.6
Averages	75.55	75.9	76.065	76.01	75.985	76.13	75.945	75.87	75.83

techniques, Table 8. In all experiments, each image in the Corel-1K database is used as a query image. The retrieval performance of tested techniques is measured in terms of average retrieval precision (ARP) and average retrieval recall (ARR). The higher ARP and ARR values mean the better performance.

According to the results in Table 8, the best result is achieved by the proposed approach for both top-10 and top-20. All the tested state-of-the-art techniques, except technique in Al-Jubouri and Du [7], they tested their approach either for top-10 or for top-20, and this is why in Table 8 some cells do not contain the ARP and ARR results.

TABLE 7: Precision results for different number of M in the range 100–300 for top-10

Categories	Different number of M								
	M=100	M=125	M=150	M=175	M=200	M=225	M=250	M=275	M=300
Africa	83.7	83.2	82.7	82.4	82.4	82.1	82.3	82.7	82.8
Beaches	63.4	63.6	63.6	63.9	64	64.1	63.1	63.1	63
Buildings	68.4	68.3	68.3	67.7	67.5	67.4	66.7	66.8	66.6
Buses	97.5	97.3	97.2	97.2	96.8	96.9	96.6	96.6	96.3
Dinosaur	100	100	100	100	100	100	100	100	100
Elephant	77.8	78.6	78.4	78.3	78.4	79.1	79	79.4	79.4
Roses	94.4	94.4	94.6	94.5	94.2	94.1	93.9	93.9	94
Horses	97.6	97.4	97.4	97.3	97.4	97.3	97.4	97.4	97.4
Mountains	60.4	61	61.5	61.7	62.4	62.7	62.5	61.8	61.3
Food	77.8	77.8	78.5	77.8	78.6	79	79	78.7	79
Averages	82.1	82.16	82.22	82.08	82.17	82.27	82.05	82.04	81.98

TABLE 8: ARP results of tested CBIR techniques

Approaches	Top-10		Top-20	
	ARP	ARR	ARP	ARR
Proposed approach	82.27	8.22	76.13	15.22
[7]	64.00	6.40	59.60	11.92
[18]	-	-	73.5	14.7
[21]	67.60	6.76	-	-
[22]	-	-	70.48	14.09

6. CONCLUSION

This paper has developed an effective CBIR technique for retrieving images from a wide range of images in the dataset. The proposed approach involves two layers; in the first layer, all images in the database are compared to the query image based on the BoF technique, and as a result, 225 most similar images to the query image are selected. Color, texture, and shape features are used in the second layer to extract significant features from the selected 225 images to retrieve the most similar images to the query image. The obtained results depicted that the proposed approach has reached an optimal average precision of 82.27% and 76.13% for top 10 and top 20, respectively.

REFERENCES

[1] Z. F. Mohammed and A. A. Abdulla. "Thresholding-based white blood cells segmentation from microscopic blood images". *UHD Journal of Science and Technology*, vol. 4, no. 1, pp. 9-17, 2020.

[2] M. W. Ahmed and A. A. Abdulla. "Quality improvement for exemplar-based image inpainting using a modified searching mechanism". *UHD Journal of Science and Technology*, vol. 4, no. 1, pp. 1-8, 2020.

[3] A. A. Abdulla, H. Sellahewa and S. A. Jassim. *"Secure Steganography Technique Based on Bitplane Indexes"*. 2013 IEEE International Symposium on Multimedia, United States, pp. 287-291, 2013.

[4] A. A. Abdulla. "*Exploiting Similarities Between Secret and Cover Images for Improved Embedding Efficiency and Security in Digital Steganography*, *PhD Thesis*". 2015. Available from: http://www.bear.buckingham.ac.uk/149. [Last accessed on 2020 Dec 15].

[5] A. Sarwar, Z. Mehmood, T. Saba, K. A. Qazi, A. Adnan and H. Jamal. "A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine". *Journal of Information Science*, vol. 45, no. 1, pp. 117-135, 2019.

[6] S. Hossain and R. Islam. "A new approach of content based image retrieval using color and texture features". *Current Journal of Applied Science and Technology*, vol. 51, no. 3, pp. 1-16, 2017.

[7] J. Pradhan, A. Ajad, A. K. Pal and H. Banka. "Multi-level colored directional motif histograms for content-based". *The Visual Computer*, vol. 36, pp. 1-22, 2019.

[8] L. K. Pavithra and T. S. Sharmila. "Optimized feature integration and minimized search space in content based image retrieval". *Procedia Computer Science*, vol. 165, pp. 691-700, 2019.

[9] H. F. Atlam, G. Attiya and N. El-Fishawy. "Comparative study on CBIR based on color feature". *International Journal of Computer Applications*, vol. 78, no. 16, pp. 9-15, 2013.

[10] Y. D. Mistry. "Textural and color descriptor fusion for efficient content-based image". *Iran Journal of Computer Science*, vol. 3, pp. 1-15, 2020.

[11] T. Kato. "Database architecture for content-based image retrieval". *International Society for Optics and Photonics*, vol. 1662, pp. 112-123, 1992.

[12] C. H. Lin, R. T. Chen and Y. K. Chan. "A smart content-based image retrieval system based on color and texture feature". *Image and Vision Computing*, vol. 27, no. 6, pp. 658-665, 2009.

[13] Z. C. Huang, P. P. Chan, W. W. Ng and D. S. Yeung. "Content-based image retrieval using color moment and Gabor texture feature". *2010 International Conference on Machine Learning and Cybernetics*, vol. 2, pp. 719-724, 2010.

[14] M. Singha and K. Hemachandran. "Content based image retrieval using color and texture". *Signal and Image Processing*, vol. 3, no. 1, p. 39, 2012.

[15] J. Yu, Z. Qin, T. Wan and X. Zhang. "Feature integration analysis of bag-of-features model for image retrieval". *Neurocomputing*, vol. 120, pp. 355-364, 2013.

[16] S. Somnugpong and K. Khiewwan. "*Content-based Image*

Retrieval Using a Combination of Color Correlograms and Edge Direction Histogram". 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Thailand, pp. 1-5, 2016.

[17] H. Al-Jubouri and H. Du. "A content-based image retrieval method by exploiting cluster shapes". *Iraqi Journal for Electrical And Electronic Engineering*, vol. 14, no. 2, pp. 90-102, 2018.

[18] A. Nazir, R. Ashraf, T. Hamdani and N. Ali. "*Content Based Image Retrieval System by Using HSV Color Histogram, Discrete Wavelet Transform and Edge Histogram Descriptor*". 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Pune, pp. 1-6, 2018.

[19]. H. Qazanfari, H. Hassanpour and K. Qazanfari. "Content-based image retrieval using HSV color space features". *International Journal of Computer and Information Engineering*, vol. 13, no. 10, pp. 537-545, 2019.

[20] A. Rashno and E. Rashno. "Content-based image retrieval system with most relevant features among wavelet and color features". *arXiv preprint arXiv*, vol. 2019, pp. 1-18.

[21] S. P. Rana, M. Dey and P. Siarry. "Boosting content based image retrieval performance through integration of parametric and nonparametric approaches". *Journal of Visual Communication and Image Representation*, vol. 58, pp. 205-219, 2019.

[22] F. Sadique, B. K. Biswas and S. M. Haque. "*Unsupervised Content-based Image Retrieval Technique Using Global and Local Features*". 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Bangladesh, pp. 1-6, 2019.

[23] S. Jabeen, Z. Mehmood, T. Mahmood, T. Saba, A. Rehman and M. T. Mahmood. "An effective content-based image retrieval technique for image visuals representation based on the bag-of-visual-words model". *PloS One*, vol. 13, no. 4, pp. 1-24, 2018.

[24] J. Zhou, X. Liu, W. Liu and J. Gan. "Image retrieval based on effective feature extraction and diffusion process". *Multimedia Tools and Applications*, vol. 78, no. 5, pp. 6163-6190, 2019.

[25] F. Rajam and S. Valli. "A survey on content based image retrieval". *Life Science Journal*, vol. 10, no. 2, pp. 2475-2487, 2013.

[26] J. Olaleke, A. Adetunmbi, B. Ojokoh and I. Olaronke. "An appraisal of content-based image retrieval (CBIR) methods". *Asian Journal of Research in Computer Science*, vol. 3, pp. 1-15, 2019.

[27] M. Sharma and A. Batra. "Analysis of distance measures in content based image retrieval". *Global Journal of Computer Science and Technology*, vol. 14, no. 2, p. 7, 2014.

[28] J. Li and J. Z. Wang. "Automatic linguistic indexing of pictures by a statistical modeling approach". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1075-1088, 2003.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Modeling Groundwater Potential Zones across Sulaimani Governorate Using Geographic Information System and Multi-influencing Factor Techniques

Haveen Muhammed Rashid*

Department of Water Resources, College of Engineering, University of Sulaimani, KRG, Iraq

ABSTRACT

Groundwater is one of the most important natural resources in the world. The presence of groundwater is the result of interaction of several factors such as: hydrology, geology, climate, ecology, and physiography. The purpose of this paper is to produce groundwater potential zones which are useful in determining the amount of groundwater available in Sulaimani Governorate, North of Iraq. Geographic information system database for six different thematic layers (digital elevation model, rainfall, soil texture, drainage density, slope and land use/land cover) were generated. The study approach involved integration of six layers carried out based on the multiplication of each data raster values with specific weight using weighted overlay analysis method. Raster maps of all the layers assigned a fixed score and weight using multi-influencing factor technique. Based on the resulted map the study area has been divided into four zones that had very high potential zone (1%), high potential zone (14%), moderate zone potential (79%) and low potential zone (6%). About 50% of the high groundwater potential zone were located in Halabja, Rania, and Pshdar districts. Obtained results can be useful for exploration in regional areas, preventing excessive exploitation of groundwater and planning for suitable sites of artificial groundwater.

Index Terms: Geographic information system, Groundwater potential, Multi-influencing factor, Sulaimani, Spatial distribution

1. INTRODUCTION

Groundwater is a substantial source of freshwater; however, it does not always available sufficiently where it is needed. Population growth, industry development and climate change increased demand of fresh water, which causes the decline of groundwater table [1]. Therefore, the lower availability of groundwater by side of increase in its development, require sustainable groundwater management [2] and [3].

Geographic information system (GIS) and remote sensing are the two mechanisms that have been used remarkably in groundwater management researches. GIS is widely used for modeling process by preparing the input parameter for the model through generation of digital geographic database [4]. Assessment of groundwater resources of an area requires preparation and integration of many factors such as: slope, drainage, land use (LU), rainfall, elevation, soil texture, geological structures and geomorphic features [4]. Integration of various geospatial information is carried out by taking values of each data raster and multiplying them

Access this article online

DOI:10.21928/uhdjst.v5n1y2021.pp13-20 E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Haveen Muhammed Rashid, Department of Water Resources, College of Engineering, University of Sulaimani, Kurdistan Region – F.R. Iraq. E-mail: haveen.rashid@univsul,edu.iq

Received: 10-11-2020 Accepted: 11-02-2021 Published: 15-02-2021

with the specific weight. Weights determination is intuitive and is obtained from the values used in the literatures or from the experience.

Many researches were studied this issue utilizing both GIS and remote sensing techniques [5]-[17]. In 2012 Jani published a research used GIS to model groundwater flow for the aquifer called Lincolnshire Limestone in the Slea catchment, United Kingdom. The resulted model has confirmed the advantage of using GIS tool in modeling groundwater [2].

Rawal and Vyas (2016) presented a research to delimit the water logging in Mehsana district, India using GIS application in groundwater modeling and identify the perched aquifers in their selected area. Their study used factors such as soil layer, LU layer, and the temporal distribution of water logging. The resulted model suggests convenient method to control water logging and identifies the groundwater regime to estimates the total recharge [18].

In a study conducted by Singh, GIS techniques were used to identify the recharge zones of groundwater in New Zealand. Many data sets were prepared such as soil layer, LU layer, aspect, slope, lithology, and drainage density with 500 m × 500 m spatial resolution. The data set overlies to develop potential zones of the groundwater. The output results explain; the groundwater has low potential in both residential and high elevation areas, while areas of water body and low elevations fall in the high potential zones [19].

Western Ghats river basin/India was selected to demonstrate the groundwater model. To delineate groundwater zones, 12 layers were utilized such as geology, drainage density layer, rainfall, soil, slope, geomorphology, lineament density, LU/land cover [LC], roughness, topographic position index, topographic wetness index, and curvature. The relative weights to each class in all maps were allocated using Analytical Hierarchy Process [20].

Karim and Al-Manmi selected Halabja Said Sadiq sub-basin as a study area to model the groundwater recharge zones using GIS and geophysical mechanisms. Eight thematic maps were used such as hydrogeology, LU/LC, topography, drainage density, lineaments, soil type, slope, and rainfall with four geoelectrical resistivity profiles. Three zones of groundwater potential delineated which are low, moderate and high and cover 33%, 24%, and 42% of the total area, respectively. Spatially, the highest zone is located along with the Quaternary deposits which characterized by high lineament density, low slop, and pediment deposition [13].

Groundwater recharge zones in a tropical river basin of Kerala, India were identified using integration of different layers. Ten thematic layers were prepared for mapping groundwater recharge. Four classes have been identified and approximately half of the basin area is located in two recharge zones; very high and high zones which are appropriate for groundwater recharge [3].

2. STUDY AREA

The proposed study area is located in Sulaimani Governorate, north of Iraq – Kurdistan Region, as shown in Fig. 1. According to the topographic map, the elevation is ranged from 182 to 3430 m. The latitudes are between (34° 32' 15" N and 36° 34' 15" N), and the longitudes are between (44° 30' 30 E and 46° 20' 30 E). The total area calculated as 18,525 km^2 and the average annual rainfall of the study area is (618 mm) for 15 years from 2000 to 2014 taken from 14 metrological stations distributed within the proposed area. The study area has hot and cold weather in summer and winter, respectively, with long summer and winter season in compare to spring and autumn season [21], [22].

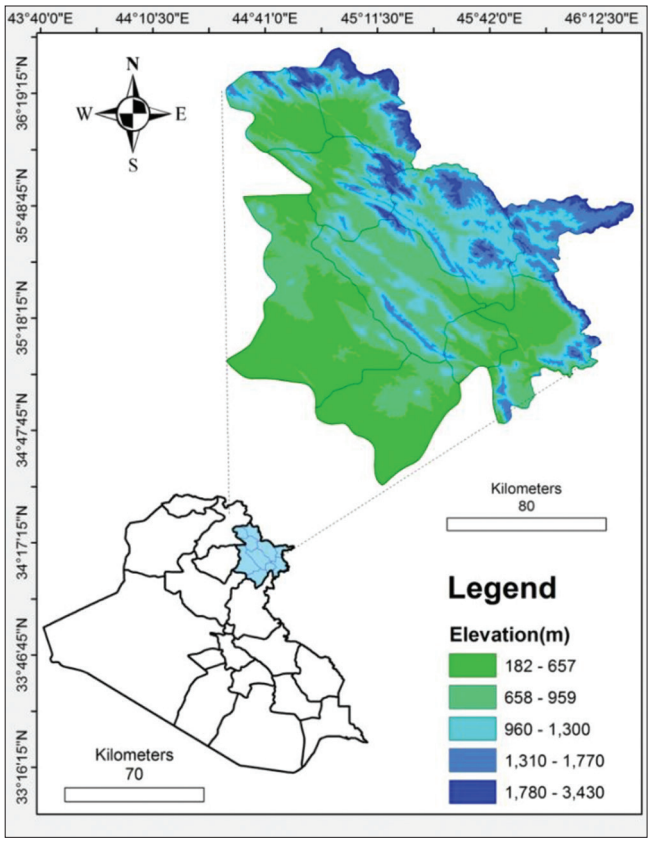

Fig. 1. Study area location and digital elevation model.

3. MATERIALS AND METHODS

Different factors (layers) were used for the modeling process. Selection of factors was based on data availability across Sulaimani Governorate. The data were collected from various sources and include: Digital Elevation Model (DEM), soil texture, LU/LC, rainfall, slope, and drainage density. All the raster layers were projected using Universal Transverse Mercator UTM Projection Zone 38, Datum WGS84 with 30 × 30 m resolution.

The method used in this study consists two main steps: (a) Geospatial database generation, (b) generation of weight for groundwater prospecting factors and associated features. The complete process of the groundwater potential zone delineation is shown as flowchart in Fig. 2. All the factors prepared for the current study and their impact on the occurrence of groundwater in Sulaimani are illustrated below:

3.1. DEM

The DEM data for Sulaimani as shown in Fig. 1 with elevations ranged from (182 to 3430 m) classified into five classes spatially distributed in the area as (35%, 31%, 19%, 12%, and 4%) from low to high elevations. The data were downloaded from USGS earth explorer website using Shuttle Radar Topography Mission1 Arc Second-Global [23].

3.2. LU/LC

LC marks the physical land type such as water or forest; however, the LU provides information how people are using the land [24]. LULC gives the primary information on infiltration, soil moisture, surface water, and groundwater, additionally to give a signal on groundwater demand [25]-[29].

The LU map was created through image processing of remote sensing data using Arc GIS software (Iso Cluster Unsupervised Classification), by downloading 3 Landsat8 images (acquired on October 2019) with 7 bands for each image [23]. The study area was classified to 5 classes of LU/LC: Water, Forest, Crop, Urban area, and Bare Soil which is shown in Fig. 3.

3.3. Soil Map

Soil properties influence groundwater recharge. The soils with coarse texture has high permeability which increase's groundwater recharge, while fine texture soils with low permeability decrease the groundwater recharge [30]. Therefore, soil properties such as permeability, porosity, texture, and structure have considerable effect on groundwater recharge [3] and [31].

The soil map of Sulaimani area was designed and classified to 5 classes based on percent of fine and course aggregates (soil texture) as shown in Fig. 4. This map shows the distribution of the soil texture in the study area as: (Sandy Loam 1%, Clay 17%, Loam 36%, Sandy clay 9%, and Sandy Clay Loam 37%). The major soil types in the study area were grouped as: (Leptosol LP, Vertisol VR, Calcisol CL, and Gypsisol GY) [32]. The data were

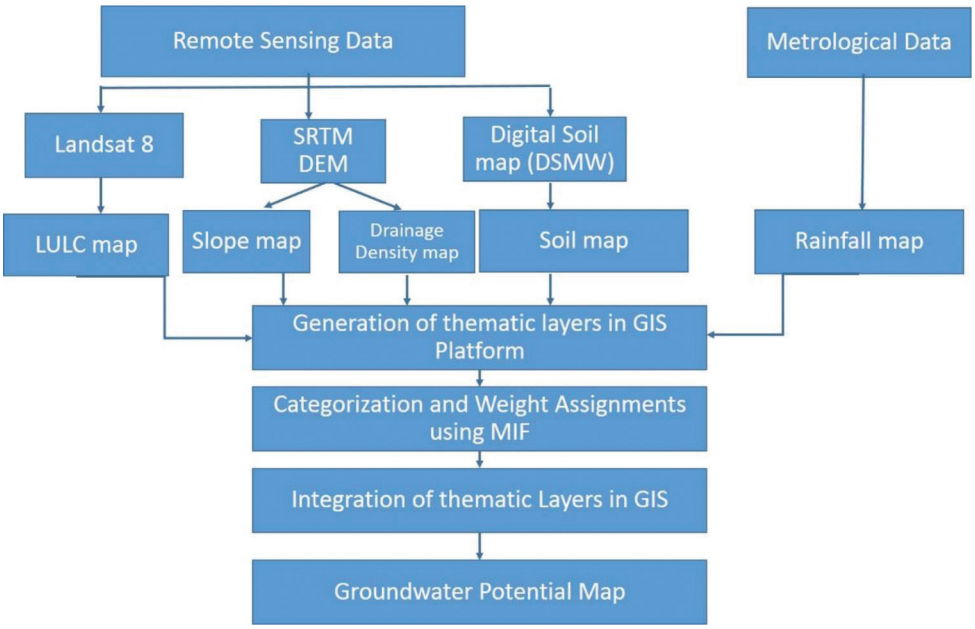

Fig. 2. Flowchart for delineating groundwater potential.

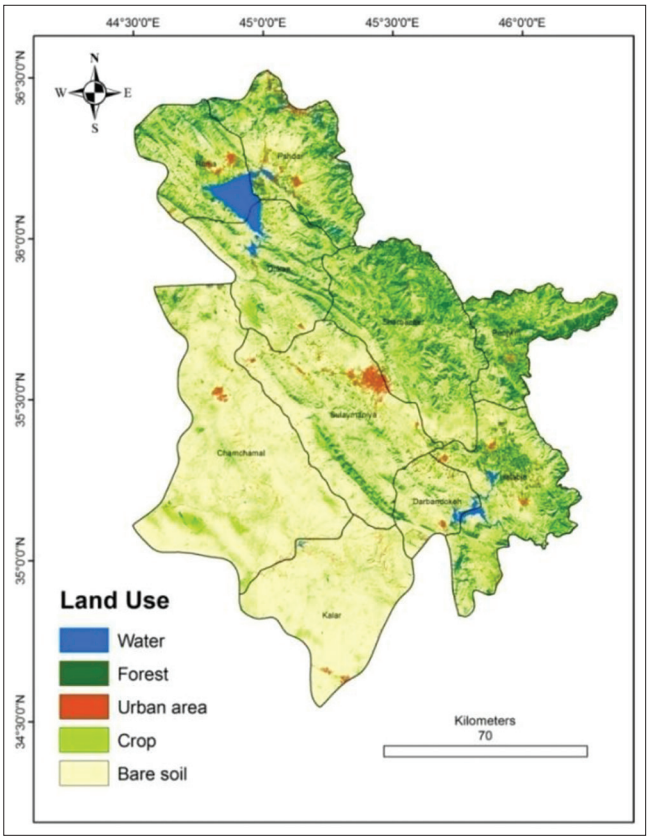

Fig. 3. Land use/land cover of the study area.

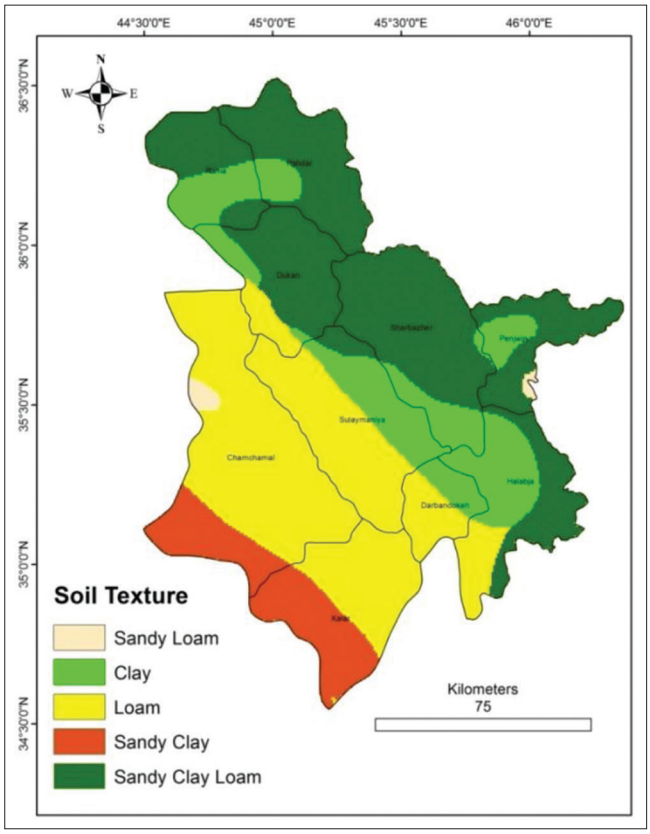

Fig. 4. Soil texture map of the study area.

downloaded from the FAO-UNESCO Soil Map of the World and the present version is (3.6) of the digitized Soil Map of the World [33].

3.4. Rainfall Map

The groundwater recharge is remarkably affected by patterns of rainfall during recharge seasons [34], [35]. The average annual rainfall for the 14 stations from 2000 to 2014 in Sulaimani is represented in Fig. 5a. Rainfall map was created by Arc GIS utilizing Inverse Distance Weighted interpolation method with 5 classes in which each class represent different ranges of rainfall distributed as (22%, 21%, 27%, 23%, and 7%) from low to high ranges shown in Fig. 5b.

3.5. Slope Map

The slope map was prepared from the DEM using 5 classes in degree and the spatial distribution in the study area are (very low slope 42%, low slope 26%, moderate slope 17%, high slope 11%, and very high slope 3%) as shown in Fig. 6. Slope plays a notable effect on groundwater flow. The slope is considered as an essential factor for runoff generation due to its control of the division of precipitation into runoff and infiltration. A higher slope results in low recharge and fast runoff. Hence, the slope is inversely correlated with groundwater potential [8], [36]-[39].

3.6. Drainage Density

Drainage density can be calculated by taking the ratio of total length of all the streams in a drainage basin to the total area of the basin [40]. The drainage density has a reciprocal relation with the permeability. Drainage density with high values produces less permeability whereas low drainage density shows the areas of high permeability. Accordingly, it is a significant factor for modeling the groundwater potential zone [20]. Drainage density is calculating using the equation below [27]:

$$Drainage\ density = \frac{total\ length\ channels\ (km)}{basin\ area\ (km^2)} \quad (1)$$

Map of the drainage density (Fig. 7) was generated using ArcGIS tool called line density analysis and 5 classes have been identified ranged between (0.21 and 2.9 km/km^2).

3.7. Multi-influencing Factors (MIF)

Six factors, namely, DEM, LU/LC, soil texture, rainfall, slope, and drainage density have been specified to model the

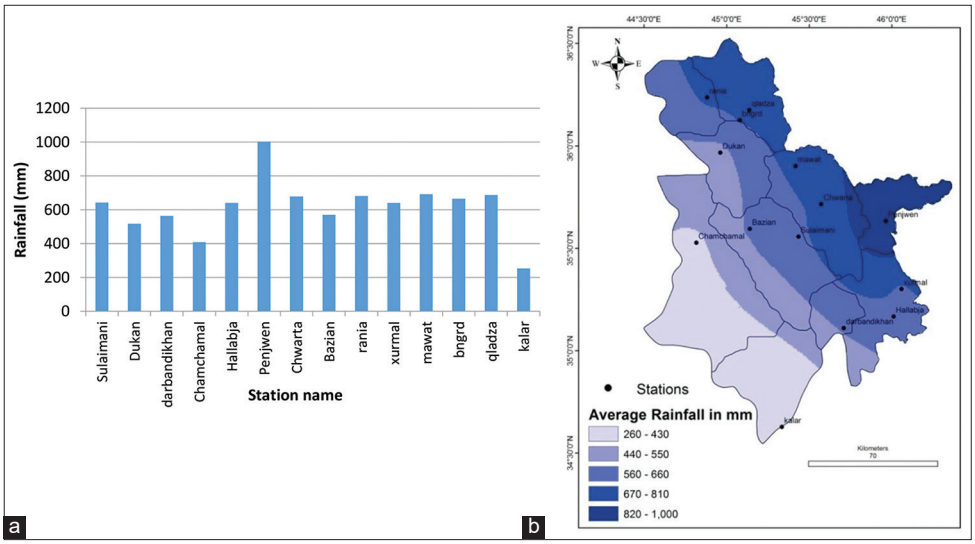

Fig. 5. (a). Average annual rainfall histogram of the stations. (b) Rainfall map of the study area.

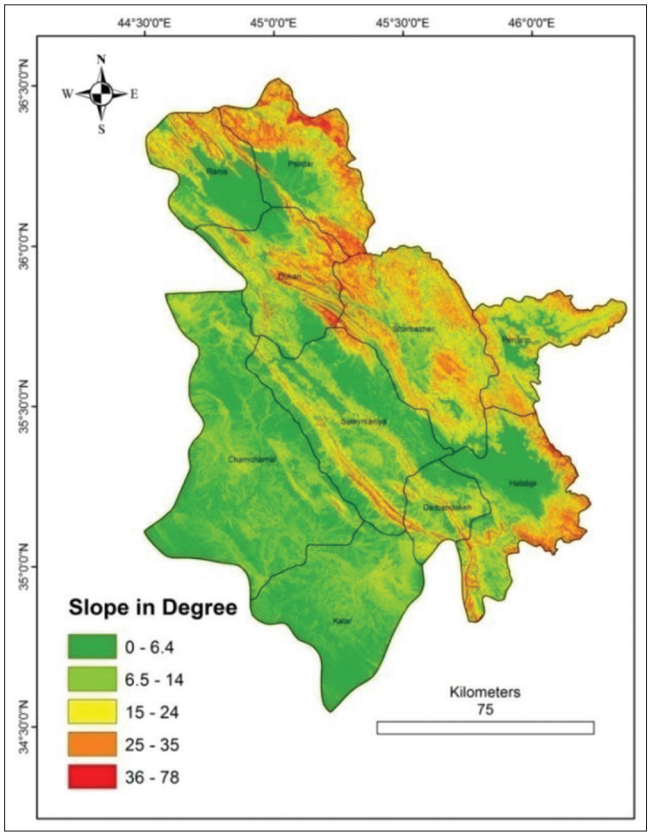

Fig. 6. Slope of the study area.

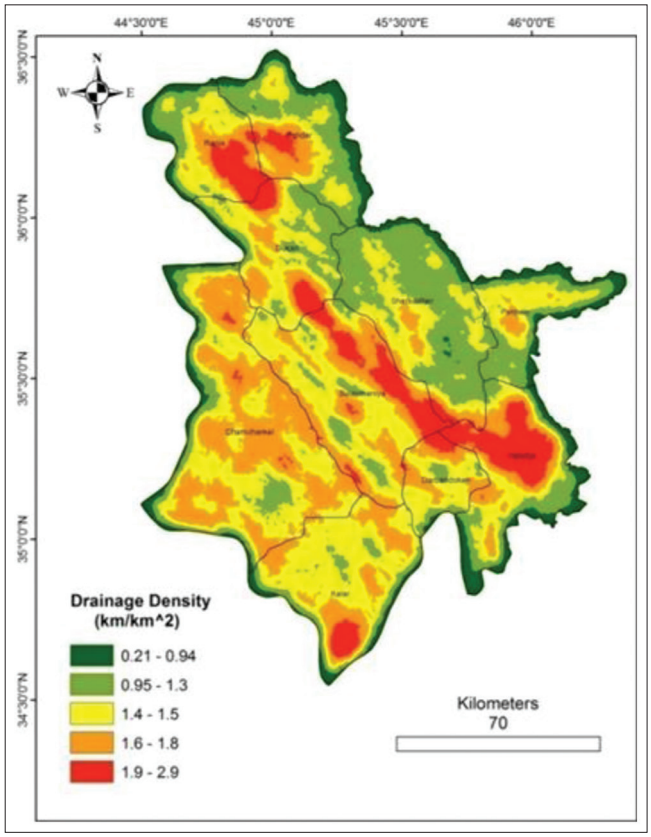

Fig. 7. Drainage density of the study area.

groundwater recharge potential zones. The mutual relations between these factors and their effect are shown in Fig. 8. These factors are interdependent and their influence on the recharge potential of groundwater has been estimated using MIF technique. The major (*Mj*) and minor (*Mn*) effects of the factors are designated as a numerical value of 1.0 and 0.5, respectively (Table 1). The selected weight for all factors has been calculated using the following equation:

TABLE 1: Effect of the impacting factor, relative rates, and proposed weight using MIF technique (modified after [3], [39], [41])

Factor	Major effect (Mj)	Minor effect (Mn)	Proposed relative rates (Mj+Mn)	Proposed weight for each impacting factor
DEM	1+1	0.5	2.5	21
LULC	1+1	0.5	2.5	21
Soil Texture	1	0	1	8
Rainfall	1	0.5+0.5	2	17
Slope	1+1	0.5	2.5	21
Drainage Density	1	0.5	1.5	12
Total (∑)			12	100

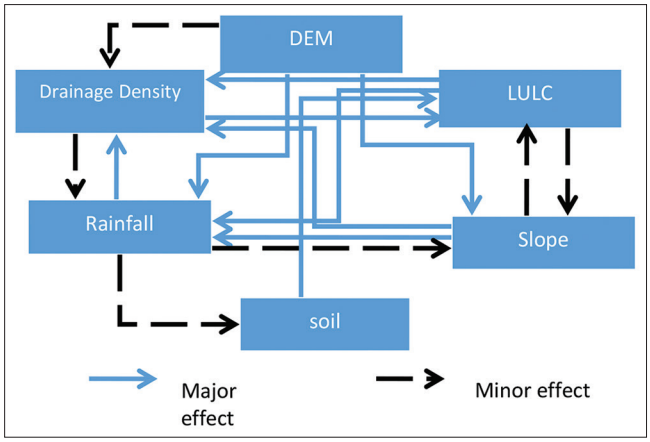

Fig. 8. Mutual relations between all the factors influencing the groundwater potential zone.

$$Proposed\ weight = \frac{(Mj + Mn)}{(\sum Mj + Mn)} X100 \quad (2)$$

3.8. Layers Overlay

The Weighted Overlay tool in Arc GIS was used for creating the potential groundwater zones in Sulaimani from the generated raster maps. The characteristics of each of the raster maps used in this study were gave weight age of 1–5, depending on the layers effect on the occurrence of groundwater. The weight age factor 1 indicates low groundwater potentiality and 5 indicates high groundwater potentiality. The classifications of weighted factors impacting the potential zones are shown in Table 2.

4. RESULTS AND DISCUSSION

Based on the assignment of all layers influence and weight of the individual features of the thematic layers, a potential groundwater zone map was produced as shown in Fig. 9. According to the results, about 79% (14423.45 km^2) of the study area can be classified as moderate groundwater recharge

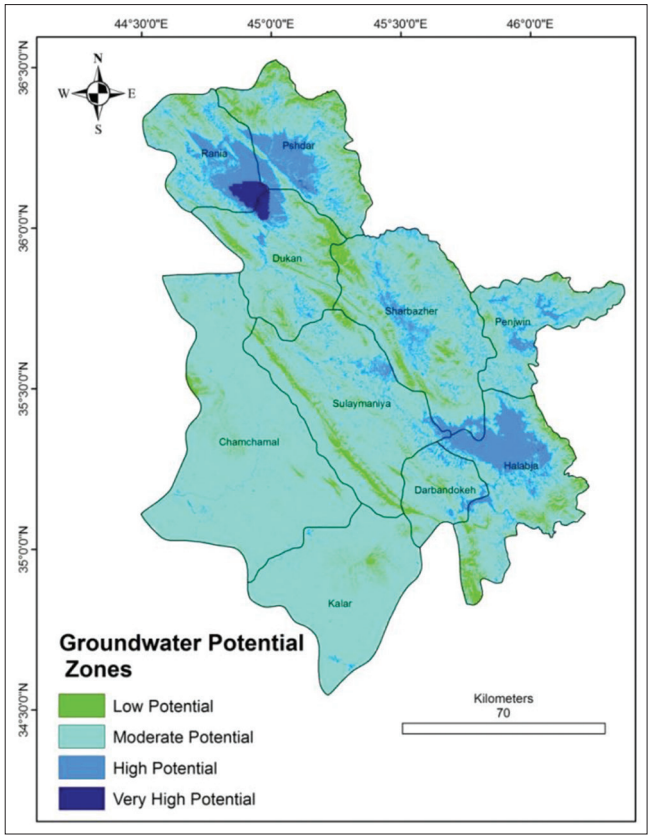

Fig. 9. Groundwater potential zones.

zone. About 1% (112.3746 km^2) has very high groundwater potential zone and about 14% (2576.78 km^2) has high groundwater potential zone, while only 6% (1168.015 km^2) of the total study area is of low groundwater potentials. Based on Fig. 8, the high groundwater potential zones were covers about 50% of Halabja, Rania, Pshdar districts.

The results showed that the higher the elevations the less the groundwater potential, and in lower elevations the groundwater is more as the water accumulates in low lands. For slope arrangement, it shows that areas with steep slopes

TABLE 2: Classification of weight age influencing the potential zones

Factor	Influence% (proposed weight)	Factotr of effect	Weight age
DEM (m)	21	182–657	5
		658–959	4
		960–1300	3
		1310–1770	2
		1780–3430	1
Land use	21	Water	5
		Forest	2
		Crop	4
		Urban area	3
		Bare soil	1
Soil texture	8	Sandy clay loam	1
		Clay	2
		Loam	4
		Sandy Loam	5
		Sandy clay	3
Rainfall (in mm)	17	260–430	1
		440–550	2
		560–660	3
		670–810	4
		820–1000	5
Slope (in degree)	21	0–6.4	5
		6.5–14	4
		15–24	3
		25–35	2
		36–78	1
Drainage density (in km/km^2)	12	0.21–0.94	5
		0.95–1.3	4
		1.4–1.5	3
		1.6–1.8	2
		1.9–2.9	1

have less groundwater than with mild slope. Rainfall is also affects the potential groundwater zones; the areas that have more rainfall have more groundwater than those with less rainfall. Different LU/LC leave distinctive signatures on groundwater recharge, the areas of water body in the study area produced high groundwater recharge zone.

5. CONCLUSIONS

In this study, GIS and MIF techniques have been used for modeling groundwater potential zones by integrating various factors that have been chosen based on the availability of data for Sulaimani Governorate. Traditional data, topographic maps, and satellite imageries were used to prepare the thematic layers of (DEM), soil texture, LU/LC, slope, rainfall, and drainage density. The resulted map of groundwater recharge zones for Sulaimani Governorate was classified into four zones that had very high potential zone, high potential zone, moderate zone potential, and low potential zone and cover (1%), (14%), (79%), and (6%) of the total area, respectively. About 50% of the high groundwater potential zone were located in Halabja, Rania, and Pshdar districts. The groundwater potential zonation presented here can be applied only for regional studies for the purpose of groundwater development, providing quick prospective guides for groundwater exploration and exploitation, while individual site selection for groundwater development should take into consideration other site-specific conventional ground-truthing methods.

REFERENCES

[1] V. P. Dinesan, G. Gopinatha and M. K. Ashitha. "*Application of Geoinformatics for the Delineation of Groundwater Prospects Zones-a Case Study for Melattur Grama Panchayat in Kerala, India*", 2015.

[2] J. Jani. GIS as a tool for modelling groundwater flow. In: "*2012 IEEE Symposium on Business, Engineering and Industrial Applications*", IEEE, United States, 2012.

[3] A. Ashok, R. Reghunath and J. Thomas. "*Mapping of Groundwater Recharge Potential Zones and Identification of Suitable Site-Specific Recharge Mechanisms in a Tropical River Basin, Earth Systems and Environment*", 2020.

[4] R. Gogu, G. Carabin, V. Hallet, V. Peters and A. Dassargues. "GIS-based hydrogeological databases and groundwater modelling". *Hydrogeology Journal*, vol. 9, no. 6, pp. 555-569, 2001.

[5] J. Ghayoumian, M. M. Saravi, S. Feiznia, B. Nouri and A. Malekian. "Application of GIS techniques to determine areas most suitable for artificial groundwater recharge in a coastal aquifer in southern Iran". *Journal of Asian Earth Sciences*, vol. 30, no. 2, pp. 364-374, 2007.

[6] M. I. Adham, C. S. Jahan, Q. H. Mazumder, M. A. Hossain and A. M. Haque. "Study on groundwater recharge potentiality of Barind tract, Rajshahi district, Bangladesh using GIS and remote sensing technique". *Journal of the Geological Society of India*, vol. 75, no. 2, pp. 432-438, 2010.

[7] V. Singhal and R.Goyal. "GIS based methodology for groundwater flow estimation across the boundary of the study area in groundwater flow modeling". *Journal of Water Resource and Protection*, vol. 3, no. 11, p. 824, 2011.

[8] N. S. Magesh, N. Chandrasekar and J. P. Soundranayagam. "*Delineation of Groundwater Potential Zones in Theni District, Tamil Nadu, Using Remote Sensing, GIS and MIF Techniques*", 2012.

[9] S. Kaliraj, N. Chandrasekar and N. S. Magesh. "Identification of potential groundwater recharge zones in Vaigai upper basin, Tamil Nadu, using GIS-based analytical hierarchical process (AHP) technique". *Arabian Journal of Geosciences*, vol. 7, no. 4, pp. 1385-1401, 2014.

[10] P. K. Ghosh, S. Bandyopadhyay and N. C. Jana. "*Mapping of Groundwater Potential Zones in Hard Rock Terrain Using Geoinformatics: A Case of Kumari Watershed in Western Part of West Bengal*", 2016.

[11] S. Kumar, B. K. Bhadra and R. Paliwal. "Evaluating the impact of artificial groundwater recharge structures using geo-spatial techniques in the hard-rock terrain of Rajasthan, India". *Environmental Earth Sciences*, vol. 76, no. 17, p. 613, 2017.

[12] A. G. Selvarani, G. Maheswaran and K. Elangovan. "Identification of artificial recharge sites for Noyyal river basin using GIS and remote sensing". *Journal of the Indian Society of Remote Sensing*, vol. 45, no. 1, pp. 67-77, 2017.

[13] H. A. Karim and D. A. Al-Manmi. "*Integrating GIS-Based and Geophysical Techniques for Groundwater Potential Assessment in Halabja Said Sadiq Sub-Basin, Kurdistan, NE Iraq*", 2019.

[14] C. Serele, A. Perez-Hoyos and F. Kayitakire. "*Mapping of Groundwater Potential Zones in the Drought-Prone Areas of South Madagascar Using Geospatial Techniques*", 2019.

[15] T. Dar, N. Rai and A. Bhat. "*Delineation of Potential Groundwater Recharge Zones Using Analytical Hierarchy Process (AHP), Geology, Ecology, and Landscapes*", 2020.

[16] S. Arya, T. Subramani and K. Duraisamy. "Delineation of groundwater potential zones and recommendation of artificial recharge structures for augmentation of groundwater resources in Vattamalaikarai Basin, South India". *Environmental Earth Sciences*, vol. 79, p. 102, 2020.

[17] M. O. Al-Djazouli, K. Elmorabiti, A. Rahimi, O. Amellah and O. A. Mohammed. "Delineating of groundwater potential zones based on remote sensing, GIS and analytical hierarchical process: A case of Waddai, eastern Chad". *GeoJournal*, vol. 246, p. 5, 2020.

[18] D. Rawal and A. Vyas. "Application in GIS and groundwater modeling techniques to identify the perched aquifers to demarkate water logging conditions in parts of MEHASAN". *ISPRS annals of photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 3, no. 8, pp. 173-180, 2016.

[19] S. K. Singh, M. Zeddies, U. Shankar and G. A. Griffiths. "*Potential Groundwater Recharge Zones within New Zealand*", 2018.

[20] P. Arulbalaji, D. Padmalal and K. Sreelash. "GIS and AHP techniques based delineation of groundwater potential zones: A case study from Southern Western Ghats, India". Scientific Reports, vol. 9. p. 2082, 2019.

[21] A. M. Rasheed. "Analysis of rainfall drought periods in the North of Iraq". *Al-Rafidain Engineering*, vol. 18, no. 2, pp. 60-72, 2010.

[22] N. F. Mustafa, H. M. Rashid and H. M. Ibrahim. "Aridity index based on temperature and rainfall data for kurdistan Region-Iraq". *Journal of University of Duhok*, vol. 21, no. 1, pp. 65-80, 2018.

[23] "*United States Geological Survey Online Database*", 2019. Available from: https://www.earthexplorer.usgs.gov. [Last accessed on 2019 Sep 20]

[24] S. Zakaria, N. Al-Ansari, Y. T. Mustafa, S. Knutsson, P. S. Ahmad and B. D. Ghafour. "Rainwater harvesting at koysinjaq (Koya), Kurdistan region, Iraq". *Journal of Earth Sciences and Geotechnical Engineering*, vol. 3, no. 4, pp. 25-46, 2013.

[25] M. L. Collin and A. J. Melloul. "*Combined Land-Use and Environmental Factors for Sustainable Groundwater Management*". *Urban Water*, vol. 3, no. 3, pp. 229-237, 2001.

[26] D. N. Lerner and B. Harris. "The relationship between land use and groundwater resources and quality". *Land Use Policy*, vol. 26, pp. S265-S273, 2009.

[27] H. F. Yeh, Y. S. Cheng, H. I. Lin and C. H. Lee. "Mapping groundwater recharge potential zone using a GIS approach in Hualian River, Taiwan". *Sustainable Environment Research*, vol. 26, pp. 33-43, 2016.

[28] K. Ibrahim-Bathis and S. A. Ahmed. "Geospatial technology for delineating groundwater potential zones in Doddahalla watershed of Chitradurga district, India". *The Egyptian Journal of Remote Sensing and Space Science*, vol. 19, pp. 223-234, 2016.

[29] S. L. Martin, D. B. Hayes, A. D. Kendall, and D. W. Hyndman. "The land-use legacy effect: Towards a mechanistic understanding of time-lagged water quality responses to land use/cover". Science of the Total Environment, vol. 579, pp. 1794-1803, 2017.

[30] R. W. Healy. "*Estimating Groundwater Recharge*". Cambridge University Press, Cambridge. 2010.

[31] H. Bouwer. "Artificial recharge of groundwater: Hydrogeology and engineering". *Hydrogeology Journal*, vol. 10, no. 1, pp. 121-142, 2002.

[32] "*Harmonized World Soil Database, Version 1.2*", 2012. Available from: http://www.fao.org/fileadmin/templates/nr/documents/hwsd/hwsd_documentation.pdf. [Last accessed on 2019 Sep 20]

[33] Food and Agriculture Organization of the United Nations (FAO). "*FAO-Geonetwork, Online Database. Digital Soil Map of the World*", 2007. Available from: http://www.fao.org/geonetwork/srv/en/metadata.show?id=14116. [Last accessed on 2019 Oct 18]

[34] K. E. Keese, B. R. Scanlon and R. C. Reedy. "Assessing controls on diffuse groundwater recharge using unsaturated flow modelling". *Water Resources Research*, vol. 41, no. 6, p. W06010, 2005.

[35] C. Mohan, A. W. Western, Y. Wei and M. Saft. "Predicting groundwater recharge for varying land cover and climate conditions-a global meta-study". *Hydrology and Earth System Sciences*, vol. 22, no. 5, pp. 2689-2703, 2018.

[36] G. D. Fontana and L. Marchi. "Slope-area relationships and sediment dynamics in two alpine streams". *Hydrological Processes*, vol. 17, no. 1, pp. 73-87, 2003.

[37] V. M. Rokade, P. Kundal and A. K. Joshi. "Groundwater potential modeling through remote sensing and GIS: A case study of Rajura Taluka, Chandrapur District, Maharashtra". *Journal of the Geological Society of India*, vol. 69, no. 5, pp. 943-948, 2007.

[38] J. M. Detty and K. J. McGuire. "Topographic controls on shallow groundwater dynamics: Implications of hydrologic connectivity between hillslopes and riparian zones in a till mantled catchment". *Hydrological Processes*, vol. 24, no. 16, pp.2222-2236, 2010.

[39] S. Selvam, N. S. Magesh, P. Sivasubramanian, J. P. Soundranayagam, G. Manimaran and T. Seshunarayana. "Deciphering of groundwater potential zones in Tuticorin, Tamil Nadu, using remote sensing and GIS techniques". *Journal of the Geological Society of India*, vol. 84, pp. 597-608, 2014.

[40] D. Greenbaum. "*Review of Remote Sensing Applications to Groundwater Exploration in Basement and Regolith*". British Geological Survey, Nottingham, United Kingdom. p. 63, 1985.

[41] S. Lakhwinder. "*Groundwater Potential Zones*, *online course*", 2018. Available from: https://www.udemy.com/course/groundwater-potential-zones-using-gis-full-project-arcgis-tutorial/learn/lecture/12788101#overview. [Last accessed on 2019 Oct 15]

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Offline Writer Recognition for Kurdish Handwritten Text Document Based on Proposed Codebook

Twana Latif Mohammed[1]*, Ahmed Abdullah Ahmed[2]

[1]Department of Information Technology, Technical College of Informatics, Sulaimani Polytechnic University, Sulaymaniyah, Kurdistan Region, Iraq, [2]Department of Software Engineering, Faculty of Engineering and Computer Science, Qaiwan International University (QIU)/Raparin, Sulaymaniyah, Kurdistan Region, Iraq

ABSTRACT

Handwritten text recognition has been an ongoing attractive task to research in the field of document analysis and recognition with applications in handwriting forensics, paleography, document examination, and handwriting recognition. In the present research, an automatic method of writer recognition is presented using digitized images of unconstrained texts. Despite the increasing efforts by prior literature on the different methods used for the same purpose, such methods performance, particularly their accuracy, has not been promising, leaving plenty of room for improvements. This method made use of codebook-based writer characterization, with each writing sample represented by a group of computed features from a primary and secondary codebook. The writings were then represented through the computation of the probability of codebook patterns occurrence, and the probability distribution was employed for each writer's characterization. Writer identification process involved comparing two writings through the computation of the distances between their respective probability distribution. The study carried out experiments to determine the performance of the implemented method in light of rates of identification with the help of standard datasets, namely, KRDOH and IAM, the former being the most current and largest Kurdish handwritten datasets with 1076 writers, and the latter being a dataset containing 650 writers. The outcome of the experiments was promising with a rate of identification of 94.3%, with the proposed method outperforming the state-of-the-art methods by 2–3%.

Index Terms: Writer Identification, Feature Extraction, Text Independent, Codebooks, Feature Combination

1. INTRODUCTION

An individual is distinguished from another through the distinct identities that he possesses. Such identities may be physical or behavioral and they can be employed to identify individuals in a scientific field called biometrics. Biometrics plays a key role in researches dedicated to forensic science, where forensic experts make use of physical or behavioral biometrics to recognize and identify individuals. Physical biometrics includes DNA as illustrated in the study by Holland and Parsons [1], fingerprints as presented by Abu-Faraj *et al.* [2], ear prints [3], [4], irises [5], and soft and hard tissues as illustrated in the study by Zewail *et al.* [6]. Examples of behavioral biometrics are speech [7], [8] and gait [9], [10]. This also includes keystroke intervals as in Delac and Grgic [11], and signatures, and handwriting. In this thesis, the researcher focuses and examines handwriting.

Whether handwriting is alphabetical or pictographic based, it has been employed as a significant communication means

Access this article online

DOI: 10.21928/uhdjst.v5n1y2021.pp21-27

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Twana Latif Mohammed, Technical College of Informatics, Sulaimani Polytechnic University, Sulaymaniyah, Kurdistan Region, Iraq. E-mail: twana.latif@spu.edu.iq

Received: 19-10-2020 Accepted: 27-03-2021 Published: 31-03-2021

from the beginning of time and has made certain evolutions. According to Huber and Headrick [12], writing styles are developed based on local culture, geographical location, historical background, and temporal situations. The earlier handwriting record keeping came from China, dated around 2000 years ago, following the invention of the first inks and papers. During that time, early handwritings and the following handwriting were written in some standard writing models. In general, writers have a tendency not to follow standard writing models, and thus, handwritings show deviation. Individual writer characteristics are invaluable in distinguishing one writer from another.

2. LITERATURE REVIEW

This section presents a comprehensive review of the techniques developed for writer identification on offline writing samples, and this is one of the main topics addressed in the research. The section presents an extensive review of offline handwritten datasets, outlines the existing methods on text-dependent writer identification methods while the subsection discusses the significant contributions to the text-independent writer recognition domain which is also the primary focus of this research. Finally, the last part of this section conducts a comparative analysis among the methods with detailed presentations of their performances.

A recent study [13] uses a simplified and rapid methodology by avoiding character appearances and making a distinction from approaches similar to those in traditional codebook-based methods. Here, researchers present new descriptors that are derived from different scales of geometrical interest points. This is achieved by documenting the geometric associations between parts of the script, such as strokes, loops, endings, and junctions. These descriptors are easier to use, more effective, provide better results on unseen datasets, and reduce processing time dramatically over existing methods. In addition to these benefits, this method has a drawback in terms of the amount of data needed to build a model with consistent results.

Some researchers in Al-Maadeed *et al.* [14] showed the identification of various writers using their proposed collection of curvature, direction, and tortuosity-based geometrical features. They also suggested improvement of edge-based directional features using a filled moving window instead of an edge moving window alongside chain code-based features using a fourth-order chain code list for enhancing its recognizing ability. This method was tested in the handwriting databases of IAM and QUWI. In addition, the authors [15] proposed a new method called (DLS_CNN) for writer recognition so, in this research used the combination between neural network (NN) with line segmentation. On the other hand in Chahi *et al.* [16], the authors proposed a new algorithm called LSTP but at the classification step base on NN achieved Hamming distance.

While great interest and substantial progress have been observed in the field of handwriting based biometrics and its applications [17], the identification of writers is based on relatively complex scripts, that is, Arabic [18] and Chinese [19] which remains a less investigated area [20], rare comparable between each of the scripts among researchers have resulted in vague results which are not proportional to its widespread use. Most of the researches in this area share the same purpose of determining a script's authorship through the acquisition of individual handwriting characteristics. In the current process, all document features are found and created, after which the feature vector distance is compared between the query and the library image. Nevertheless, this performance is considered far from being achieved, and it is computationally very costly, particularly a significant problem in document image analysis and retrieval is the search for the relevant document from large and complex document image repositories. Apart from issues of database size (scale), there is a problem of data heterogeneity. The smooth incorporation of such techniques with the current knowledge in forensic handwriting is still unknown. Such approaches are not obsolete and are still used today. Some researchers in the same field have used the principal component analysis method to extract the most important information. The data performed as testing and training on the grayscale's images like 12,500 images [21].

Ghiasi and Safabakhsh [22] generated feature vectors for each manuscript by taking advantage of the normalized and resampled contours of connected segments. Then, for writer recognition, they used these feature vectors to form a codebook. They also solve cursive handwriting, the method utilizes the occurrence histogram of the shapes in a codebook, connected complements can be too long and may have a wide range of shapes. To prevent complex patterns, the authors used small fragments of the connected components and implemented two effective methods for extracting code from contours. One of the techniques uses the actual pixel coordinates of contoured fragments, while the other utilizes linear piecewise approximation using segment angles and lengths and removes some of the unnecessary information. It helps to identify and group similar shapes. The shorter length of this code allows it to be applied faster and also helps the

quicker generation of the codebook. The authors tested this code on two English databases and one Persian and found better performance than other contemporary techniques in 2013. It may be quicker to generate codebooks, but the computational times of this technique are long.

3. METHODOLOGY

In this study, the author brings forward a methodology characterizing the writer through the division of the text into smaller fragments, and potential clusters are searched within it. This is based on premise that writer recognition is related to the physical stroke's generation by the writer. Rather than dividing the writing in graphemes, the method divides them into small fragments that are enough to be utilized in writer recognition. Further detailed elaboration of the modules is provided in the next sections.

3.1. Binarization

Under this process, the digitized document images are scanned in the form of grayscale images. The research addresses handwritten scanned documents, and as such, there are two objects to the image, namely, handwriting and background, and in this research, the primary object of interest is handwriting. Therefore, handwriting is separated from the background through the use of binarization that is categorized into two classes. The study employs the popular Otsu's thresholding logarithm (known as bae benchmark) for the calculation of the threshold from the grayscale image. A grayscale original image along with its binarized version is demonstrated in Fig. 1.

3.2. Componentization of Writing

Before handwriting fragmentation, the applied method entails the division of handwriting into related components in what is known as componentization. This forms clusters of the entire related black pixels based on their connectivity. In an individual connected component, the pixels adjacencies are gauged through the use of 8-side connectivity, after which they are labeled with sequential numbers. For every pixel with the same label, a component obtained from the image is highlighted for their fragmentation. The connecting image components are depicted in Fig. 2.

3.3. Fragmentation of Components

This study brings forward a writer characterization method through a specific sample by examining small invariant fragments and exploiting the writing redundancy. The step entailing the division of handwriting into fragments is a significant one in the applied method, and it considers them along with the adjacent fragments connected to them. More specifically, the adjacent fragments are acquired through the writing division into windows after which, the main and adjacent fragments are categorized into individual codebooks for the ultimate writer characterization method.

3.4. Feature Extraction

It is the comparison among the fragments through pattern matching or by representing them with a set of features. Although pattern matching is a simple process, it calls for maintaining the fragment's pixel values, otherwise, the comparison outcome may lose its robustness to both noise and distortions. A comparison of features mitigates the representation space of the dimension but it is susceptible to distortions. Thus, the applied method represents each writing fragments (main and adjacent) using a set of features including vertical and horizontal projections, upper and lower profiles along with a group of familiar shape descriptors (i.e., elongation, solidity, rectangularity, orientation, and perimeter).

3.4.1. Horizontal and vertical projections

Projections provide the number of black pixels present in the fragmented image, within each row and column. More specifically, the horizontal projection is produced by determining the number of black pixels in every column of the image, while the vertical projection is produced by determining the number of black pixels in every role of the same.

3.4.2. Upper and lower profile

For the upper and lower profile, the former is described as the distance of the first black pixel from the top of every fragmented image, while the latter is the distance of the first black pixel from the bottom of every fragment. Both upper and lower fragment profiles are calculated by determining the column of the fragment and the distance between the upper black pixels to the lower one.

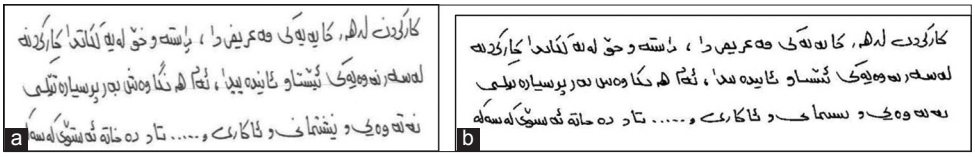

Fig. 1. Image binarization. (a) Grayscale handwriting image before binarization, (b) image after binarization.

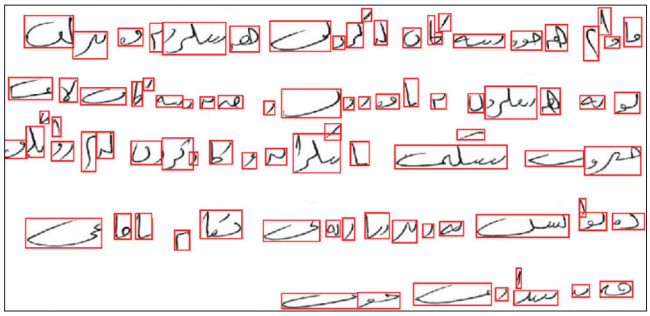

Fig. 2. Bounding box of connected components.

3.4.3. Orientation

The direction of a stroke (or its slope) in a fragmented image is calculated through its orientation feature, specifically by the angle between the X-axis and the major axis of an ellipse that approximates the fragment. It is evident from Fig. 3. 18b that an ellipse comprises a group of points that move around the black pixels of the fragmented stroke, whose sum constitutes the distance from two fixed points, namely, F1 and F2, and it remains constant.

3.4.4. Rectangularity

This feature refers to the ratio of the object area to the bounding box area, the latter of which is the smallest rectangle encapsulating the writing shape in a fragment. Rectangularity is mathematically defined as follows:

$$Rec\tan gularity = \frac{A_{FW}}{A_{BB}} \quad (1)$$

In the above equation, A_{FW} denotes the number of pixels in the fragmented window area, while A_{BB} denotes the bounding box area containing the stroke region.

3.4.5. Elongation

This feature refers to the ratio between the bounding box height and its width. A bounding box was obtained from a fragment enclosing a stroke. The stroke elongation is mathematically represented by the following equation:

$$Elongation = \frac{l_b}{s_b} \quad (2)$$

From the above equation, l_b represents the bounding box longer side and the S_b represents the bounding box shorter side.

3.4.6. Perimeter R

This feature represents the shape boundary's total length. More specifically, the boundary of the shape comprises a group of pixels in the boundary having a non-shape pixel as an adjacent pixel. Mathematically, the perimeter can be calculated by tracking the stroke's boundary pixel after which the steps are summed up.

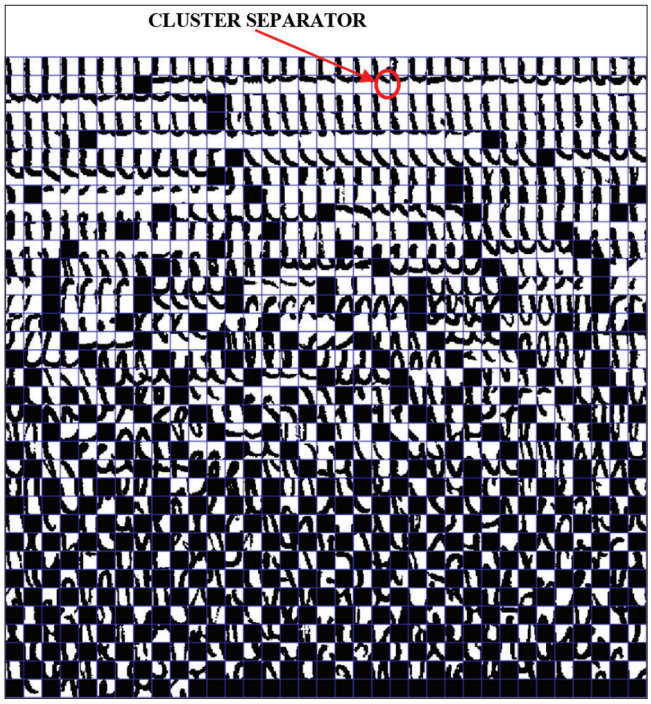

Fig. 3. Primary codebook obtained from the main fragmented windows on a writing sample (Sample: W0010Para2).

3.4.7. Solidity

This feature is useful in measuring the fragment's density and is calculated as the ratio between the fragment areas and is corresponding to convex. The solidity value ranges from 0 to 1 and a solidity value that is near to 0 depicts an irregular object, while that is near to 1 is a solid one. Solidity can be mathematically represented as:

$$Solidity = \frac{A_{FW}}{A_{CR}} \quad (3)$$

In the above equation, A_{FW} denotes the fragment area, while A_{CR} denotes the convex region area.

3.5. Clustering of Fragments

The present section proceeds to present the grouping of similar fragments, extracted through the use of main and adjacent windows, into clusters referred to as codebook. The features are used to make clusters, in that closely related fragmented patterns are clustered together to make a class. In each class, patterns are distinct from those in other classes, and in each cluster, every individual class contains a group of invariant writing patterns. The implemented method produces two distinct cluster sets, by matching the

invariant patterns features. The entire main strokes extracted through the use of main windows are clustered to develop a primary cluster, while the entire adjacent windows fragments are clustered to develop a secondary cluster. Figs. 3 and 4 illustrate the primary and secondary codebooks generated from the main and adjacent fragmented windows.

4. RESULTS AND DISCUSSION

This section discusses the experimental evaluation of the applied technique. Many experiments were performed to evaluate the performance of the system and study the sensitivity of the performance to different parameters. Since two codebooks, primary and secondary, have been presented, writer recognition results on each of these codebooks are presented. All experiments are conducted on the latest, largest, and standard Kurdish handwritten documents database known as KRDOH [23]. Moreover, the implemented method is also benchmark against the best and up-to-date methods found in the literature of writer identification that has used IAM [24] data set. Sample forms from the database are shown in Fig. 5.

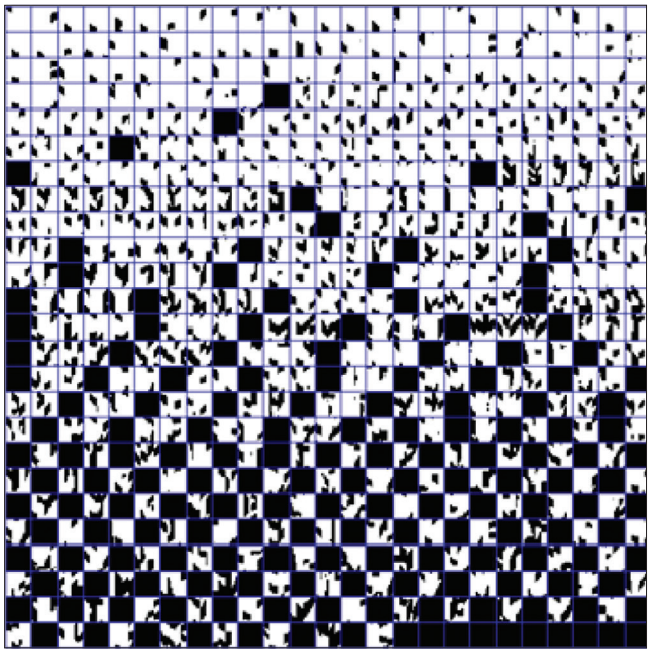

Fig. 4. Secondary codebook obtained from the adjacent fragmented windows on a writing sample (Sample: W0010Para2).

This study first evaluates the performance of primary and secondary codebook separately and then merges both codebooks. Initial experiments were conducted on 210 random writers from the KRDOH dataset. Table 1 summarizes the results of a primary codebook, secondary codebook, and merged codebooks. Using the primary codebook, an identification rate of 87.14% (Top-5: 91.03% and Top-10: 94.08%) is achieved with an EER of 5.92%. The secondary codebook achieves slightly better identification rate of 89.26% (Top-5: 92.14% and Top-10: 96.17%) with 3.83% EER. By merging the two codebooks, the overall identification rate is increased to 91.87% (Top-5: 93.3% and Top-10: 97.6%) and EER drops to 2.4%.

Later, Table 2 provides a performance comparison of the latest writer identification techniques. Oriented basic image features and the concept of graphemes codebook were employed by Durou *et al.*, 2019 [25], achieved 92% identification rate on the IAM dataset. Later (Nguyen *et al.*, 2019) [26], the author used a CNN-based method for text-independent writer identification on the

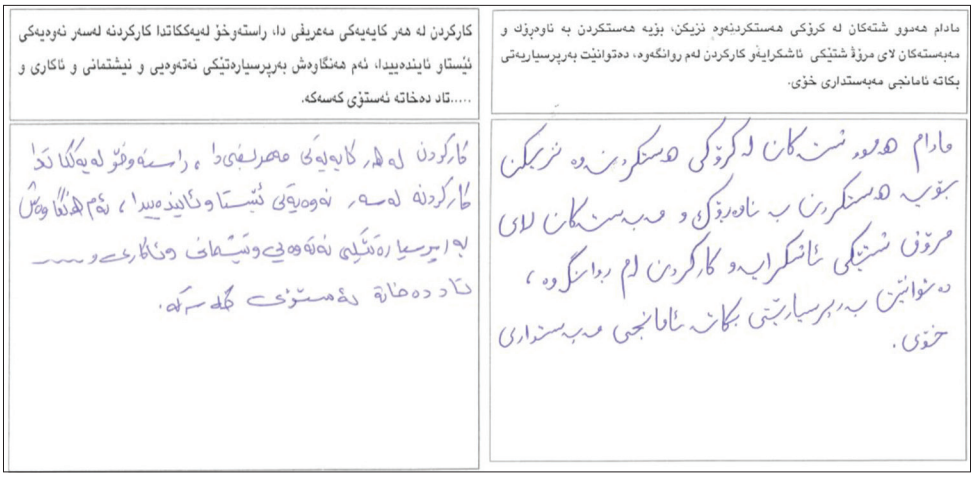

Fig. 5. Examples of the scanned forms of KRDOH dataset.

TABLE 1: Applied method results on 210 writers from KRDOH dataset

Mission	Identification			Verification
Codebook	Top 1 (%)	Top 5 (%)	Top 10 (%)	EER (%)
Primary	87.14	91.03	94.08	5.92
Secondary	89.26	92.14	96.17	3.83
Merge	91.87	93.3	97.6	2.4

TABLE 2: Performance comparison of writer identification methods

Authors	Year	Dataset	Writers	Performance (%)
Durou *et al*.	2019	IAM	650	92
Nguyen *et al*.	2019	IAM	650	91.81
Proposed method	2020	IAM	650	94.37

same database of offline handwritten English text and achieved 90.12% identification rate. Using the same 650 sets of writers of the IAM dataset, the proposed study achieved an identification rate of 94.37% which is the best identification rate on this dataset so far using the writer-specific codebook technique.

5. CONCLUSION

This research primary aims to apply and test an automatic writer recognition method on offline Kurdish handwritten text. Such objective was achieved through the implementation of a new method addressing the issues of state-of-the-art methods and outperforming them. The approach involved the extraction of small writing fragments through the positioning of windows over the writing and clustering writing fragments that are similar, forming a codebook. In contrast to classical methods that produce a codebook of graphemes, the codebook of small writing fragments is script independent and is applicable to text of different languages. Moreover, the applied method extracts main strokes along with the adjacent strokes linked to the former and both fragments (main and adjacent) are separately clustered to generate the primary and the secondary codebooks. Following the generation of the codebooks, each writing sample is represented as a probability patterns distribution in them, after which two writings are compared by calculating the distance between their respective codebooks. Standard datasets (IAM and KRDOH) using both the codebooks (primary and secondary) and their integration reflect the optimal performance of the approach over the existing approaches.

REFERENCES

[1] M. M. Holland and T. J. Parsons. "Mitochondrial DNA sequence analysis validation and use for forensic casework". *Forensic Sci. Rev.*, vol. 11, no. 1, pp. 21-50, 1999.

[2] Z. Abu-faraj, D. P. A. Atie, K. Chebaklo, S. Member and Z. E. Khoukaz. "*Fingerprint Identification Software for Forensic Applications*". Electronics, Circuits and Systems, 2000. ICECS 2000. The 7th IEEE International Conferenceno. May, 2010.

[3] S. Black and T. J. U. Thompson. "*Body Modification*". CRC Press, Boca Raton, 2007.

[4] W. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfeld. "Face recognition: A literature survey". *ACM Computing Surveys*, vol. 35, no. 4, pp. 399-458, 2003.

[5] J. Daugman. "How iris recognition works". *IEEE Journal*, vol. 14, no. 1, pp. 21-30, 2004.

[6] R. Zewail, A. Elsafi, M. Saeb and N. Hamdy. "*Soft and Hard Biometrics Fusion for Improved Identity Verification*". The 2004 47th Midwest Symposium on Circuits and Systems, pp. 225-228, 2004.

[7] C. Champod and D. Meuwly. "The inference of identity in forensic speaker recognition". *Speech Communication*, vol. 31, pp. 193-203, 2000.

[8] G. R. Joaquin and D. Ramos. Forensic automatic speaker classification in the coming paradigm shift. *In: Speaker Classification I. Springer, Berlin, Heidelberg*, pp. 205-217, 2007.

[9] J. K. Aggarwal and Q. Cai. "Human motion analysis: A review". *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428-440, 1999.

[10] M. G. Grant, J. D. Shutler, M. S. Nixon and J. N. Carter. "*Analysis of a Human Extraction System for Deploying Gait Biometrics*". 6th IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 46-50, 2004.

[11] K. Delac and M. Grgic. "*A Survey of Biometric Recognition Methods*". 46th International Symposium Electronics in Marineno, pp. 16-18, 2004.

[12] R. A. Huber and A. M. Headrick. "*Handwriting identification: Facts and fundamentals*". CRC Press, Boca Raton, Florida, 1999.

[13] A. Garz, M. Würsch and A. Fischer. "*Simple and Fast Geometrical Descriptors for Writer Identification*". Society for Imaging Science and Technology, Springfield, Virginia, pp. 1-12, 2016.

[14] S. Al-Maadeed, A. Hassaine, A. Bouridane and M. A. Tahir. "Novel geometric features for off-line writer identification". *Pattern Analysis and Applications*, vol. 19, no. 3, pp. 699-708, 2016.

[15] C. Shi-Ming and W. Yi-Song. "A robust off-line writer identification method". *Renhe Test*, vol. 46, no. 1, pp. 108-116, 2020.

[16] A. Chahi, Y. Ruichek and R. Touahni. "Local gradient full-scale transform patterns based off-line text-independent writer identification". *Applied Soft Computing*, vol. 2020, p. 106277, 2020.

[17] A. Forn, D. Albert and G. Josep. "CVC-MUSCIMA: A ground-truth of handwritten music score images for writer identification and staff removal". *International Journal on Document Analysis and Recognition*, vol. 15, pp. 243-251, 2012.

[18] A. A. Ahmed and G. Sulong. "Arabic writer identification: A review of literature". *Journal of Theoretical and Applied Information Technology*, vol. 69, no. 3, pp. 474-484.

[19] G. J. T. Rahim and M. S. M. Rahim. "Off-line text-independent writer recognition for chinese handwriting: A review". *Jurnal Teknologi*, vol. 2, pp. 39-50, 2015.

[20] S. M. Awaida and S. A. Mahmoud. "State of the art in off-line writer identification of handwritten text and survey of writer identification of Arabic text". *Educational Research Review*, vol. 7, no. 20, pp. 445-463, 2012.

[21] A. Junaidi, S. Trianingsih and M. Iqbal. "Writer identification of lampung handwritten documents based on selected characters". *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika*, vol. 6, no. 1, pp. 1-8.

[22] G. Ghiasi and R. Safabakhsh. "Offline text-independent writer identification using codebook and efficient code extraction methods". *Image and Vision Computing*, vol. 31, no. 5, pp. 379-391, 2013.

[23] T. L. Mohammed, A. A. Ahmed and O. I. Al-Sanjary. "*KRDOH: Kurdish Offline Handwritten Text Database*". In: 2019 IEEE 7th Conference on Systems, Process and Control (ICSPC), pp. 86-89, 2019.

[24] U. V. Marti and H. Bunke. "The IAM-database: An English sentence database for offline handwriting recognition". *International Journal on Document Analysis and Recognition*, vol. 5, no. 1, pp. 39-46, 2002.

[25] A. Durou, I. Aref, S. Al-Maadeed, A. Bouridane and E. Benkhelifa. Writer identification approach based on bag of words with OBI features". *Information Processing and Management*, vol. 56, no. 2, pp. 354-366, 2019.

[26] H. T. Nguyen, C. T. Nguyen, T. Ino, B. Indurkhya and M. Nakagawa. "Text-independent writer identification using convolutional neural network". *Pattern Recognition Letters*, vol. 121, pp. 104-112, 2019.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

An Efficient Two-layer based Technique for Content-based Image Retrieval

Fawzi Abdul Azeez Salih[1], Alan Anwer Abdulla[2,3]*

[1]Department of Computer Science, College of Science, University of Sulaimani, 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

The rapid advancement and exponential evolution in the multimedia applications raised the attentional research on content-based image retrieval (CBIR). The technique has a significant role for searching and finding similar images to the query image through extracting the visual features. In this paper, an approach of two layers of search has been developed which is known as two-layer based CBIR. The first layer is concerned with comparing the query image to all images in the dataset depending on extracting the local feature using bag of features (BoF) mechanism which leads to retrieve certain most similar images to the query image. In other words, first step aims to eliminate the most dissimilar images to the query image to reduce the range of search in the dataset of images. In the second layer, the query image is compared to the images obtained in the first layer based on extracting the (texture and color)-based features. The Discrete Wavelet Transform (DWT) and Local Binary Pattern (LBP) were used as texture features. However, for the color features, three different color spaces were used, namely RGB, HSV, and YCbCr. The color spaces are utilized by calculating the mean and entropy for each channel separately. Corel-1K was used for evaluating the proposed approach. The experimental results prove the superior performance of the proposed concept of two-layer over the current state-of-the-art techniques in terms of precision rate in which achieved 82.15% and 77.27% for the top-10 and top-20, respectively.

Index Terms: CBIR, Feature Extraction, Color Descriptor, DWT, LBP

1. INTRODUCTION

Beside content-based image retrieval (CBIR), digital image processing plays a vital role in numerous areas such as processing and analyzing medical image [1], image inpainting [2], pattern recognition [3], biometrics [4], multimedia security [5], and information hiding [6]. In the area of image processing and computer vision, CBIR has grown increasingly as an advanced research topic. CBIR refers to the system which retrieves similar images of a query image from the dataset of images without any help of caption and/or description of the images [7]. There are two main mechanisms for image retrieval which are text-based image retrieval (TBIR) and CBIR [8]. TBIR was first introduced in 1970 as search and retrieve images from image dataset [9]. In such a kind of image retrieval mechanism, the images are denoted by text and then the text is used to retrieve or search for the images. The TBIR method depends on the manual text search or keyword matching of the existing image keywords and the result has been relied on the human labeling of the images. TBIR approach requires information such as image keyword, image location, image tags, image name, and other information related to the image. Human involvement is needed in the challenging process of entering information of the images in the dataset. The drawbacks of TBIR are as follows: 1- It leads to inaccurate results if human has been

Access this article online

DOI:10.21928/uhdjst.v5n1y2021.pp28-40 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, Sulaimani, Iraq; Department of Information Technology, University College of Goizha, Sulaimani, Iraq. E-mail: alan.abdulla@univsul.edu.iq

Received: 07-01-2021 Accepted: 30-03-2021 Published: 05-04-2021

doing datasets annotation process incorrectly, 2- single keyword of image information is not effective to transfer the overall image description, and 3- it is based on manual annotation of the images, which is time consuming [10]. Researchers introduced CBIR as a new mechanism for image retrieval to overcome the above-mentioned limitations of TBIR. It is considered as a popular technique to retrieve, search, and browse images of query information from a broad dataset of images. In CBIR, the image information, visual features such as low-level features (color, texture, and/or shape), or bag of features (BoF) have been extracted from the images to find the most similar images in the dataset [11]. Fig. 1 illustrates the general block diagram of the CBIR mechanism [12].

Fig. 1 shows the block diagram of basic CBIR system that involves two phases: Feature extraction and feature matching. The first phase involves extracting the image features while the second phase involves matching these features [13]. Feature extraction is the process of extracting features from the dataset of images and stored in feature vector and also extracting features from the query image. On the other hand, feature matching is the process of comparing the extracted features from the query image to the extracted features from images in the dataset using similarity distance measurement. The corresponding image in the dataset is considered as a match/similar image to the query image if the distance between feature vector of the query image and the image in the dataset is small enough. Thus, the matched images are then ranked based on the similarity index from the smallest distance value to the largest one. Eventually, the retrieved images are selected according to the lowest distance value. The essential objective of the CBIR systems is improving the efficiency of the system by increasing the performance using the combination of features [9]. Image features can be categorized into two types: Global features and local features. Global features extract information from the entire image while local features work locally which are focused on the key points in images [14]. For the large image dataset, image relevant to the query image are very few. Therefore, the elimination of irrelevant images is important. The main contribution of this research is first eliminating the irrelevant images in the dataset and then finds the most similar/matches images from the rest of the remained images. The reminder of the paper is organized as follows: Section 2 discusses the related work. Section 3 introduces a background about the techniques used in the proposed approach. Section 4 presents the proposed CBIR approach. Section 5 shows the experimental results. Finally, section 6 gives the conclusions.

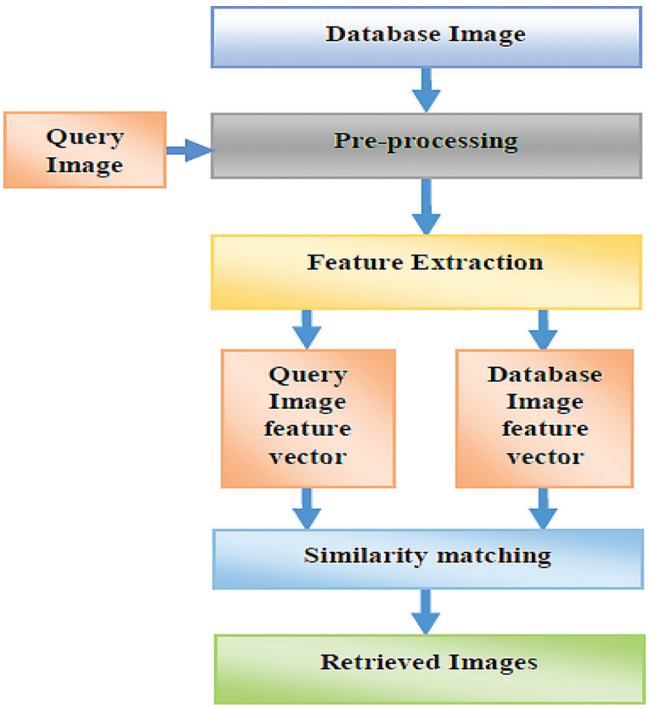

Fig. 1. General block diagram of CBIR mechanism.

2. RELATED WORK

Studies related to the developed techniques of CBIR have been researched a lot and they mainly focused on analyzing and investigating the interest points/areas such as corners, edges, contours, maxima shapes, ridges, and global features [15]. Some of those developed approaches are concerned on combining/fusing certain types of the extracted features, since such kind of strategy has an impact on describing the image content efficiently [13], [16]. This section reviews the most important and relevant existing works on CBIR. The main competition in this research area is increasing the precision rate that refers to the efficiency of retrieving the most similar images correctly. Kato *et al.* were first to investigate this field of study in 1992, who developed a technique for sketch retrieval, similarity retrieval, and sense retrieval to support visual interaction [17]. Sketch retrieval accepts the image data of sketches, similarity retrieval evaluates the similarity based on the personal view of each user, and sense retrieval evaluates based on the text data and the image data at content level based on the personal view. Yu *et al.*, in 2013, proposed an effective image retrieval system based on the (BoF) model, depending on two ways of integrating [18]. Scale-invariant feature transform (SIFT) and local binary pattern (LBP) descriptors were integrated in one hand, and histogram of oriented gradients (HOG) and LBP descriptors were integrated on the other hand. The first proposed integration, namely SIFT-LBP, provides better precision rate in which reached 65% for top-20 using Jaccard similarity measurement.

Shrivastava *et al.*, in 2014, introduced a new scheme for CBIR based on region of interest ROI codes and an effective feature set consisting of a dominant color and LBP were extracted from both query image and dataset of images. This technique achieved, using Euclidean distance measurement, a precision rate of 76.9% for top-20 [19]. DWT as a global feature, and gray level co-occurrence matrix (GLCM), as local feature, was extracted and fused in the algorithm introduced by Gupta *et al.*, in 2015, and as a result, a precision rate of 72.1 % was obtained for top-20 using Euclidean distance [20]. Another technique was introduced by Navabi *et al.*, in 2017, for CBIR which based on extracting color and texture features. The technique used color histogram and color moment as color feature. The principal component analysis (PCA) statistical method was applied for the dimension's reduction. Finally, Minkowski distance measurement was used to find most similar images. As reported, this technique achieved the precision rate of 62.4% for top-20 [21]. Nazir *et al.*, in 2018, proposed a new CBIR technique by fusing the extracted color and texture features [22]. Color Histogram (CH) was used to extract a color information, and DWT as well as edge histogram descriptor (EDH) were used to extract texture features. As authors claimed, this technique achieved a precision rate of 73.5% for top-20 using Manhattan distance measurement. Pradhan *et al.*, in 2019, developed a new CBIR scheme based on multi-level colored directional motif histogram (MLCDMH) [23]. This scheme extracts local structural features at three different levels. The image retrieval performance of this proposed scheme has been evaluated using different Corel/natural, object, texture, and heterogeneous image datasets. For the Corel-1k, the precision rate of 64% and 59% was obtained for top-10 and top-20, respectively. Recently, Sadique *et al.*, in 2019, developed a new CBIR technique by extracting global and local features [7]. A combination of speeded up robust features (SURF) descriptor with color moments, as local feature, and modified GLCM, as global feature, leads this technique to obtain 70.48% of the precision rate for top-20 using Manhattan similarity measurement. Continuously, in 2019, Khawaja *et al.* proposed another technique for CBIR using object and color features [24]. Authors claimed that this technique outperformed in certain categories of the benchmark datasets Caltech-101 and Corel-1000, and it gained 76.5% of the precision rate for top-20 using Euclidean distance. Different from the previous techniques discussed above, Qazanfari *et al.*, in 2019, investigated HSV color space for developing CBIR technique [25]. As reported in this work, the human visual system is very sensitive to the color as well as edge orientation, and also color histogram and color difference histogram (CDH) are two kinds of low-level feature extraction which are meaningful representatives of the image color and edge orientation information. This proposed technique used Canberra distance measurement to measure the similarity between the extracted feature of the both query image and images in the dataset. This technique achieved 74.77% of the precision rate for the top-20 using Euclidean distance similarity measurement. Rashno *et al.*, in 2019, developed an algorithm in which HSV, RGB, and norm of low frequency components were used to extract color features, and DWT was used to extract texture features [26]. Accordingly, ant colony optimization (ACO) feature selection technique was used to select the most relevant features. Eventually, Euclidian distance measurement was used to measure the similarity between query and images in the dataset. The results reported in this work showed that this approach reached the precision rate of 60.79% using Euclidean distance for the top-20. Finally, Aiswarya *et al.*, in 2020, proposed a CBIR technique which uses a multi-level stacked Autoencoders for feature selection and dimensionality reduction [27]. A query image space is created first before the actual retrieval process by combining the query image as well as similar images from the local image dataset (images in device gallery) to maintain the image saliency in the visual contents. The features corresponding to the query image space elements are searched against the characteristics of images in a global dataset. This technique achieved the precision rate of 67% for top-10.

3. BACKGROUND

This section aims to provide detailed background information about important techniques, used in the proposed approach presented in this paper, such as SURF feature descriptor, color-based features, texture-based features, and feature matching techniques.

3.1. SURF Feature Descriptor

There are many feature descriptors available and SURF is one of the most common and significant feature descriptors in which can be considered as a local feature. In comparison with global features such as color, texture, and shape; local features can provide more detailed characters in an image. The rotation and scale invariant descriptor can perform better in terms of distinctiveness, repeatability, and robustness [12]. SURF is used in many applications such as bag of feature (BoF) which is used and success in image analysis and classification [28]. In the BoF technique, the SURF descriptor is sometimes used first to extract local features. Then K-means clustering is used to initialize M center point to create M visual words. The K-means clustering algorithm takes the feature space as input and reduces it to the M cluster as output. Then, the image is represented as a code word histogram by mapping the local features into

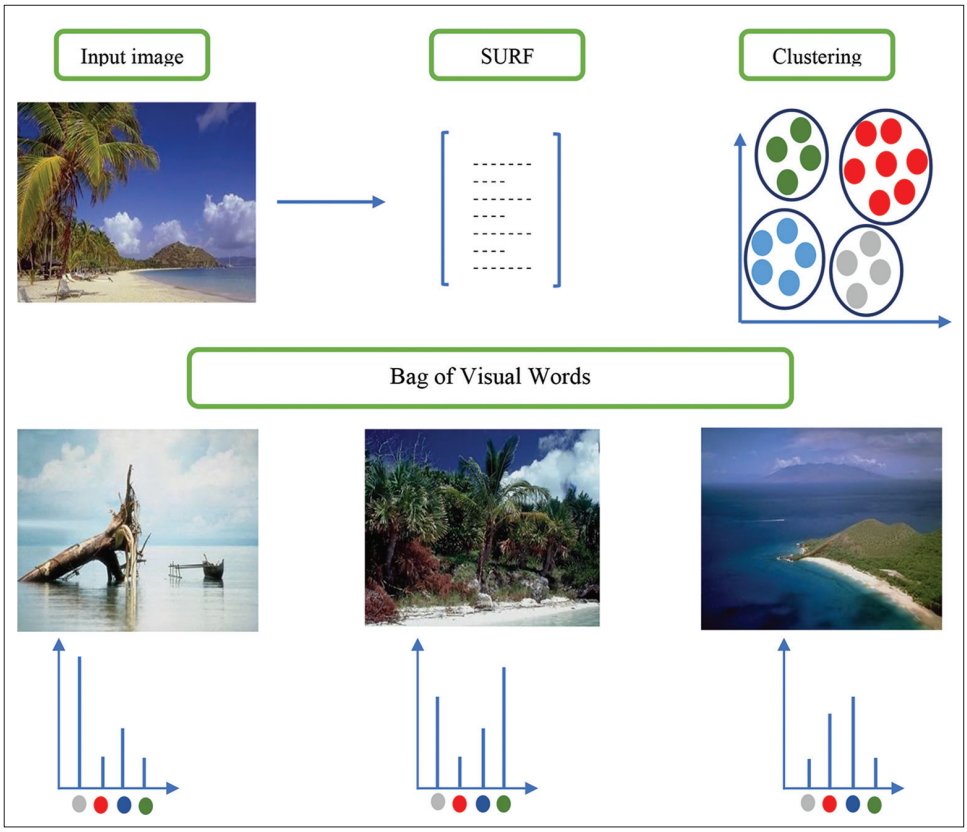

Fig. 2. Methodology of the BoF technique for representing Image in CBIR.

a vocabulary [28]. Fig. 2 illustrates the methodology of the image representation based on the BoF model.

SURF features are extracted from database images, then the k-means clustering algorithm takes feature space as input and reduces it into clusters as output. The center of each cluster is called a visual word and the combination of visual words formulates the dictionary, which is also known as codebook or vocabulary. Finally, using these visual words of the dictionary, the histogram is constructed using visual words of each image. The histogram of v visual words is formed from each image. After that, resultant information in the form of histograms is added to the inverted index of the BoF model [25].

3.2. Texture-based Features Extraction

Texture-based features can be considered as a powerful low-level feature for image search and retrieval applications. There are many works have been developed on texture analysis, classification, and segmentation for the last four decades. Yet, there is no unique definition for the texture-based features. Texture is an attribute representing the spatial arrangement of the gray levels of the pixels in a region or image. In other words, texture-based features can be used to separate and extract prominent regions of interest in an image and apply to the visual patterns that have properties of homogeneity independent of a single color or intensity [9]. Texture analysis methods can be categorized into statistical, structural, and spectral [15]. DWT and LBP are the two methods of texture feature extraction used in this work.

3.2.1. Discrete wavelet transform (DWT)

The DWT is considered to be an efficient multiresolution technique and it is easy to compute [29]. The signal for each level decomposed into four frequency sub-bands which are: Low of low (LL), low of high (LH), high of low (HL), and high of high (HH) [30]. DWT is used to change an image from the spatial domain into the frequency domain, the structure of the DWT is illustrated in Fig. 3 [22], [26].

Wavelet transform could be applied to images as 2-dimensional signals. To refract an image into k level, first the transform is applied on all rows up to k level while columns of the image are kept unchanged. Then, this task is applied on columns while keeping rows unchanged. In this manner, frequency components of the image are obtained up to k level. These frequency components in various levels let us to better analyze original image or signal [26]. For more details about DWT, you can see [31].

3.2.2. Local binary pattern (LBP)

The concept of LBP was originally proposed by Ojala *et al.* in 1996 [29], [32]. LBP can be considered as a texture analysis approach unifying structural and statistical models. The characteristic of LBP is that LBP operator is invariant to monotonic gray-level changes [33]. In the process of LBP calculation, firstly a 3 × 3 grid of image is selected, and then the intensity value of the center pixel can be computed using the intensity values of its neighboring pixels based on the following equations [34]:

$$\mathrm{Lbp}_{n\ \{1,2,3,..,9\}} = \sum_{k=0}^{n-1} 2^k \times \mathrm{fn}\left(I_k - I_c\right) \quad (1)$$

$$\mathrm{fn}\left(a-b\right) = \begin{cases} 1, & \text{if } (a-b) \geq 0 \\ 0, & \text{Otherwise} \end{cases} \quad (2)$$

where, n is the number of neighboring pixels around the center pixel. I_k is the intensity value of the k^{th} neighboring pixel, and I_c is the intensity value of the center pixel. An example of LBP is presented in Fig. 4 [34].

Fig. 4 shows the LBP spectrum of the Lena image with different circular domain radius and sampling points. Correspondingly, fineness of the texture information in the obtained LBP spectrum is different. Taking the Lena image as an example, with the increase of sampling radius, the gray scale statistical value of the LBP map is sparser [34].

In the proposed approach presented in this paper, after LBP is applied on the LH and HL sub-bands of the DWT, 512 features are extracted to represent the image.

3.3. Color-based Features Extraction

Color is considered as a basic feature observation in viewing an image to reveal a variety of information [12]. Color is extremely used feature for image retrieval techniques [35], [36].

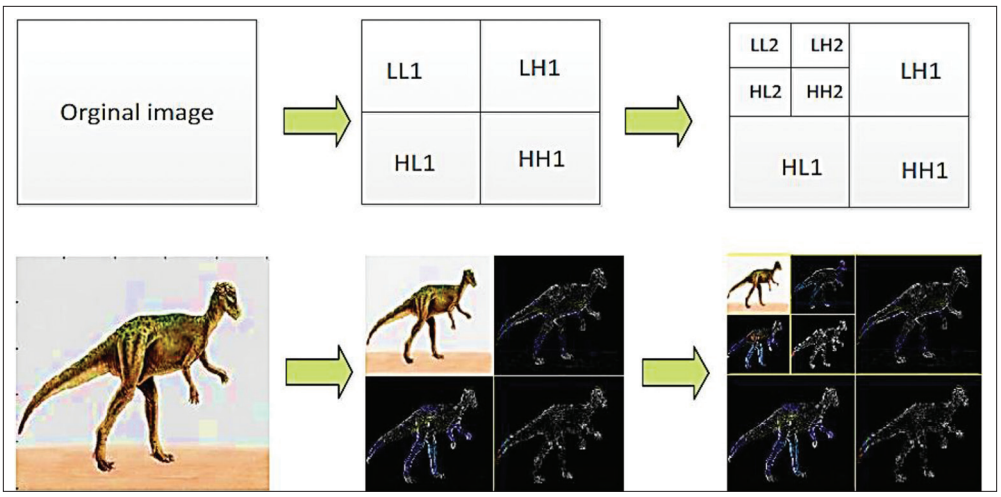

Fig. 3. DWT sub-bands.

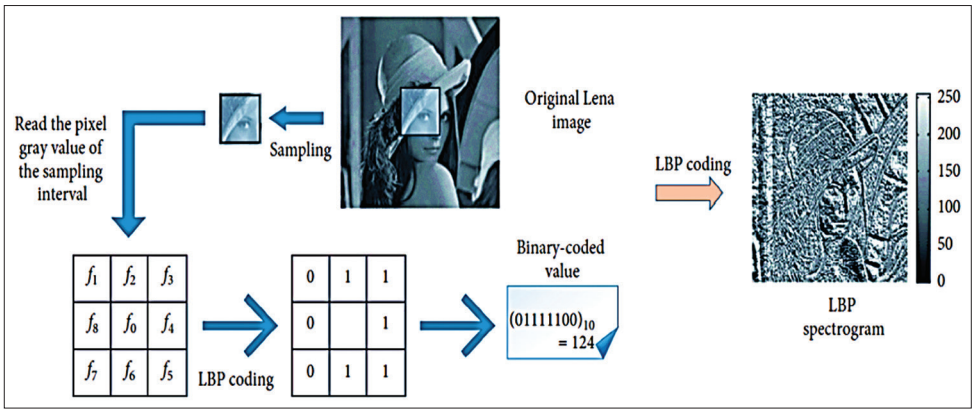

Fig. 4. An example of LBP operator.

Color points create color space and various color spaces based on the perceptual concepts are used for color illustration [23]. Among all color spaces, YCrCb and HSV have the mentioned perceptual characteristic. In YCbCr, the Y represents the luminance while the color is represented by Cb and Cr [37]. In our proposed approach, the mean and entropy of each component of the color spaces RGB, HSV, and YCbCr have been calculated as color features. Meanwhile, totally 18 color-based features are extracted.

3.4. Feature Matching

There are variety of similarity measurements used to determine the similarity between the query image and the images in the dataset [9]. Manhattan distance is used as a similarity measurement for both layers of the proposed approach in this work, equation (3) [36]:

$$\text{Manhattan distance (MD)} = \sum_{i=1}^{k} |x_i - y_i| \quad (3)$$

Where x is the feature vector of query image, y is the feature vector of the dataset of images, and k is the dimension of image feature. Manhattan distance is also known as City block distance. In general, the Manhattan distance is non-negative where zero defines an identical point, and other means little similarity [9].

3.4.1. Proposed approach

This section describes the details of the proposed two-layer approach in the following steps:

1. Let the query image is denoted by Q, and $I = \{I_1, I_2, \ldots, I_n\}$ refers to the dataset which consists of n images.
2. First layer of the proposed approach involves the following steps:
   a. Q_{BoF} and I_{BoF} represent the feature vector of Q and I, respectively, after BoF technique is implemented on.
   b. To find the similarity between Q_{BoF} and I_{BoF}, Manhattan similarity measurement is used, and as a result, M most similar images to the query image are retrieved.
3. Second layer of the proposed approach, which includes the following steps, implements on the query image Q as well as the M most similar images that gained in the first layer.
   a. Extracting the following features from Q and M_i:
      - Let $L = \{l_1, l_2, \ldots, l_{512}\}$ be the vector of 512 extracted texture-based features after LBP is applied on the LH and HL sub-bands of the DWT, 256 features extract from each sub-band.
      - Let $C = \{c_1, c_2, \ldots\ldots, c_{18}\}$ be the extracted 18 color-based features that represent the mean and entropy of the three components of RGB, HSV, and YCbCr color spaces. Meanwhile, 6 features are extracted from each of the mentioned color spaces.
      - Let $F = L + C$ represents the feature vector of the fused of all the 530 extracted features from the previous steps.
      - Finally, Q_F and M_{Fi} represent the fused feature vector of Q and M_i, respectively.
   b. To find the similarity between Q_F and M_{Fi}, Manhattan similarity measurement is used, to retrieve the most similar images to the query image.

The block diagram of the proposed two-layer approach is illustrated in Fig. 5.

4. EXPERIMENTAL RESULTS

Experiments are conducted comprehensively in this section to evaluate the performance of the proposed approach in terms of precision rate, the most common confusion matrix measurement used in the research area of CBIR. In addition, the proposed approach is compared to the current existing works.

4.1. Dataset

Corel-1K dataset of images has been used, which is a public and well-known dataset, that contains 1000 images in the form of 10 categories and each category consists of 100 images with resolution sizes of (256 × 384) or (384 × 256) [37], [38]. The categories are arranged as follows: African, people, beaches, buildings, buses, dinosaurs, elephants, flowers, horses, mountains, and foods [37].

4.2. Evaluation Measurements

To evaluate the performance of the proposed approach, precision confusion matrix measurement has been used which determines the number of correctly retrieved images to the total number of the retrieved images from the tested dataset of images. Meanwhile, it measures the specificity of image retrieval system based on the following equation [38], [39]:

$$\text{Precision} = \frac{R_c}{R_t} \quad (4)$$

where R_c represents the total number of correctly retrieved images and R_t represents the total number of retrieved images. In this study, top-10 and top-20 have been tested. Top-10 indicates the total number of retrieved images is 10 images, and top-20 indicates the total number of retrieved images is 20 images.

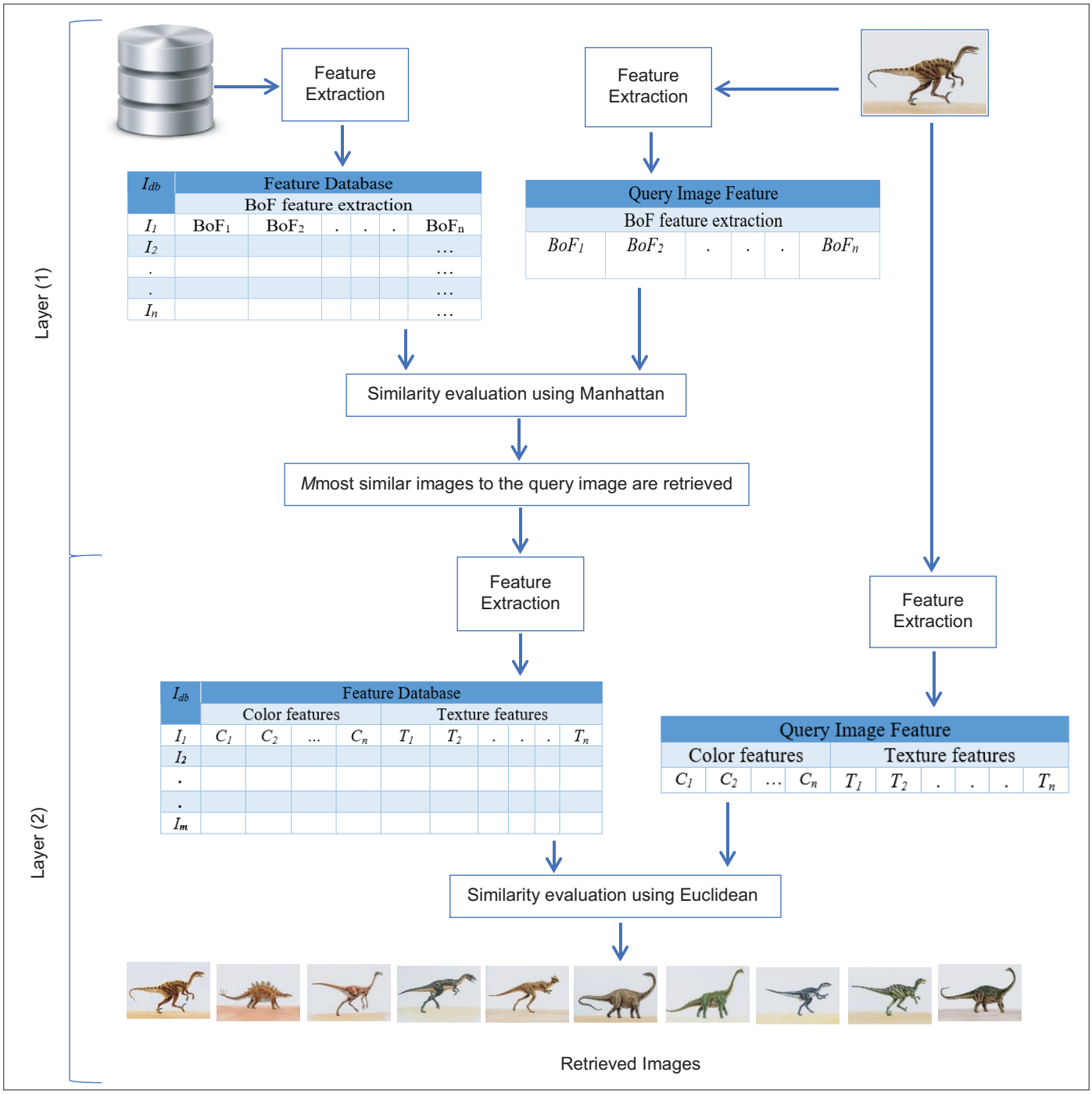

Fig. 5. Block diagram of the proposed two-layer CBIR approach.

4.3 Results

The experiments carried out in this work include two parts: (a) Single layer CBIR model and (b) Two-layer CBIR model. First part evaluates the single layer model (i.e., BoF technique) alone, and on the other hand, CBIR technique based on extracting texture and color features is evaluated. In the second part, the proposed two-layer model has been assessed. The experiments are detailed in the following steps:

1. BoF-based CBIR technique is tested using different number of clusters, as BoF technique relies on the K-means clustering algorithm to create clusters, which is commonly called visual words. The number of clusters

cannot be selected automatically; manual selection is needed. To select the proper number of clusters, (i.e., value of k-means), the different number of clusters have been tested to obtain the best precision result of BoF technique. The precision results of different number of clusters are illustrated in the following tables.

From Tables 1 and 2, it is quite obvious that the best result is achieved when k = 500 for both top-10 and top-20.

2. The DWT sub-bands, and the concatenation of the sub-bands, have been tested as a texture feature as presented in the following tables.

From Tables 3 and 4, one can observe that the best result is obtained when the LH and HL sub-bands are concatenated for both top-10 and top-20.

3. In this step, LBP is implemented on DWT sub-bands, implementation of LBP on LH and HL sub-bands.

In the proposed method, LBP is extracted from DWT sub-bands to form a sub-novel local feature descriptor. To achieve this, we performed DWT decomposition and consider the high frequency sub-bands HL, and LH. However, the sub-bands HL and LH also contain edge and contour details of image's significant in extracting pose and expression relevant features with the aid of LBP. We ignored the low-frequency LL and the high-frequency HH sub-band as it mostly contains the noise with negligible feature details. To preserve the spatial characteristics and to form a robust local feature descriptor, multi-region LBP pattern-based features [4] are obtained from non-overlapping regions of DWT sub-bands {HL, LH}, are statistically significant and offer reduced dimensionality with increased robustness to noise. Each of the sub-band {HL, LH} is equally divided into m non-overlapping rectangle regions R0; R1;… ; Rm, each of size (x,y) pixels. From each of these m regions, we extract local features LBP each with 256 labels separately. Local features from successive regions are concatenated to form a combining the two results in one vector with 512 features, the results in Tables 5 and 6.

From Tables 5 and 6, one can observe that the best result is obtained when the implementation of LBP on LH sub-band as well as HL sub-band is concatenated.

TABLE 1: Precision rate of BoF technique for different number of clusters for top-10

Different number of clusters	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
K=100	61.1	47.2	50.1	88.9	100	68.9	87.3	89.6	46.4	53.8	71.056
K=200	65.7	49.3	55.2	88.9	100	70.9	87	91.8	49.1	53.3	73.1
K=300	65.2	49.9	57.1	89.2	100	72.3	87.4	93.1	48.4	54.4	73.622
K=400	65.5	47.9	56.7	90.9	100	72.1	87.6	92.4	48.9	53.4	73.556
K=500	67.1	49.3	60.4	89.2	100	70.1	87.7	93.6	51.5	53.9	74.322
K=600	64.9	47.9	59.1	89.4	100	70.6	88.6	93.6	51.7	53.4	73.978
K=700	67.1	49.3	60.4	88.2	100	70.1	86.7	93.6	51.5	53.9	74.1
K=800	64.6	48	59.4	89.2	100	68.2	88.3	94.2	53.5	51	73.933
K=900	64.7	46.2	62	88.1	100	68	87.5	93.4	55.6	52.7	73.944
K=1000	64.5	47	61.5	87.6	100	66.9	88.3	94.2	53.4	50.9	73.711

TABLE 2: Precision rate of BoF technique for different number of clusters for top-20

Different number of clusters	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
K=100	54.75	46.5	42.4	84.5	100	58.55	84.9	82	39.9	40.15	63.365
K=200	56.95	48.65	42.5	86.35	100	59.85	85.5	87.05	41	42.3	65.015
K=300	58.55	48.55	45	85.7	100	61.05	85.5	88.4	42.45	41.95	65.715
K=400	58.45	48.25	47.3	87.35	100	59.2	85.65	87.65	44.65	41.35	65.985
K=500	60.5	48.75	50.3	85.6	99.95	59.05	85	87.9	47.15	40.55	66.475
K=600	60.2	48.5	50	84.8	99.95	58.85	84.7	87.7	47.1	40.05	66.185
K=700	57.95	48.5	51.6	84.85	99.95	57.5	84.75	88.45	44.75	39.65	65.795
K=800	58.3	48.7	50.55	84.3	99.95	58.95	85.95	89.3	47.35	39.15	66.25
K=900	57.65	47.95	51.7	84.35	99.95	57.6	85.05	88.75	47.7	39.3	66
K=1000	56.8	48.2	52.95	82.6	99.85	56.95	85.15	88.4	46.15	37.55	65.46

TABLE 3: Precision rate for the DWT sub-bands for top-10

DWT sub-bands	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
LL	17.2	48.4	44.6	24.85	95.1	29.4	27.25	16.5	26.15	20.25	34.97
LH	25.05	36.05	31.8	24.3	97.35	29.85	29.05	22	27.95	27.95	35.14
HL	11.55	21.55	22.55	23.15	93.3	23.2	24.25	21.25	21.65	19.7	28.22
HH	21.2	12.55	17.85	22.8	92.95	22.85	13.9	20.9	11.3	15.35	25.17
LL_LH	20.2	33.15	29.8	26.9	98	34.5	34.2	20.55	30.6	23.95	35.19
LL_HL	30.2	32.9	30.1	26.8	98.55	33.35	34.45	21.15	31.1	24.85	36.35
LL_HH	29	31.25	28.2	25.05	96.8	31.9	32.6	19.85	29.1	23.2	34.7
LH_HL	26.85	34.05	34.05	33.7	98.55	27.8	39.05	25.2	26.15	23.35	36.88
LH_HH	15.61	22.86	22.46	15.86	98.16	21.51	19.26	15.71	21.56	22.76	27.58
HL_HH	11.85	27.8	22.5	8.1	92.4	23.75	21.5	17.05	15.8	19	25.98
LL_LH_HH	16.05	32	26.7	12.3	96.6	27.95	25.7	21.25	20	23.2	30.18
LL_HL_HH	19.85	33.05	30.3	25.9	99.25	33.1	34.2	19.9	31.25	24.25	35.11
LL_LH_HH	24.66	33.86	32.11	30.71	96.2	31.91	31.01	20.71	30.06	22.06	35.329
LL_LH_HL_HH	17.6	29.2	29.5	24.3	99.7	31.8	30.6	15	30.3	23.4	33.14

TABLE 4: Precision rate for the DWT sub-bands for top-20

DWT sub-bands	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
LL	15.2	29.7	26.75	22.85	94.1	29.4	27.25	16.5	26.15	20.25	30.82
LH	20.05	23.5	22.9	15.3	97.35	21.85	19.05	15	23.95	22.95	28.19
HL	11.55	12.9	18.2	13.15	93.3	13.2	14.25	11.25	11.65	15.7	21.52
HH	17.55	8.9	14.2	19.15	89.3	19.2	10.25	17.25	7.65	11.7	21.52
LL_LH	16.55	29.5	26.15	23.25	94.35	30.85	30.55	16.9	26.95	20.3	31.54
LL_HL	25.95	28.65	25.85	22.55	95.3	29.1	30.2	16.9	26.85	20.6	32.2
LL_HH	26.45	28.7	25.65	22.5	94.25	29.35	30.05	17.3	26.55	20.65	32.15
LH_HL	22.3	29.5	29.9	29.55	94.4	23.65	34.9	21.2	22	19.2	32.66
LH_HH	11.85	19.1	18.7	12.1	94.4	17.75	15.5	11.95	17.8	19	23.82
HL_HH	7.85	15.1	14.7	8.1	92.4	13.75	11.5	7.95	13.8	15	20.02
LL_LH_HH	11.85	19.1	18.7	12.1	96.4	17.75	15.5	11.95	17.8	19	24.02
LL_HL_HH	15.95	29.15	26.4	22	95.35	29.2	30.3	16	27.35	20.35	31.21
LL_LH_HH	17.35	30.55	27.8	23.4	95.75	30.6	31.7	17.4	28.75	21.75	32.51
LL_LH_HL_HH	11.95	25.4	23.3	18.55	95.2	27.05	27.2	11.65	23.95	15.85	28.01

TABLE 5: Precision rate for LBP for top-10

DWT sub-bands	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
LL	37.35	50.55	47.8	43.4	115.75	50.6	51.7	37.4	48.75	41.75	52.51
LH	56.8	49.5	54.1	85.2	99.5	45.8	91.5	70.2	39.8	47.8	64.02
HL	74.1	54.2	58	92.9	99.9	56.3	89.1	78.1	46.5	63.8	71.29
HH	53.5	40.9	54.9	90.1	98.6	39.8	92.3	57.1	37.8	46.8	61.18
LL_LH	63.4	50.4	61.8	92.3	98.7	44.6	92.9	74.6	39.1	51.4	66.92
LL_HL	62.64	49.54	61.02	91.74	98.18	43.7	92.34	73.91	38.17	50.55	66.18
LL_HH	61.19	48.19	59.59	90.09	96.49	42.39	90.69	72.39	36.89	49.19	64.71
LH_HL	70.7	61.4	69.7	95.5	99.2	55.1	93.4	83	46.4	65	73.94
LH_HH	64.45	51.45	62.85	93.35	99.75	45.65	93.95	75.65	40.15	52.45	67.97
HL_HH	63.45	50.45	61.85	92.35	98.75	44.65	92.95	74.65	39.15	51.45	66.97
LL_LH_HH	64.55	51.75	63.05	92.95	99.75	45.95	94.05	75.45	40.15	52.45	68.01
LL_HL_HH	62.75	49.95	61.25	91.15	97.95	44.15	92.25	73.65	38.35	50.65	66.21
LL_LH_HH	74.35	57.95	67.75	95.75	99	54.05	94.15	81.85	47.55	65.05	73.75
LL_LH_HL_HH	72	55.8	65.5	93.6	98.3	52	91.9	79.7	45.5	62.9	71.72

4. Before selecting an appropriate color description, selection of color space is important and needs to choose a color model for color feature extraction process [35]. This step evaluates the impact of extracting the color

feature by testing different color space components such as: RGB, YCbCr, and HSV. In other words, mean and entropy for each color components have been calculated and the results are presented in the following tables.

The results presented in Tables 7 and 8 demonstrates that combining the extracted features of all the tested color spaces provides best precision rate.

5. Finally, all the extracted features are fused. Meanwhile, by concatenating the extracted LBP in step 3 and the extracted color feature in step 4, Table 9.
6. Eventually, the proposed two-layer approach has been tested. It includes two layers: The first layer implements BoF technique (for K=500) and M most similar images are retrieved, M is user defined. In the second layer, color and texture features are extracted from the query image and the M remained images, as a result, N most similar images are retrieved. The following tables investigate the best value of M. In other words, Tables 10 and 11 show investigating different number of M for top-10 and top-20, respectively.

Results in Tables 10 and 11 demonstrate that the best precision results are obtained for M = 100 and M=200. For

TABLE 6: Precision rate for LBP for top-20

DWT sub-bands	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
LL	35.15	48.35	45.6	41.2	113.55	48.4	49.5	35.2	46.55	39.55	50.31
LH	48.95	40.7	45.25	79.65	96.6	36.85	87.8	64.75	32.85	40.75	57.42
HL	67.55	48.7	49.2	89.85	98.4	44.7	86.05	68.9	40.7	57.35	65.14
HH	44.8	36	46.25	84.4	98.1	32.1	89.9	49.9	30.75	38.95	55.12
LL_LH	53	43.3	51	87.95	98.1	35.95	90.5	62.85	34.3	42.15	59.91
LL_HL	52.16	42.39	50.15	87.36	97.58	34.99	89.93	62.08	33.33	41.24	59.12
LL_HH	50.79	41.09	48.79	85.74	95.89	33.74	88.29	60.64	32.09	39.94	57.7
LH_HL	63.8	54.45	59.85	93.2	99.15	43.75	90.65	72.75	41	58.2	67.68
LH_HH	54.05	44.35	52.05	89	99.15	37	91.55	63.9	35.35	43.2	60.96
HL_HH	53.05	43.35	51.05	88	98.15	36	90.55	62.9	34.35	42.2	59.96
LL_LH_HH	54.1	44.65	52.15	88.65	99.15	36.9	91.55	63.8	35.4	43.35	60.97
LL_HL_HH	52.3	42.85	50.35	86.85	97.35	35.1	89.75	62	33.6	41.55	59.17
LL_LH_HH	64.5	50.6	57.8	93.3	98.9	43.35	92.45	69.75	40.7	56	66.74
LL_LH_HL_HH	62.25	48.4	55.9	90.9	97.65	41.2	90.2	67.6	38.45	53.7	64.625

TABLE 7: Precision rate for color-based features for top-10

Color spaces	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
RGB	45.8	35.5	33.2	41.6	99.4	49.6	62.6	83.7	39.4	56.4	54.72
HSV	21.1	22.65	32.8	26.7	39.5	23.9	19.9	29.45	20	22.1	25.81
YCbCr	47.6	35.9	40.1	46.9	100	53.4	65.7	56.6	39.3	51	53.65
RGB_HSV	48	37.3	41.7	48.8	99.3	52.3	60.3	84	38.5	56.7	56.69
RGB_YCbCr	50.7	37.4	38.5	50.8	99.7	55.2	68	86.4	38.8	55.6	58.11
HSV_YCbCr	41.5	33.5	41.1	46.8	99.6	48.65	59.65	56.1	35.2	44.2	50.63
RGB_HSV_ YCbCr	47.8	38.9	47.6	60.3	99.7	52.4	63.9	82.2	44.6	56.8	59.42

TABLE 8: Precision rate for color-based features for top-20

Color spaces	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
RGB	40	28.8	27.55	34.5	99.5	45.45	54.6	77.7	34.5	48.15	49.075
HSV	27.1	29.3	37.7	31.2	50.4	28.1	26.3	37.1	25.2	27.5	31.99
YCbCr	39.2	30.55	34.5	39.65	99.85	47.8	60.25	49.55	35.1	44	48.045
RGB_HSV	41.05	30.85	34.55	41.05	99.15	46.15	51.8	77.25	36.25	48.2	50.63
RGB_YCbCr	42.55	30.3	31	42.25	99.55	49	58.9	78.45	33.85	48.05	51.39
HSV_YCbCr	44.4	33.4	45.5	48.9	95.2	49.8	60.5	61.4	35.3	47.2	52.16
RGB_HSV_ YCbCr	41.9	33.3	37.15	55.35	99.5	47.25	56.05	73.85	38.55	48.25	53.115

this reason, different numbers of M in the range of M=100 to M=200 have also been investigated to gain better precision result, Tables 12 and 13.

From Tables 12 and 13, it is quite clear that the best result is obtained when M =110 for both top-10 and top-20. More experiments have been done to compare the proposed approach with the state-of-the-art techniques, Table 14.

According to the results presented in Table 14, the best performance (precision rate) is achieved by the proposed approach for both top-10 and top-20. All the tested

TABLE 9: Concatenation of texture-based feature and color-based feature

Top image retrieved	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
Top-20	67.45	49.1	60.65	88.8	99.45	61.7	90.5	81.8	49	61.55	71
Top-10	67.45	56.25	69.5	88.8	99.45	61.7	90.5	81.8	49	61.55	72.6

TABLE 10: Precision rate for different number of M for top-10

Different Number of M	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
M=100	81.65	63	71.9	95.95	99.85	77.6	96.35	92.75	67.55	73.45	82.005
M=200	78.85	62.2	73.65	95.7	99.85	76.6	95.95	91.6	66.65	75.1	81.615
M=300	77.05	61.1	74.65	94.8	99.85	75.4	96.05	90.65	65.4	75.35	81.03
M=400	76.6	60.1	74.7	94.65	99.65	74.8	95.95	90.3	64.45	75.15	80.635
M=500	76.5	60.2	74.75	94.35	99.65	74.2	95.85	90.25	63.45	74.85	80.405
M=600	76.2	60.1	74.3	94	99.65	74	95.65	90	63.3	74.75	80.195
M=700	76.15	60	74.1	93.85	99.65	73.85	95.45	90	63.15	74.55	80.075
M=800	76	59.4	74	93.6	99.2	73.25	95.15	89.4	63	74.2	79.72
M=900	76.35	59.9	75.2	94.15	99.65	73.4	95.85	90	62.1	75.05	80.165
M=1000	76.3	59.3	75.2	94.1	99.25	73.2	95.55	89.7	62	74.85	79.945

TABLE 11: Precision results for different number of M for top-20

Different number of M	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
M=100	76.35	58.8	61.85	94.95	99.9	70.35	94.1	90.3	61.45	64.15	77.22
M=200	75.25	57.45	63.85	94.35	99.6	70.1	93.95	87.45	60.6	66.25	76.885
M=300	73.1	55.5	64.4	92.9	99.55	68.9	93.8	86.35	58.85	66.75	76.01
M=400	72.1	54.47	65.4	92.4	99.55	68.15	93.6	85.95	57.75	66.5	75.587
M=500	72	54.45	65.4	92	99.55	67.85	93.55	85.75	57.25	66.3	75.41
M=600	71.7	54	65.1	91.6	99.55	67.75	93.3	85.65	57.15	66.2	75.2
M=700	71.6	53.8	65	91.4	99.45	67.6	93	85.4	57.05	66.2	75.05
M=800	71.5	53.5	64.8	91.2	99	67.1	93	85.1	56.8	66	74.8
M=900	71.7	54.05	65.5	91.45	99.55	66.75	93.45	85.5	54.05	66.45	74.845
M=1000	71.2	54.05	65.2	91.15	99.05	66.65	93.15	85.4	53.85	66.35	74.605

TABLE 12: Precision rate of the proposed approach for different number of M for top-10

Different Number of M	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
M=30	82.89	64.64	70.39	95.44	99.74	76.79	96.24	94.19	68.94	69.94	81.92
M=50	83.15	64.9	70.65	95.7	100	77.05	96.5	94.45	69.2	70.2	82.18
M=70	81.9	63.95	70.85	95.6	100	77.2	96.65	93.45	68.5	72.65	82.08
M=90	81.55	62.65	71.75	95.9	99.95	77.85	96.6	93	68.45	72.65	82.04
M=110	80.95	64.25	72.6	96	99.95	77.75	96.35	92.55	68	73.05	82.15
M=130	81.85	63.8	71.05	95.8	99.92	77.65	96.4	92.4	68.2	72	81.91
M=150	81.49	63.38	71.86	95.8	99.9	77.5	96.4	92.55	68	72.9	81.98
M=170	81.1	62.25	71.29	95.33	99.58	77.3	96	92.2	67.7	72	81.48
M=190	80	62.6	70.9	94.6	99.2	76.45	95.3	91.8	67	71.7	80.96

TABLE 13: Precision rate of the proposed approach for different number of M for top-20

Different number of M	Categories										
	Africa	Beaches	Buildings	Buses	Dinosaur	Elephant	Roses	Horses	Mountains	Food	Average
M=30	75.14	57.84	59.19	94.09	99.74	69.29	93.29	90.64	58.54	59.24	75.70
M=50	75.40	58.10	59.45	94.35	100.00	69.55	93.55	90.90	58.80	59.50	75.96
M=70	76.80	59.00	60.65	94.45	99.95	69.80	94.65	90.65	59.65	62.10	76.77
M=90	76.95	58.70	61.35	94.85	99.95	69.65	94.15	90.45	61.00	63.70	77.08
M=110	76.00	59.10	62.35	95.00	99.90	70.35	93.85	90.20	61.55	64.40	77.27
M=130	76.35	58.6	60.7	94.74	99.2	69.45	94.1	90.4	61	62.2	76.674
M=150	76.4	58.9	61.5	94.7	99.35	70	94.3	90.1	61.1	61.58	76.793
M=170	76.38	58	61	94.4	99	69.4	93.2	89.9	61.69	61.99	76.496
M=190	75.25	57.6	60.5	93.7	98.93	69.33	93.1	89	60.87	61.93	76.021

TABLE 14: Precision rate of the tested CBIR techniques

Approaches	Top-10	Top-20
Proposed approach	82.15	77.27
Atif Nazir, 2018	-	73.5
Pradhan Jitesh, 2019	64.00	59.60
Khawaja and Ahmed , 2019	-	76.5
Hamed Qazanfari, 2019	-	74.77
Aiswarya, 2020	67	

state-of-the-art techniques, except technique in Ahmed and Naqvi [24], they tested their approach either for top-10 or for top-20, and this is why in Table 14 some cells are not contained the precision rate.

5. CONCLUSION

The two-layer based CBIR approach for filtering and minimizing the dissimilar images in the dataset of images to the query image has been developed in this study. In the first layer, the BoF has been used and as a result, M most similar images are remained for the next layer. Meanwhile, the most dissimilar images are eliminated and, hence, the range of search is narrowed for the next step. The second layer concentrated on concatenating the extracted both (texture and color)-based features. The results obtained by the proposed approach devoted the impact of exploring the concept of two-layer in improving the precision rate compared to the existing works. The proposed approach has been evaluated using Corel-1K dataset and the precision rate of 82.15% and 77.27% for top-10 and top-20 is achieved, respectively. In the future, certain feature extractors need to be investigated as well as feature selection techniques need to be added to select the most important feature which reflects increasing precision rate.

REFERENCES

[1] Z. F. Mohammed and A. A. Abdulla. "Thresholding-based white blood cells segmentation from microscopic blood images". *UHD Journal of Science and Technology*, vol. 4, no. 1, p. 9, 2020.

[2] M. W. Ahmed and A. A. Abdulla. "Quality improvement for exemplar-based image inpainting using a modified searching mechanism". *UHD Journal of Science and Technology*, vol. 4, no. 1, p. 1, 2020.

[3] H. Liu, J. Yin, X. Luo and S. Zhang. "*Foreword to the Special Issue on Recent Advances on Pattern Recognition and Artificial Intelligence*". Springer, Berlin, p. 1, 2018.

[4] A. Wojciechowska, M. Choraś and R. Kozik. "*Evaluation of the Pre-processing Methods in Image-Based Palmprint Biometrics*". Springer, International Conference on Image Processing and Communications, p. 1, 2017.

[5] A. A. Abdulla, S. A. Jassim and H. Sellahewa. "*Secure Steganography Technique Based on Bitplane Indexes*". 2013 IEEE International Symposium on Multimedia, 2013.

[6] A. A. Abdulla. "*Exploiting Similarities between Secret and Cover Images for Improved Embedding Efficiency and Security in Digital Steganography*". Department of Applied Computing, The University of Buckingham, United Kingdom, pp. 1-235, 2015.

[7] S. Farhan, B. K. Biswas and R. Haque. "*Unsupervised Content-Based Image Retrieval Technique Using Global and Local Features*". International Conference on Advances in Science, Engineering and Robotics Technology, p. 2, 2019.

[8] R. S. Patil, A. J. Agrawal. "Content-based image retrieval systems: A survey". *Advances in Computational Sciences and Technology*, vol. 10, 9, pp. 2773-2788, 2017.

[9] H. Shahadat and R. Islam. "A new approach of content based image retrieval using color and texture features". *Current Journal of Applied Science and Technology*, vol. 21, no. 1, pp. 1-16, 2017.

[10] A. Sarwar, Z. Mehmood, T. Saba, K. A. Qazi, A. Adnan and H. Jamal. "A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine". *Journal of Information*, vol. 45, pp. 117-135, 2019.

[11] L. K. Paovthra and S. T. Sharmila. "Optimized feature integration and minimized search space in content based image retrieval". *Procedia Computer Science*, vol. 165, pp. 691-700, 2019.

[12] A. Masood, M. A. Shahid and M. Sharif. "Content-based image retrieval features: A survey". *The International Journal of Advanced Networking and Applications*, vol. 10, no. 1, pp. 3741- 3757, 2018.

[13] S. Singh and S. Batra. "*An Efficient Bi-layer Content Based Image*

Retrieval System". Springer, Berlin, p. 3, 2020.

[14] Y. D. Mistry. "Textural and color descriptor fusion for efficient content-based image". *Iran Journal of Computer Science*, vol. 3, pp. 1-15, 2020.

[15] K. T. Ahmed, A. Irtaza, M. A. Iqbal. "*Fusion of Local and Global Features for Effective Image Extraction*". Elsevier, Amsterdam, Netherlands, vol. 51, pp. 76-99, 2019.

[16] M. O. Divya and E. R. Vimina. "*Maximal Multi-channel Local Binary Pattern with Colour Information for CBIR*". Springer, Berlin, p. 2, 2020.

[17] T. Kato. "*Database Architecture for Content-based Image Retrieval*". International Society for Optics and Photonics, vol. 1662. pp. 112-123, 1992.

[18] J. Yu, Z. Qin, T. Wan and X. Zhang. "Feature integration analysis of bag-of-features model for image retrieval". *Neurocomputing*, vol. 120, pp. 355-364, 2013.

[19] N. Shrivastava. "*Content-based Image Retrieval Based on Relative Locations of Multiple Regions of Interest Using Selective Regions Matching*". Elsevier, Amsterdam, Netherlands, vol. 259, pp. 212-224, 2014.

[20] E. Gupta and R. S. Kushwah. "*Combination of Global and Local Features Using DWT with SVM for CBIR in Reliability*". Infocom Technologies and Optimization (ICRITO) Trends and Future Directions, 2015.

[21] E. Gupta and R. S. Kushwah. "Content-based image retrieval through combined data of color moment and texture". *International Journal of Computer Science and Network Security*, vol. 17, pp. 94-97, 2017.

[22] A. Nazir, R. Ashraf, T. Hamdani and N. Ali. "*Content Based Image Retrieval System by using HSV Color Histogram, Discrete Wavelet Transform and Edge Histogram Descriptor*". 2018 International Conference on Computing, Mathematics and Engineering Technologies, p. 4, 2018.

[23] P. Jitesh, A. Ashok, P. A. Kumarand and B. Haider. "Multi-level colored directional motif histograms for content-based". *The Visual Computer*, vol. 36, pp. 1847-1868, 2020.

[24] K. T. Ahmed and S. H. Naqvi. "*Convolution, Approximation and Spatial Information Based Object and Color Signatures for Content Based Image Retrieval*". 2019 International Conference on Computer and Information Sciences, 2019.

[25] H. Qazanfari, H. Hassanpour and K. Qazanfari. "Content-based image retrieval using HSV color space features". *International Journal of Computer and Information Engineering*, vol. 13, no. 10, pp. 537-545, 2019.

[26] E. Rashno. "Content-based image retrieval system with most relevant features among wavelet and color features". *Iran University of Science and Technology*, vol. pp. 1-18, 2019.

[27] K. S. Aiswarya, N. Santhi and K. Ramar. "Content-based image retrieval for mobile devices using multi-stage autoencoders". *Journal of Critical Reviews*, vol. 7, pp. 63-69, 2020.

[28] J. Zhou, X. Liu, W. Liu and J. Gan. "Image retrieval based on effective feature extraction and diffusion process". *Multimedia Tools and Applications*, vol. 78, no. 5, pp. 6163-6190, 2019.

[29] P. Srivastava. "*Content-Based Image Retrieval Using Multiresolution Feature Descriptors*". Springer, Berlin, pp. 211-235, 2019.

[30] I. A. Saad. "An efficient classification algorithms for image retrieval based color and texture features". *Journal of AL-Qadisiyah for Computer Science and Mathematics*, vol. 10, no. 1, pp. 42-53, 2018.

[31] M. S. Haji. "Content-based image retrieval: A deep look at features prospectus". *International Journal of Computational Vision and Robotics*, vol. 9, no. 1, pp. 14-37, 2019.

[32] V. Geetha, V. Anbumani, S. Sasikala and L. Murali. "*Efficient Hybrid Multi-level Matching with Diverse Set of Features for Image Retrieval*". Springer, Berlin, pp. 12267-12288, 2020.

[33] R. Boukerma, S. Bougueroua and B. Boucheham. "*A Local Patterns Weighting Approach for Optimizing Content-Based Image Retrieval Using a Differential Evolution Algorithm*". 2019 International Conference on Theoretical and Applicative Aspects of Computer Science, 2019.

[34] Y. Cai, G. Xu, A. Li and X. Wang. "A novel improved local binary pattern and its application to the fault diagnosis of diesel engine". *Shock and Vibration*, vol. 2020, p. 9830162, 2020.

[35] G. Xie, B. Guo, Z. Huang, Y. Zheng and Y. Yan. "*Combination of Dominant Color Descriptor and Hu Moments in Consistent Zone for Content Based Image Retrieval*". IEEE Access, vol. 8, pp. 146284-146299, 2020.

[36] A. C. Nehal and M. Varma. "*Evaluation of Distance Measures in Content Based Image Retrieval*". 2019 3rd International conference on Electronics, Communication and Aerospace Technology, pp. 696-701, 2019.

[37] S, Bhardwaj, G. Pandove and P. K. Dahiya. "A futuristic hybrid image retrieval system based on an effective indexing approach for swift image retrieval". *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 12, pp. 1-13, 2020.

[38] S. P. Rana, M. Dey and P. Siarry. "Boosting content based image retrieval performance through integration of parametric and nonparametric approaches". *Journal of Visual Communication and Image Representation*, vol. 58, pp. 205-219, 2019.

[39] M. K. Alsmadi. "*Content-Based Image Retrieval Using Color, Shape and Texture Descriptors and Features*". Springer, Berlin, pp. 1-14, 2020.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A State-of-the-Art Review on Machine Learning-based Methods for Prostate Cancer Diagnosis

Ari Mohammed ali Ahmed[1], Aree Ali Mohammed[2]

[1]Department of Information Technology, Technical College of Informatics, Sulaimani Polytechnic University, KRG, Sulaimani, Iraq, [2]Department of Computer Science, College of Science, University of Sulaimani, Sulaymaniyah, Iraq

ABSTRACT

Prostate cancer can be viewed as the second most dangerous and diagnosed cancer of men all over the world. In the past decade, machine and deep learning methods play a significant role in improving the accuracy of classification for both binary and multi classifications. This review is aimed at providing a comprehensive survey of the state of the art in the past 5 years from 2015 to 2020, focusing on different datasets and machine learning techniques. Moreover, a comparison between studies and a discussion about the potential future researches is described. First, an investigation about the datasets used by the researchers and the number of samples associated with each patient is performed. Then, the accurate detection of each research study based on various machine learning methods is given. Finally, an evaluation of five techniques based on the receiver operating characteristic curve has been presented to show the accuracy of the best technique according to the area under curve (AUC) value. Conducted results indicate that the inception-v3 classifier has the highest score for AUC, which is 0.91.

Index Terms: Prostate cancer, Machine learning, Deep learning, Algorithm, and Datasets

1. INTRODUCTION

Cancer is a category of diseases that includes cell growth which is irregular with the ability to spread to other areas of the body. Physicists have concentrated on continuous advancement in imaging methods over the past decades, enabling radiologists to improve cancer detection and diagnosis. However, the human diagnosis still suffers from poor repeatability, associated with false identification or perception in clinical decisions of anomalies. Two factors influence these inaccuracies: The ability to observe is limited, for example, perception of human vision is constrained, fatigue duty, or confusion, and the second factor is the clinical case complexity, for instance, unbalanced data which are the mean number of healthy cases are more than a malignant case. Different machine learning-based techniques for cancer detection and classification have introduced a new area of research for early cancer detection. The researches will lead to the ability to reduce the manual system impairments [1]. Another reason, modality that has various analysis techniques such as inappropriate diagnostics, handling, and complicated history is leading to increasing mortality [2].

In the past decades, the field of digital pathology has dramatically developed due to the improvement of algorithms in image processing, machine learning, and advancements in computational power. Within this sector, countless approaches have been suggested to analyze and classify automated pathological images. At present, many

Access this article online

DOI:10.21928/uhdjst.v5n1y2021.pp41-47 E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: ari.m.ali@spu.edu.iq

Received: 19-10-2020 Accepted: 27-03-2021 Published: 31-03-2021

smart and powerful features are added to the microscope and digital images to convert slides of stained tissues into entire digital images. These facilities make a more efficient computer diagnosis system to analyze histopathology and helping early diagnosis. Moreover, they treat cancer by avoiding the increase of cancer cells and easily controlling the tumors from spreading to other parts of the body [3].

In addition, analysis of medical imaging could be significantly involved in identifying defects in various body organs, such as prostate cancer (PCa), blood cancer, skin cancer, breast cancer, brain cancer, and lung cancer. The abnormality of the organ is mainly the result of rapid tumor development, which is the world's leading cause of death. As mentioned by GLOBOCAN statistics, around 18.1 million new cancer cases have appeared in 2018 that gave rise to 9.6 million cancer deaths [2]. PCa is considered the most dangerous disease type of cancer, and it is viewed as the second most commonly diagnosed cancer [3], [4]. The most ubiquitous form of cancer in men is PCa and it has been reported to be the second leading cause of death in men [5].

In the USA, the occurrence of PCa ranks first in men whereas in South Korea, is the fifth most common cancer among males, and the expected cancer deaths in 2018 were 82,155 [3]. PCa is the most leading cancer among men, after lung cancer. It is estimated that about 174,650 new cases and 31,620 PCa-related deaths were recorded in the United States in 2019. PCa considers about 1 in 5 new cancer diagnoses among men. One of the difficulties of PCa is grading that can be considered as a part of the classification problem. Therefore, accurate prediction of PCa grade is crucial to guarantee the quick treatment of malignancy [6]. Furthermore, early diagnosis and treatment planning can significantly reduce the mortality rate due to PCa [6], [7].

Technologies lead to having a crucial role in helping the medical community to diagnose cancer quickly [8]. On the one hand, there are many differences between images attained with modalities of analytic imaging and other image types that related to features and management of procedures. On the other hand, challenges are arising from the use of the different types of scanners, protocols of imaging, variety of noising, and other issues related to image attainment [5].

Different computer-aided techniques have been proposed using a radiomics method or deep learning network to accurately classify the PCa on magnetic resonance imaging (MRI) images [8]. Several studies have shown that computer-aided systems have a remarkable role in PCa detection and diagnostic evaluation. The methods proposed so far are based on handcrafted features, using a classifier on top to determine whether a PCa lesion is present or to assess its severity by assigning a specific class label. Recently, different techniques such as convolutional neural networks (CNN), support vector machine (SVM), iterative random (random forest [RF]), and J48 in the field of machine learning are proposed for locating and identifying cancer cells and normal cells. They have shown an impressive performance in various computer vision tasks following training with large image databases [5], [9].

This paper aims to propose a state-of-the-art review that surveys several techniques for PCa diagnosis, moreover, the techniques which are mostly based on machine learning are comparing in terms of performance accuracy.

The structure of the paper is as follows: In Section II, a review of some related works is represented while in Section III, the methodology of the literature review is described. Section IV shows a comparison among the aforementioned methods. Finally, a conclusion and future direction of the research survey are given in Section V.

2. SURVEY OF PCA TECHNIQUES

Several techniques have been suggested by many researchers for improving and developing PCa detection. In this survey, we mainly focus on the researcher's techniques that have been implemented with the machine learning field between 2015 and 2020.

Sammouda *et al*., 2015, worked on malignant PCa cells using near-infrared optical imaging technique that uses the high absorption of hemoglobin in PCa cells. Two algorithms (k-mean and fuzzy clustering mean) are used to segment and extract the cancer region in the prostate's infrared images. Using the Student's t-test to measure the accuracy between these two clusters, *P* value of K-means "cluster 3" is < 0.0001, and the standard error = 0.0002 is less than *P* value of ferric carboxymaltose (FCM)<0.0252 and standard error = 0.004. As the result, the K-mean is more accurate than FCM based on statistical analysis [10].

Mohapatra and Chakravarty, 2015, suggested a model using three classifiers SVM, Naive Bayes, and KNN to classify PCa. In this model, microarray is used as a dataset. The area under the curve (AUC) and accuracy have been measured to compare and evaluate the performances of these classifiers,

with taken the entire datasets and selected optimal features separately as the input to the classifiers one by one. As the result, the SVM technique performs more efficacy with higher accuracy of 95.5% [11].

In the same year, Bouazza *et al.*, 2015, proposed a classification method that performed a comparative study of four feature selection methods Fisher, T-Statistics, SNR, and ReliefF, using two classifiers K-nearest neighbors and SVM. Test results indicated that the best classification accuracy is obtained with SVM classifier and SNR method [12].

Dash *et al.*, 2016, worked on the microarray medical datasets and, two variations of kernel ridge regression (KRR) are used which are WKRR and RKRR to classify the datasets. To achieve a high rate of accuracy, this model is comparing the accuracy test among several techniques such as KRR, SVM-RBF, SVM-POLY, and RF. As the result, KRR (WKRR and RKRR) has been higher than all of them, especially RKRR which has an accuracy rate of 97%. However, this model has some drawbacks related to feature extraction which are ignoring the interaction with the classifier, features are considered independently which is mean ignoring these features which are dependencies. Another difficulty is related to determine the point of threshold to rank the features [13].

Imani *et al.*, 2016, proposed an approach to integrate mp-MRI with temporal ultrasound for PCa classification, in vivo. CNN technique has been utilized in this approach. A combination of mp-MRI and temporal ultrasound is used to reduce the missing regions of tumors. The AUC of 0.89 has been achieved for the classification of cancer with higher grades. Despite the importance of this model, there are some drawbacks because of the heterogeneous of PCa and it is difficult to determine tissue signature consistently [14].

Ram *et al.*, 2017, proposed an iterative RF (iRF) algorithm as a classifier model to separate cancer from the controlled samples of PCa. The method worked on microarray and next-generation sequencing (NGS) data. However, having a large number of gene expression data make it difficult of how to identify the biomarkers related to cancer. The RF has been used to select the genes which can diagnose and treat cancer effectively. RF method is used to extract very small sets of genes while it is taken predictive performance. Genes of *SNRPA1* are selected for PCa with the obtained accuracy of 73.33% [15].

Sun *et al.*, 2017, suggested a model investigate the performance of SVM algorithm and to predict the prostate tumor location using multiparametric MRI data. The capability of best predictive is achieved by optimizing model parameters using leave-one-out cross-validation. A binary SVM classifier utilizes to find a plane in feature space, frequently identified as a decision boundary, which splits the data into two parts. Furthermore, this algorithm is used to search for a decision boundary that maximizes the margin between the two groups. The final model gives results of classification by predicting the higher accuracy of 80.5%. However, only signal intensities and values from both T2-weighted (T2w) images and parametric maps are incorporated as features, respectively [16].

Liu and An, 2017, suggested a model based on deep learning and CNN for image classification of PCa, they used diffusion-weighted MRI (DWI) images that are selected images from a number of patients including positive and negative images. However, a small dataset makes a difficult for training a model that achieves higher accuracy. The proposed model has yielded an accuracy of 78.15% [17].

Reda *et al.*, 2018, presented a model using CNN based on computer-aided design (CAD) system for early diagnosis of PCa from DWI. They achieved accuracy rate of 95.65% [18].

Bhattacharjee *et al.*, 2019, developed a system for digitized histopathology images using a supervised learning method. SVM has been presented and used to classify malignant and benign PCa Grade 3, achieved accuracy was 88.7%. In SVM classification, 2-fold cross-validation has been used to train the model. Both of linear and Gaussian kernel are used for classifying samples as benign and malignant. Furthermore, a binary classification approach has been used which divides the multitype classification into two-category groups. Each partition characterizes distinct and independent classifications which are malignant and benign [3].

Yoo *et al.*, 2019, proposed a model of CAD system based on CNN and RF techniques for MRI (DWI) images. Five individually trained CNNs have been used to categorized DWI slices to extract the features and RF classifier has been used to classify patients into two groups patient with PCa and without PCa with achieved 0.84 as an AUC. The main limitation of this model is intrinsically biased which is mean these patients take MRIs who have symptoms of PCa. On the other hand, by depending on the reports of radiology, these slices with no biopsy consider as a negative sample while slices with biopsy consider as a positive sample based on pathology reports [19].

Cahyaningrum *et al*., 2020, proposed a method of artificial neural network (ANN) that optimized by genetic algorithm (GA) for PCa detection. This approach gives 76.4% of accurate detection. ANN has some limitations because of involving a huge number of parameters. Consequently, there have been many efforts to fix some of these limitations by joining ANN with another algorithm to address this problem. GA is an algorithm that is compatible and adapted with the ANN algorithm [20].

Besides, Duran-Lopez *et al*., 2020, presented a novel CAD system based on a deep learning algorithm (CNN) for distinguishing between malignant and normal tumors in whole slide images (WSIs). Cross-validation technique has been used with patches extracted from WS images. In this approach, the higher accuracy rate has achieved 99.98% [21].

Liu *et al*., 2020, stated a model of deep learning that integrates S-Mask R-CNN with Inception-v3 in ultrasound images to diagnose PCa. Furthermore, the AUC for Inception-v3 is 0.91. According to this model, there is a lot of traditional classifiers that can be used such as SVM and K-nearest neighbor. Due to minor variation between the ultrasonic PCa images and serious noise interface, some miss classifications might happen. Therefore, the CNN was presented in deep learning to achieve the best improvement of the classification accuracy in ultrasound images of the prostate without needing to describe the features manually and target image extraction [22].

3. METHODOLOGY

This review paper has conducted various studies in the field of PCa that is based on machine and deep learning techniques. First, the datasets that have been used by the researchers are described in subsection A. While in subsection B, the methodologies that are related to the performance accuracy are explained.

3.1. Datasets

Different datasets that have been used by researchers were investigated. Table 1 shows the modality of the dataset types that have been used by the authors in this survey. Moreover, sample numbers associated with each patient were given.

3.2. Performance parameter criteria

In the classification process, performance measurement is very important and essential, which determines the accuracy of the model. For this purpose, receiver operating characteristic (ROC) and AUC are proposed as effective evaluation metrics based classification model's performance.

TABLE 1: Datasets types with number of samples

Authors	Modality	No. of samples
Mohapatra and Chakravarty[11]	Microarray	136
Dash *et al*. [13]	Microarray	136
Ram *et al*. [15]	Microarray Gene Expression Omnibus (GSE71783)	30
Sun *et al*. [16]	MRI (T2w, DWI and DCE)	5
Liu and An [17]	MRI (DWI)	200
Reda *et al*. [18]	MRI (DWI)	23
Bhattacharjee *et al*. [3]	Microscopic tissue images	400
Cahyaningrum *et al*. [20]	Microarray gene expression	102
Duran-Lopez *et al*.[21]	WSI (whole slide image)	97

In statistics, a ROC curve can be defined as a graphical plot to illustrate the performance of a binary classification as it is used to distinguish varied thresholds. The true-positive rate (TPR) against the false-positive rate (FPR) is plotted to create the curve at varied threshold settings. In machine learning, the terms of recall, sensitivity, or detection probability have the same meaning as TPR. While, the term of fallout or false alarm probability has the same meaning as FPR and could be calculated as (1-specificity) [23]. Fig. 1 illustrates the relation between the ROC and AUC.

Furthermore, ROC is the probability curve whereas AUC is the degree of separable classes. ROC indicates that how much the model is capable of distinguishing amongst classes. Higher the AUC value (between 0 and 1) leads to better accuracy of the model.

This survey compares the techniques that are based on the accuracy of the proposed methods. Confusion matrix (CM) with performance metrics such as specificity and sensitivity is used to evaluate the proposed models [24]. The CM output could be either binary or multiclass. It has also a table of four different combinations between actual and predicted values. Predicted values are predicted by the model while actual values are actually in a dataset. Fig. 2 shows the CM relations.

The following formulas describe the performance accuracy metrics based on TP, TN, FP, and FN, according to CM.

TP – Values that are actually positive and predicted positive.

FP – Values that are actually negative but predicted to positive.

FN – Values that are actually positive but predicted to negative.

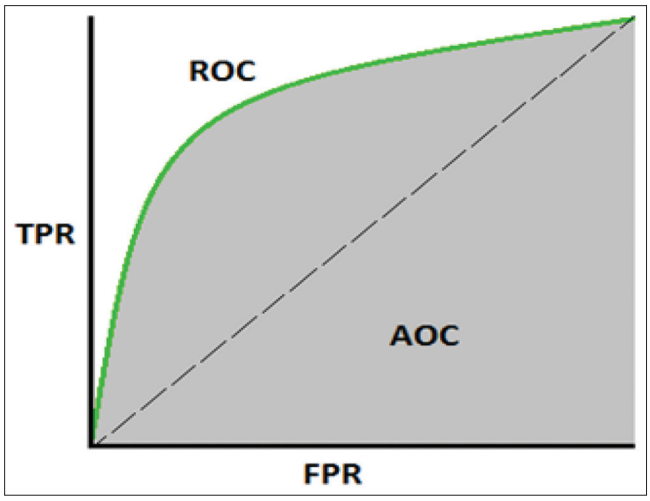

Fig. 1. Area under the curve-receiver operating characteristic curve.

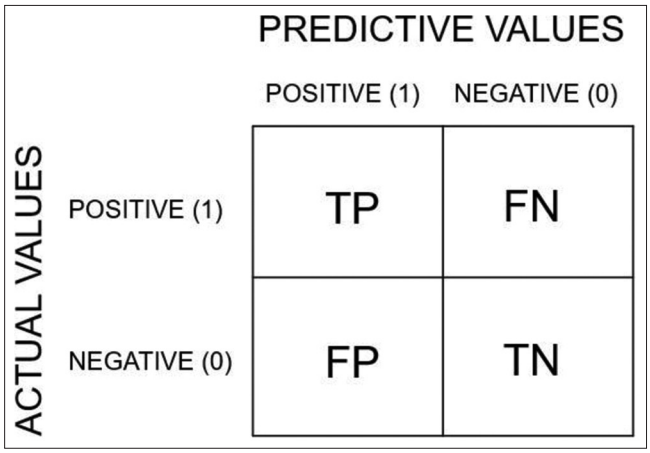

Fig. 2. Confusion matrix combinations.

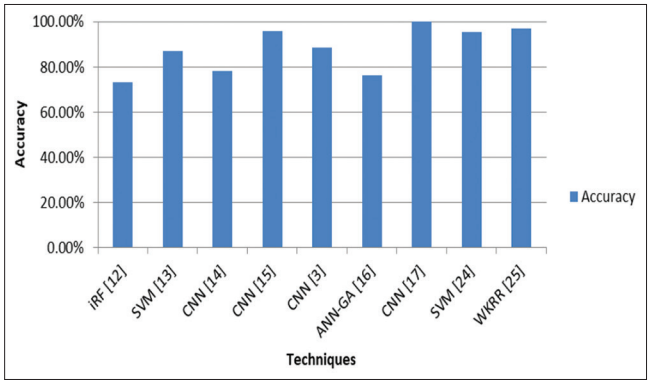

Fig. 3. Accuracy comparison of each technique.

TN – Values that are actually negative and predicted to negative.

$$\text{TPR (Sensitivity or recall)} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (2)$$

$$\text{FPR} = 1 - \text{Specificity} = \frac{FP}{FP + TN} \quad (3)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

TP and TN represent the number of correctly predicted positive and negative samples, while FP and FN are used to represent the number of incorrectly predicted positive and negative samples [25].

4. COMPARISON AND DISCUSSION

Many different techniques have been used by researchers. Each technique used a special type of dataset. Here, we compare the methods based on the accuracy with the dataset types and the year of publication, as shown in Table 2.

Fig. 3 shows the accuracy of the techniques separately.

Finally, an evaluation of five techniques based on the AUC has been performed to show the accuracy of the best technique, as depicted in Fig. 4.

As a result, according to the AUC measurements, the Inception-v3 classifier has the highest score for AUC, which

TABLE 2: Techniques with accuracy

S. No.	References	Year	Methods	Accuracy percent
1	Mohapatra and Chakravarty[11]	2015	SVM	95.5
2	Dash *et al.* [13]	2016	WKRR	97
3	Ram *et al.* [15]	2017	iRF	73.3
4	Sun *et al.* [16]	2017	SVM	80.5
5	Liu and An [17]	2017	CNN	78.15
6	Reda *et al.* [18]	2018	CNN	95.65
7	Bhattacharjee *et al.* [3]	2019	CNN	88.7
8	Cahyaningrum *et al.* [20]	2020	ANN-GA	76.4
9	Duran-Lopez *et al.* [21]	2020	CNN	99.98

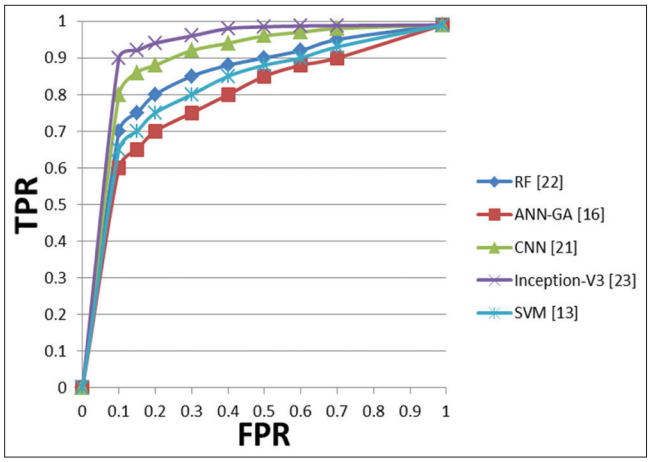

Fig. 4. Receiver operating characteristic curve for five classifiers.

is 0.91, although the type and quality of the dataset affect the ratio of the AUC scale.

5. CONCLUSION

This paper has introduced a comparison of classification methods based on machine learning techniques of the research related to PCa using various datasets including (Microarray, Microarray Gene Expression Omnibus (GSE71783), MRI (T2w, DWI, and DCE), microscopic tissue images, and WSI. In addition, the methods used in the literature have been reviewed along with the available results of the performance accuracy. The higher value of the AUC is identified amongst most five recent papers and it is 0.91.

REFERENCES

[1] Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker and F. Meriaudeau. "Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review". *Computers in Biology and Medicine*, vol. 60, pp. 8-31, 2015.

[2] T. Saba. "Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges". *Journal of Infection and Public Health*, vol. 13, no. 9, pp. 1274-1289, 2020.

[3] S. Bhattacharjee, H. G. Park, C. H. Kim, D. Prakash, N. Madusanka, J. H. So, N. H. Cho and H. K. Choi. "Quantitative analysis of benign and malignant tumors in histopathology: Predicting prostate cancer grading using SVM". *Applied Sciences*, vol. 9, no. 15, 2019.

[4] S. Liu, H. Zheng, Y. Feng and W. Li. "Prostate cancer diagnosis using deep learning with 3d multiparametric MRI". *SPIE Proceedings*, vol. 10134, pp. 3-6, 2017.

[5] N. Aldoj, S. Lukas, M. Dewey and T. Penzkofer. "Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network". *European Radiology*, vol. 30, no. 2, pp. 1243-1253, 2020.

[6] B. Abraham and M. S. Nair. "Automated grading of prostate cancer using convolutional neural network and ordinal class classifier". *Informatics in Medicine Unlocked*, vol. 17, p. 100256, 2019.

[7] L. A. Torre, B. Trabert, C. E. DeSantis, K. D. Miller, G. Samimi, C. D. Runowicz, M. M. Gaudet, A. Jemal, R. L. Siegel. "Ovarian cancer statistics, 2018". *CA: A Cancer Journal for Clinicians*, vol. 68, no. 4, pp. 284-296, 2018.

[8] M. Arif, I. G. Schoots, J. C. Tovar, C. H. Bangma, G. P. Krestin, M. J. Roobol, W. Niessen and J. F. Veenland. "Clinically significant prostate cancer detection and segmentation in low-risk patients using a convolutional neural network on multi-parametric MRI". *European Radiology*, vol. 30, pp. 6582-6592, 2020.

[9] L. Brunese, F. Mercaldo, A. Reginelli and A. Santone. "Formal methods for prostate cancer Gleason score and treatment prediction using radiomic biomarkers". *Magnetic Resonance Imaging*, vol. 66, pp. 165-175, 2020.

[10] R. Sammouda, H. Aboalsamh and F. Saeed. "*Comparison Between K Mean and fuzzy C-mean Methods for Segmentation of Near Infrared Fluorescent Image for Diagnosing Prostate Cancer*". International Conference on Computer Vision and Image Analysis Applications, 2015.

[11] P. Mohapatra and S. Chakravarty. "*Modified PSO Based Feature Selection for Microarray Data Classification*". 2015 IEEE Power, Communication and Information Technology Conference, pp. 703-709, 2015.

[12] S. H. Bouazza, N. Hamdi, A. Zeroual and K. Auhmani. "*Gene-expression-based Cancer Classification through Feature Selection with KNN and SVM Classifiers*". 2015 Intelligent Systems and Computer Vision, 2015.

[13] P. Mohapatra, S. Chakravarty and P. K. Dash. "Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system". *Swarm and Evolutionary Computation*, vol. 28, pp. 144-160, 2016.

[14] F. Imani, S. Ghavidel, P. Abolmaesumi, S. Khallaghi, E. Gibson, A. Khojaste, M. Gaed, M. Moussa, J. A. Gomez, C. Romagnoli, D. W. Cool, M. Bastian-Jordan, Z. Kassam, D. R. Siemens, M. Leveridge, S. Chang, A. Fenster, A. D. Ward and P. Mousavi. "*Fusion of Multi-parametric MRI and Temporal Ultrasound for Characterization of Prostate Cancer: In vivo Feasibility Study*". Medical Imaging 2016: Computer-Aided Diagnosis, vol. 9785, p. 97851K, 2016.

[15] M. Ram, A. Najafi and M. T. Shakeri. "Classification and biomarker genes selection for cancer gene expression data using random forest". *The Iranian Journal of Pathology*, vol. 12, no. 4, pp. 339-347, 2017.

[16] Y. Sun, H. Reynolds, D. Wraith, S. Williams, M. E. Finnegan, C. Mitchell, D. Murphy, M. A. Ebert and A. Haworth. "Predicting prostate tumour location from multiparametric MRI using Gaussian kernel support vector machines: A preliminary study". *Physical and Engineering Sciences in Medicine*, vol. 40, no. 1, pp. 39-49, 2017.

[17] Y. Liu and X. An. "*A Classification Model for the Prostate Cancer Based on Deep Learning*," Proceedings of the 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2017, pp. 1-6, 2018.

[18] I. Reda, A. Shalaby, M. Elmogy, A. A. Elfotouh, F. Kahalifa, M. A. El-Ghar, E. Hosseini-Asl, G. Gimel'farb, N. Werghi and A. El-Baz. "*A New CNN-Based System for Early Diagnosis of Prostate Cancer*". Proceedings International Symposium on Biomedical Imaging, pp. 207-210, 2018.

[19] S. Yoo, I. Gujrathi, M. A. Haider and F. Khalvati. "Prostate cancer

detection using deep convolutional neural networks". *Scientific Reports*, vol. 9, no. 1, pp. 1-10, 2019.

[20] K. Cahyaningrum, Adiwijaya and W. Astuti. "*Microarray Gene Expression Classification for Cancer Detection using Artificial Neural Networks and Genetic Algorithm Hybrid Intelligence*," 2020 International Conference on Data Science and its Applications, 2020.

[21] L. Duran-Lopez, J. P. Dominguez-Morales, A. F. Conde-Martin, S. Vicente-Diaz and A. Linares-Barranco. "PROMETEO: A CNN-based computer-aided diagnosis system for WSI prostate cancer detection". *IEEE Access*, vol. 8, pp. 128613-128628, 2020.

[22] . Liu, C. Yang, J. Huang, S. Liu, Y. Zhuo and X. Lu. "Deep learning framework based on integration of S-Mask R-CNN and Inception-v3 for ultrasound image-aided diagnosis of prostate cancer". *Future Generation Computer Systems*, vol. 114, pp. 358-367, 2021.

[23] A. Z. Shirazi, S. J. S. Mahdavi Chabok and Z. Mohammadi. "A novel and reliable computational intelligence system for breast cancer detection". Medical and Biological Engineering and Computing, vol. 56, no. 5, pp. 721-732, 2018.

[24] M. Nour, Z. Cömert and K. Polat. "A novel medical diagnosis model for COVID-19 infection detection based on deep features and bayesian optimization". *Applied Soft Computing*, vol. 97, pp. 1-13, 2020.

[25] Y. Celik, M. Talo, O. Yildirim, M. Karabatak and U. R. Acharya. "Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images". *Pattern Recognition Letters*, vol. 133, pp. 232-239, 2020.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Urban Rainwater Harvesting Assessment in Sulaimani Heights District, Sulaimani City, KRG, Iraq

Kani Namiq Gharib, Nawbahar Faraj Mustafa, Haveen Muhammed Rashid

Department of Water Resources, College of Engineering, University of Sulaimani, KRG, Iraq

ABSTRACT

Rainwater harvesting is the collection of rainwater and runoff from catchment areas such as roofs or other urban surfaces. Collected water has productive end-uses such as irrigation, industry, domestic, and can recharge groundwater. Sulaimani heights have been selected as a study area, which is located in Sulaimani Governorate in Kurdistan Region, North Iraq. The main objective of this study was to estimate the amount of harvested rainwater form Sulaimani heights urban area in Sulaimani City. Three methods for runoff calculation have been compared, the storm water management model (SWMM), the soil conservation service (SCS) method, and the runoff coefficient (RC) using daily rainfall data from 1991 to 2019. The annual harvested runoff results with the three different methods SWMM, SCS, and RC were estimated as 836,470 m^3, 508,454 m^3, and 737,381 m^3, respectively. The results showed that SWMM method has the highest runoff result and could meet 31% of the total demand of the study area and 28% and 19% for RC and SCS methods, respectively.

Index Terms: Rainwater harvesting, Storm water management model, Soil conservation service, Runoff coefficient, Runoff, Sulaimani heights

1. INTRODUCTION

Water is crucial for urban sustainability and in maintaining the sustainability of the environment [1], [2]. The extreme urbanization, industrial development, and agricultural expansion lead to increase demand of water in many parts of the world [3], [4]. Urban area development continuously reduces the groundwater recharging areas and increases depletion of groundwater [5].

Rainwater harvesting (RWH) is the collection and concentration of rainwater and runoff from catchment areas such as roofs or other urban structure and can be used for irrigation, industry, domestic, and for groundwater recharge purposes [6], [7], [8], [9]. RWH techniques have been used throughout time for the irrigation purpose by the ancient Iraqi people around 4500 BC [10], and it is an environmentally vocal decision to address issues brought out by large projects utilizing centralized water resources management approaches [11].

Many previous studies mentioned the use of RWH successfully as an effective and alternative water supply resolution [12], [13]. Patra and Gautam [4] conducted a study to assess the runoff coefficient (RC) method for RWH in Dhanbad city in India. The runoff results indicated that RH system is an economic option for where in the areas where rainfall is adequate and could supply part of the water demand of the city.

Zakaria *et al.*, 2013 [14], used Macro RWH at Koysinjaq (Koya), in Kurdistan Region based on Soil Conservation Service Curve Number (SCS-CN) method. The findings

Access this article online

DOI:10.21928/uhdjst.v5n1y2021.pp48-55 E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Nawbahar Faraj Mustafa, Department of Water Resources, College of Engineering, University of Sulaimani, KRG, Iraq. E-mail: nawbahar.mustafa@univsul,edu.iq

Received: 22-10-2020 Accepted: 25-04-2021 Published: 27-04-2021

demonstrated that the macro-RWH method can be a new source of water to reduce the problem of water scarcity and to minimize the water shortages problem.

In a research conducted by Harb [9], different RWH techniques were evaluated to identify the most significant method for METU-NCC campus on the west of North Cyprus. The runoff from roofs, pervious and impervious areas were collected and utilized and applying in two approaches: Traditional SCS method and storm water management model (SWMM) RWHS for calculation of runoff volume and the findings could meet 41.2% of the campus irrigation demand.

In 2016, a paper published by Gnecco *et al.* [15] in which SWMM was used to investigate the effect of domestic RWH and storage unit effects on control efficiency. The study area was located in neighborhood in Albaro in Italy, where covers 6000 m^2. The survey of the land use data displays that 57% of the land cover was impervious surfaces and 33% of rooftops of the total area. The findings of the software pointed that RWH can be applied in in urban water management and methods for assessment and optimization of runoff storage and use as potable water.

2. STUDY AREA

The Sulaimani heights are located in Sulaimani Governorate in Kurdistan Region, North Iraq. The latitudes are between 35°35' 55"and 35°36' 51" N and the longitudes are between 44°26'25" and 45°27'35" E. The area has a topographic with elevations ranged from 950 m to 1113 m. Sulaimani has a mean annual rainfall of 715 mm and has a mean daily temperature of 19℃ [16]. Sulaimani heights spread over an area of 2.12 km^2 and containing 2899 units of various sizes. The study area consists of three subcatchments, as shown in Fig. 1, and the detail information about each subcatchment is shown in Table 1. According to the map from Sulaimani heights authority (Qaiwan Company), the area is divided into five zones, the green areas cover 17.14 % and the water pools cover 1.1% of the total area as shown in Fig. 2.

3. MATERIALS AND METHODS

3.1. Data Sets and Data Collection

3.1.1. Climatology data

The daily precipitation data for Sulaimani city from 1991 to 2019 were used from Directorate of Meteorology and Seismology of Sulaimani (DOMSOS). As there is no rain gauge station in the studied basins, therefore the closest meteorological station should be used; Sulaimani rain gauge in Ibrahim Pasha Street which is only 4 km away from the studied area that has an acceptable distance. Daily rainfall data were used to represent the basin rainfall for the study area [17].

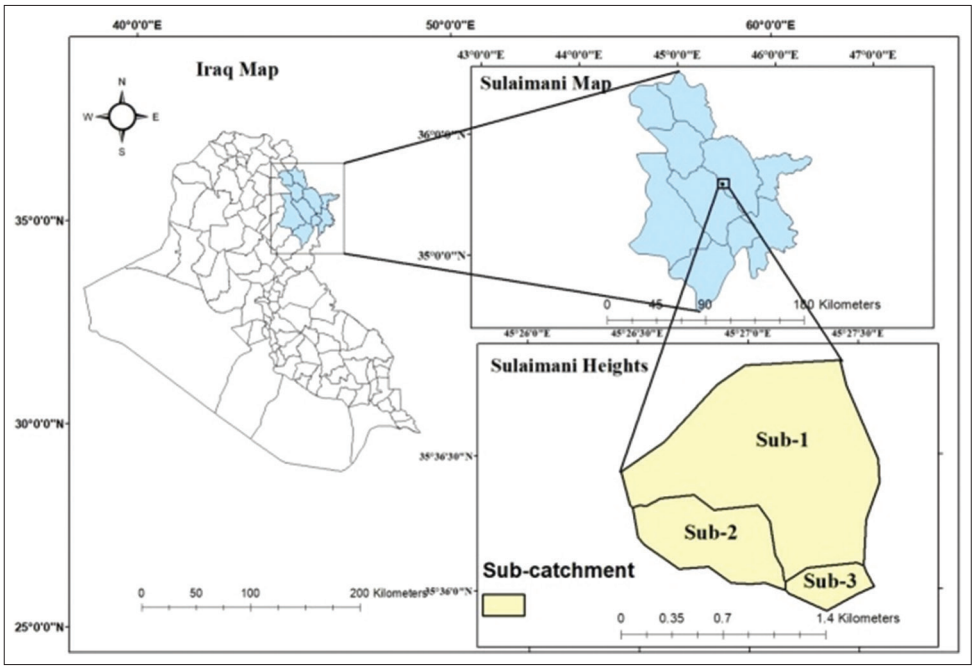

Fig. 1. Location of Sulaimani heights on Sulaimani map.

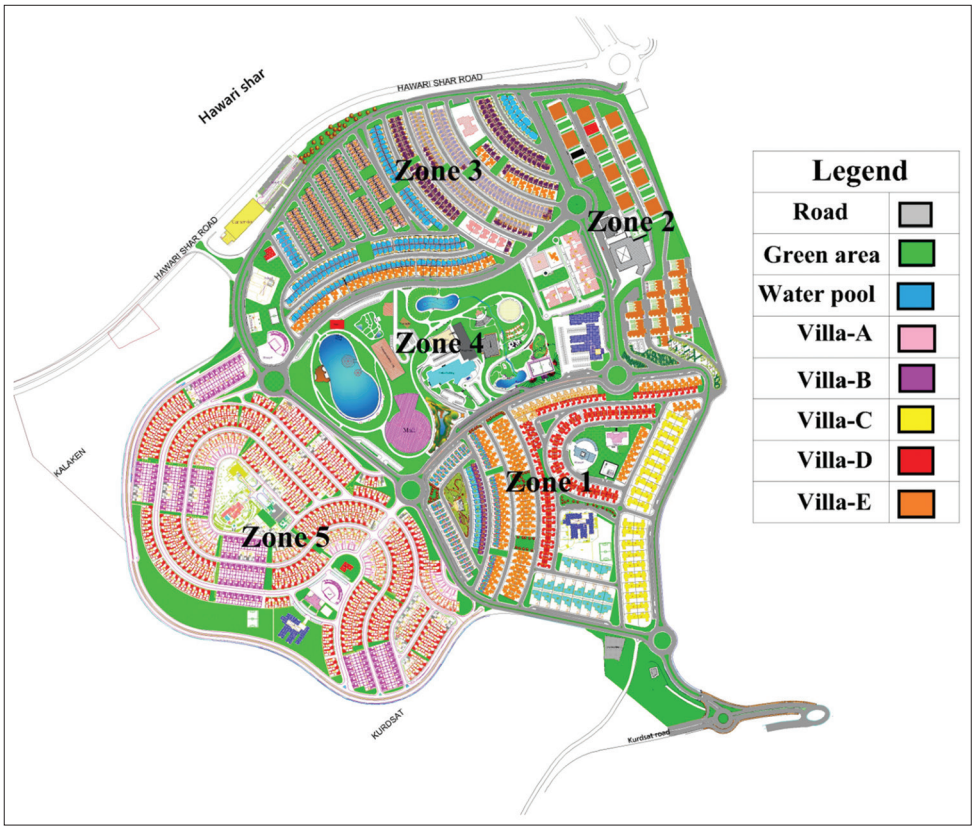

Fig. 2. Land use of Sulaimani heights.

TABLE 1: Detail information of the study area

Parameter	Value
Elevation – m	915–1113
Area km^2	2.12
Zone No.	5
Subcatchment No.	3
Residential No.	2899
Mean annual rainfall – mm	715
Mean daily temperature – °C	19

Other climatology data which have effect on the volume of the runoff should be also considered [9]; the monthly average wind speed and pan evaporation data from 1991 to 2019 were obtained from the Directorate of Meteorology and Seismology of Sulaimani.

3.1.2. Soil classification

To find the common soil characteristics of the study area, the Harmonized World Soil Database Viewer (HWSD) version 1.21 and soil map of the area were used. The software is adopted by cooperation of the Food and Agriculture Organization of the United Nations (FAO), the Chinese Academy of Sciences (CAS), the International Institute for Applied Systems Analysis (IIASA), the International Soil Reference and Information Centre (ISRIC), and the Joint Centre of the European Commission (JRC). The coordinates of the study area were pointed on the HWSD viewer software and the dominant soil group was found to be Chromic Vertisols with 100% light clay to be the most prominent soil textures. Therefore, the dominant soil texture is clay and hence this satisfies the Hydrologic Soil Group D [18].

3.2. SWMM

SWMM is widely utilized software throughout the world in associating of urban runoff quantity and quality [19]. SWMM is a rainfall-runoff simulation model developed by the US Environmental Protection Agency to assist and support local storm water management in minimizing the runoff discharges. SWMM can forecast a single event or long-term (continuous) simulation set of model outputs parameters and inputs of runoff quantity and quality from primarily urban areas [20], [21], [22], [23], [24].

In accordance with the subcatchment properties, the average monthly surface runoff can be calculated through SWMM

software, to estimate monthly results from the SWMM software, the dates of simulation should be manipulated from the Options tab and the software run gives runoff depth, infiltration depth, and runoff volumes in the form of a table.

SWMM uses the Manning equation to express the relationship between flow rate (Q), cross-sectional area (A), hydraulic radius (R), and slope (S) in all conduits [21], [25].

For standard S.I units:

$$Q = \frac{1}{n} A R^{\frac{2}{3}} S^{\frac{1}{2}} \quad (1)$$

Where, n is the manning roughness coefficient. The slope S stands for either the conduit slope or the friction slope (i.e. head loss per unit length), depending on the flow routing method used. The R is hydraulic radius, which is fraction of area to wetted parameter of the conduit or the channel.

3.3. Traditional SCS Method

SCS method is another suitable method for this case, as it includes all types of abstractions in the runoff calculation and the parameters needed for runoff estimation. The runoff volumes will be estimated based on SCS-CN method. CN method is thoroughly used for estimating direct runoff volume for a particular rainfall event [26], [27]. For the SCS, 1972 (SCS-CN) method, the CN(I) stands for dry condition, CN(III) stands for wet condition and tabulated CN is equal to CN(II), for normal (average) conditions, and can be modified for dry and wet conditions, as explained by Chow *et al.* [28] through the following Equations 2 and 3 [29]:

$$CN(I) = \frac{4.2 * CN_{II}}{10 - 0.058 * CN_{II}} \quad (2)$$

$$CN(III) = \frac{23 * CN_{II}}{10 + 0.13 * CN_{II}} \quad (3)$$

The expression used in SCS method for estimating runoff can be calculated through Equation 4 [18]:

$$Q = \frac{(P - I_a)^2}{\{(P - I_a) + S\}} \quad (4)$$

Where, Q is the accumulated storm runoff in (mm); P is accumulated storm rainfall in (mm), S is potential maximum retention of water by the soil, I_a is initial quantity of interception, infiltration, and depression which can be quantified through Equation 5.

$$S = \frac{25400}{CN} - 254 \quad (5)$$

While the data needed to calculate the runoff volume are present, SCS method is also used to compute the runoff volume. The runoff depth from subcatchments is calculated using CN and rainfall depth. After the runoff depth is calculated, the volume of the runoff from each subcatchment is computed by multiplying the area of each subcatchment as shown in Equation 6.

$$V = R*A \quad (6)$$

Where, V is the volume of runoff (m^3); R is the rainfall-runoff (m); and A is the area of the subcatchment (m^2).

3.4. Assessment of RC

RC for any catchment is the ratio of the volume of water that runs off a surface to the volume of rainfall that falls on the surface [30].

The Rc takes into account any losses due to evaporation, leakage, surface material texture overflow, transportation, and inefficiencies in the collection process [17]. The RWH potential or volume of water received from a given catchment can be obtained using the following Equation 7 [17].

$$V_r = RA_c R_c \quad (7)$$

Where, V_r the monthly volume of rainwater, R is average monthly rainfall depth, A_c is area of the catchment, and R_c runoff coefficient.

To calculate the monthly runoff produced for each subcatchment, RCs (Equation 7) is used. The average R_C for the different types of areas was selected [31], for the areas of constructed concrete and asphalt, the R_C was selected as 0.65, 0.075 for green area, and 0.9 for water bodies.

The flowchart in Fig. 3 shows the steps followed for the calculation of runoff using the mentioned three methods.

3.5. Water Demand

To determine domestic water demand for indoor and outdoor household purposes, the standard average daily water demand per capita (Sulaimani water supply directorate) which is (250 l/capita/day) is used to calculate the average monthly demand for the study area [32]. Harvested rainwater should be treated before using for drinking purpose [33]. In accordance with the study area, there are 2899 residential and the average of 5 members in a household counted to estimate the total water demand for the study area. The total population of the study area calculated and the total daily water demand found for Sulaimani Heights. Using the map of the study area, the areas for each type of vegetation group for the study area

are calculated in AutoCAD and the irrigation months were selected with the amount of water for each m^2 of vegetation area. Thus, the monthly demands for the study area calculated by multiplying the number of days of the month and then by the total daily demand. The results of water demands are shown in Tables 2-5.

4. RESULTS AND DISCUSSION

The results from the three models are shown in Table 6 which shows that the SWMM method has the largest annual runoff volume of 836,470 m^3, Rc method results with 737,381 m^3 and SCS method with 508,454 m^3 for the average annual rainfall of 719 mm. Table 7 and Table 8 represent the monthly and annual water demand, respectively, with the corresponding percent demand met.

The results showed that SWMM method has the highest runoff result and could meet 31% of the total demand of the study area and 28% and 19% for Rc and SCS methods, respectively. Comparison between respective runoff results clearly demonstrates that the runoff results are influencing by the serial urbanization [34].

TABLE 2: Total daily domestic water demand in Sulaimani heights

Sub. No.	No. of residence	Population	Water demand (m^3/day)
1	2541	12,705	3176
2	358	1790	448
3	0	0	0
Total	2899	14,495	3624

TABLE 3: Vegetation area and type of each subcatchment

Subcatchment	Green area (m^2)	Crop type	
		Ground cover area (m^2)	Trees and bushes area (m^2)
Sub-1	447,182	302,080	145,102
Sub-2	50,087	42,070	8017
Sub-3	127,650	108,400	19,250
Total area (m^2)	624,919	452,550	172,369

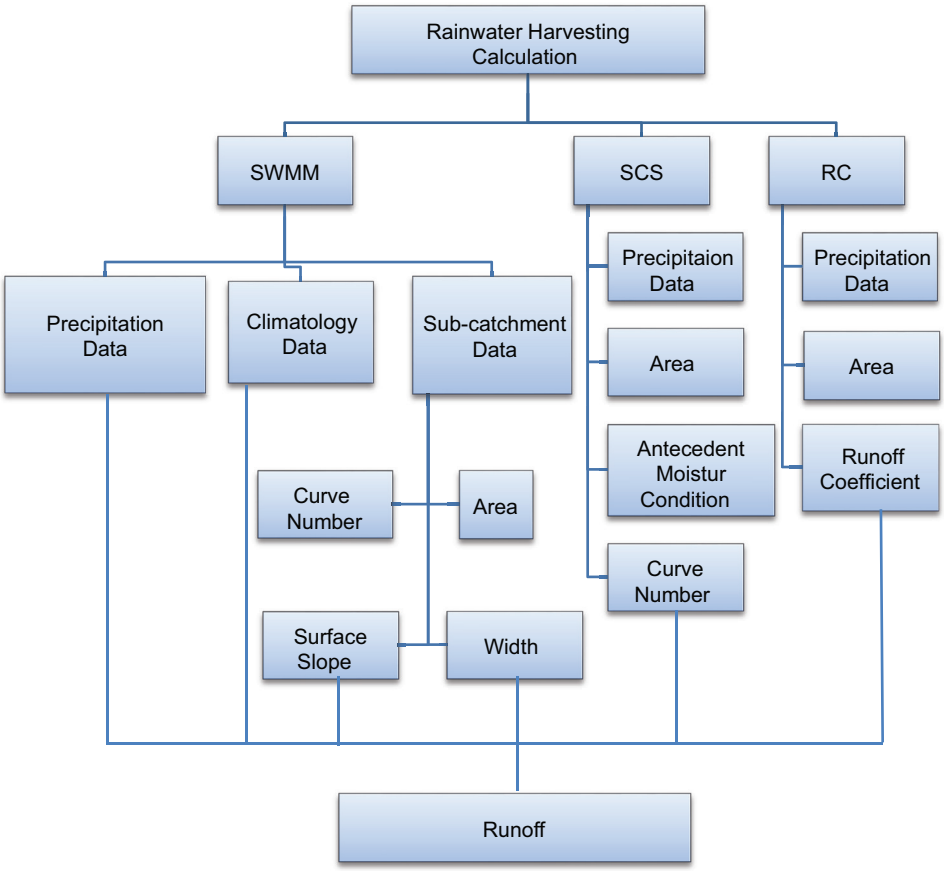

Fig. 3. Flowchart summary for runoff calculation methods.

TABLE 4: Total monthly irrigation water demand in Sulaimani heights

Crop type	Month of irrigation	Irrigation period (day/month)	Required water per (m^2) (L/day)	Water demand (m^2/month)		
				Sub-1	Sub-2	Sub-3
Ground cover	May	15	12	54,374.40	7572.60	19,512.00
	June	30	16	144,998.40	20,193.60	52,032.00
	July	31	16	149,831.68	20,866.72	53,766.40
	August	31	16	149,831.68	20,866.72	53,766.40
	September	30	16	144,998.40	20,193.60	52,032.00
	October	15	12	54,374.40	7572.60	19,512.00
Trees and bushes	May	15	8	17,412.24	962.04	2310.00
	June	30	12	52,236.72	2886.12	6930.00
	July	31	12	53,977.94	2982.32	7161.00
	August	31	12	53,977.94	2982.32	7161.00
	September	30	12	52,236.72	2886.12	6930.00
	October	15	8	17,412.24	962.04	2310.00

TABLE 5: Total demand in the three subcatchments

Month	No. of days	Water demand (m^2/month)			Total water demand (m^3/month)
		Sub-1	Sub-2	Sub-3	
January	31	98,456	13,888	0	112,344
February	28	88,928	12,544	0	101,472
March	31	98,456	13,888	0	112,344
April	30	95,280	13,440	0	108,720
May	31	170,242.6	22,422.64	21,822	214,487.28
June	30	292,515.1	36,519.72	58,962	387,996.84
July	31	302,265.6	37,737.04	60,927.4	400,930.06
August	31	302,265.6	37,737.04	60,927.4	400,930.06
September	30	292,515.1	36,519.72	58,962	387,996.84
October	31	170,242.6	22,422.64	21,822	214,487.28
November	30	95,280	13,440	0	108,720
December	31	98,456	13,888	0	112,344
Total	365	2,104,903	274,447	283,423	2,662,772

TABLE 6: The runoff volume results of the three methods

Month	Sum of average monthly rainfall (mm)	Volume of runoff by SWMM (m^3/month)	Volume of runoff by SCS (m^3/month)	Volume of runoff by Rc (m^3/month)
January	119.43	150,040	76,912.01	122,350.48
February	116.84	149,720	90,680.95	119,697.14
March	105.11	124,320	69,667.31	107,680.31
April	96.53	111,190	53,480.01	98,890.49
May	41.84	33,620	25,415.88	42,863.13
June	0	0	0	0
July	0	0	0	0
August	0	0	0	0
September	0	0	0	0
October	44.43	41,920	53,829.71	45,516.47
November	81.84	87,510	58,195.91	83,841.27
December	113.76	138,150	80,271.92	116,541.83
Total	719.78	836,470	508,453.71	737,381

SWMM: Storm water management model, SCS: Soil conservation service, RC: Runoff coefficient

From the methods discussed previously, it appears that the traditional SCS method and assessment of RC are respectable to be a combined with more losses method since the initial abstraction includes infiltration, evaporation, interception, and surface texture caused by these processes are calculated simultaneously [9], [35].

TABLE 7: Monthly water demand and corresponding percent demand met

Month	Percent of water demand met using SWMM (%)			Percent of water demand met using SCS (%)			Percent of water demand met sing RC (%)		
	Sub-1	Sub-2	Sub-3	Sub-1	Sub-2	Sub-3	Sub-3	Sub-2	Sub-3
October	18	44	5	23	49	16	20	50	2
November	68	100	100	42	100	100	65	100	100
December	100	100	100	57	100	100	88	100	100
January	100	100	100	54	100	100	92	100	100
February	100	100	100	71	100	100	100	100	100
March	94	100	100	49	100	100	81	100	100
April	86	100	100	39	94	100	77	100	100
May	15	36	4	10	27	8	19	47	2
June	0	0	0	0	0	0	0	0	0
July	0	0	0	0	0	0	0	0	0
August	0	0	0	0	0	0	0	0	0
September	0	0	0	0	0	0	0	0	0

SWMM: Storm water management model, SCS: Soil conservation service, RC: Runoff coefficient

TABLE 8: Annual water demand and corresponding percent demand met

Method	SWMM	SCS	RC
Runoff (m^3)	836,470	508,454	737,381
Annual demand (m^3)	2,662,772	2,662,772	2,662,772
Annual demand met %	31	19	28

SWMM: Storm water management model, SCS: Soil conservation service, RC: Runoff coefficient

In addition, in the SCS and Rc methods, the infiltration in the initial abstraction does not change with rainfall events variation on a subcatchment, conversely, it would stay the same before and during the rainfall event [9], [35]. Some parameters implied in the SWMM model but not computed in the SCS such as the depression storage, percent of impervious layer, the pervious roughness coefficient, and the soil drying time [9], [21], [24].

On the other hand, the SWMM model has flexibility to route runoff and external inflows through the drainage systems, and the abstractions such as evaporation and infiltration vary with changing rainfall events [20]. Due to these limitations, the SWMM model is established in the prediction of comparable runoffs [36].

The SWMM differs from the SCS and Rc approaches by that the SWMM model can perform helpful and time saver tool in designing large catchments and SWMM has better feasibility of determining peak flow and volume of runoff with in the nodes and pipes for designing urban drainage system [21], [24].

5. CONCLUSION

This research studied the feasibility of applying RWH techniques as a water resource that should be associated into the management of urban areas. RWH for different types of catchments such as roofs, roads, and open areas has been founded. Three approaches for runoff calculation were adopted, the SWMM, the traditional SCS method, and the RC. Daily rainfall data from 1991 to 2019 were used to obtain the monthly and annual volume. Moreover, to demonstrate the potential RWH system, the annual demand for the study area was found and compared with the total annual runoff volume using three methods, however, harvested rainwater harvested should be treated before using for drinking purpose.

For the estimated total yearly water demand in the study area of demand in the study area of 2,662,772 m^3, the annual runoff results with the methods SWMM, SCS, and Rc were estimated of 836,470 m^3, 508,454 m^3 and 737,381 m^3 respectively. The final results showed that SWMM method has the highest runoff result and could meet 31% of the total demand of the study area and 28% and 19% for Rc and SCS methods, respectively.

6. ACKNOWLEDGMENT

I truly want to thank Qaiwan Group for assistance in complimenting this study and my family for their support and guidance.

REFERENCES

[1] A. Daoud, K. Swaileh, R. M. Hussein and M. Matani. "Quality Assessment of Roof-harvested Rainwater in the West Bank, Palestinian Authority". *Journal of Water and Health*, vol. 9, pp. 525-533, 2011.

[2] T. M. Pinzón. "*Modelling and Sustainable Management of Rainwater Harvesting in Urban Systems*". Theses, 2013.

[3] U. WWDR. *Water and Energy, the United Nations World Water Development Report* 2014 (*2 volumes*). UN World Water Assessment Programme, Unesco, Paris. Available from: https://sustainabledevelopment.un.org/content/documents/1714Water%20Development%20Report%202014.pdf. [Last accessed on 2019 Mar 19].

[4] A. K. Patra and S. Gautam. "A pilot scheme for rooftop rainwater harvesting at Centre of Mining Environment, Dhanbad". *International Journal of Environmental Sciences*, vol. 1, pp. 1542-1548, 2011.

[5] S. N. Baby, C. Arrowsmith and N. Al-Ansari. "Application of GIS for mapping rainwater-harvesting potential: Case study Wollert, Victoria". *Engineering*, vol. 11, pp. 14-21, 2019.

[6] K. Subagyono and H. Pawitan. "*Water Harvesting Techniques for Sustainable Water Resources Management in the Catchment Area*". In: Proceedings of International Workshop on Integrated Watershed Management for Sustainable Water Use in a Humid Tropical Region, Tsukuba, 2008, pp. 18-30.

[7] D. Prinz, T. Oweis and A. Hachum. "*The Concept, Components, and Methods of Rainwater Harvesting*". In: 2nd Arab Water Forum Living With Water Scarcity, Cairo, 2011, pp. 1-25.

[8] C. Bari. "*Emerging Practices from Agricultural Water Management in Africa and the Near East*". Thematic Workshop, 2017.

[9] R. Harb. "*Assessing the Potential of Rainwater Harvesting System at the Middle East Technical University-Northern Cyprus Campus*". Middle East Technical University Library. Available from: http://www.etd.lib.metu.edu.tr/upload/12619225/index.pdf. [Last accessed on 2016 Nov 10].

[10] R. H. Handbook. "*Assessment of Best Practises and Experience in Water Harvesting*". African Development Bank, Abidjan, 2001.

[11] J. Julius, R. A. Prabhavathy and G. Ravikumar. "Rainwater harvesting (RWH)-A review". *International Journal of Innovative Research and Development*, vol. 2, p. 925, 2013.

[12] G. Freni and L. Liuzzo. "Effectiveness of rainwater harvesting systems for flood reduction in residential urban areas". *Water*, vol. 11, p. 1389, 2019.

[13] I. A. Alwan, N. A. Aziz and M. N. Hamoodi. "Potential water harvesting sites identification using spatial multi-criteria evaluation in Maysan Province, Iraq". *ISPRS International Journal of Geo-Information*, vol. 9, p. 235, 2020.

[14] S. Zakaria, N. Al-Ansari, Y. Mustafa, S. Knutsson, P. Ahmed and B. Ghafour. "Rainwater harvesting at Koysinjaq (Koya), Kurdistan Region, Iraq". *Journal of Earth Sciences and Geotechnical Engineering*, vol. 3, pp. 25-46, 2013.

[15] I. Gnecco, A. Palla and P. La Barbera. "The role of domestic rainwater harvesting systems in storm water runoff mitigation". *Eur Water*, vol. 58, pp. 497-503, 2017.

[16] N. F. Mustafa, H. M. Rashid and H. M. Ibrahim. "Aridity Index Based on Temperature and Rainfall Data for Kurdistan Region-Iraq". *Journal of Duhok University*, vol. 21, pp. 65-80, 2018.

[17] J. Worm. "*AD43E Rainwater Harvesting for Domestic Use*". Agromisa Foundation, 2006.

[18] Z. Ara and M. Zakwan. "Estimating runoff using SCS curve number method". *International Journal of Emerging Technology and Advanced Engineering*, vol. 8, pp. 195-200, 2018.

[19] J. Gironás, L. A. Roesner, L. A. Rossman and J. Davis. "A new applications manual for the Storm Water Management Model (SWMM)". *Environmental Modelling and Software*, vol. 25, pp. 813-814, 2010.

[20] W. R. C. James and L. A. Rossman. "*User's Guide to SWMM 5*". Computational Hydraulics International, 2010.

[21] L. A. Rossman. "*Storm Water Management Model User's Manual, Version 5.0*". National Risk Management Research Laboratory, Cincinnati, 2010.

[22] J. Nipper. "*Measurement and Modeling of Stormwater from Small Suburban Watersheds in Vermont*". Theses, 2016.

[23] S. Agarwal and S. Kumar. "Applicability of SWMM for semi urban catchment flood modeling using extreme rainfall events". *The International Journal of Recent Technology and Engineering*, vol. 8, pp. 245-251, 2019.

[24] M. Waikar and U. Namita. "Urban flood modeling by using EPA SWMM 5". *SRTM University's Research Journal of Science*, vol. 1, p. 20, 2015.

[25] H. Tikkanen. "*Hydrological Modeling of a Large Urban Catchment Using a Stormwater Management Model* (*SWMM*), Thesis, 2013.

[26] R. H. Hawkins, T. J. Ward, D. E. Woodward and J. A. Van Mullem. "*Curve Number Hydrology: State of the Practice*". American Society of Civil Engineers, 2008.

[27] A. Bansode and K. Patil. "Estimation of runoff by using SCS curve number method and arc GIS". *International Journal of Scientific and Engineering Research*, vol. 5, pp. 1283-1287, 2014.

[28] V. Chow, D. Maidment and W. L. Mays. "*Applied Hydrology*". MacGraw-Hill," Inc., New York, p. 149, 1988.

[29] N. Al-Ansari, S. Knutsson, S. Zakaria and M. Ezz-Aldeen. "*Feasibility of Using Small Dams in Water Harvesting, Northern Iraq*". In: ICOLD Congress 2015: International Commission on Large Dams 15/06/2015-16/07/2015, 2015.

[30] M. Awawdeh, S. Al-Shraideh, K. Al-Qudah and R. Jaradat. "Rainwater harvesting assessment for a small size urban area in Jordan". *International Journal of Water Resources and Environmental Engineering*, vol. 4, pp. 415-422, 2012.

[31] G. Dadhich and P. Mathur. "A GIS based analysis for rooftop rain water harvesting". *International Journal of Computer Science and Engineering Technology*, vol. 7, pp. 129-143, 2016.

[32] K. N. Sharief. "*Water Supply Management of the Sulaymaniyah City*". Unpublished PhD Thesis, University of Duhok, p. 24, 2013.

[33] C. A. Novak, E. Van Giesen and K. M. DeBusk. "*Designing Rainwater Harvesting Systems: Integrating Rainwater into Building Systems*". John Wiley and Sons, Hoboken, New Jersey, 2014.

[34] R. N. Eli and S. J. Lamont. "*Curve Numbers and Urban Runoff Modeling Application Limitations*". In: Low Impact Development 2010: Redefining Water in the City, pp. 405-418, 2010.

[35] D. A. Size. "*Computing Stormwater Runoff Rates and Volumes*". New Jersey Storm water Best Management Practices, pp. 5-22, 2004.

[36] G. R. Ghimire, R. Thakali, A. Kalra and S. Ahmad. "*Role of Low Impact Development in the Attenuation of Flood Flows in Urban Areas*". World Environmental and Water Resources Congress, pp. 339-349, 2016.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Kurdish Text Segmentation using Projection-Based Approaches

Tofiq Ahmed Tofiq, Jamal Ali Hussien

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

ABSTRACT

An optical character recognition (OCR) system may be the solution to data entry problems for saving the printed document as a soft copy of them. Therefore, OCR systems are being developed for all languages, and Kurdish is no exception. Kurdish is one of the languages that present special challenges to OCR. The main challenge of Kurdish is that it is mostly cursive. Therefore, a segmentation process must be able to specify the beginning and end of the characters. This step is important for character recognition. This paper presents an algorithm for Kurdish character segmentation. The proposed algorithm uses the projection-based approach concepts to separate lines, words, and characters. The algorithm works through the vertical projection of a word and then identifies the splitting areas of the word characters. Then, a post-processing stage is used to handle the over-segmentation problems that occur in the initial segmentation stage. The proposed method is tested using a data set consisting of images of texts that vary in font size, type, and style of more than 63,000 characters. The experiments show that the proposed algorithm can segment Kurdish words with an average accuracy of 98.6%.

Index Terms: Optical character recognition, Character segmentation, Kurdish text segmentation, Projection-based approach, and Cursive writing optical character recognition

1. INTRODUCTION

Optical character recognition (OCR) is an application for image recognition that can read text from images. This can be achieved by taking an image that includes text, written in a specific language to be understood by the computer, and get the final computer representation for the text. OCR techniques may vary according to the nature of the language and the purpose of the OCR application [1]. Emulating the human ability in associating symbolic identities with images of characters at a fast rate is the main goal of OCR.

Kurdish is one of the languages that present many challenges to OCR. The main challenge in Kurdish is that it is mostly cursive. Kurdish is written by connecting characters together to produce words or parts of words, as shown in Fig. 1. Kurdish text is written from right to left. The Kurdish language has 34 basic characters, of which 14 have from one to three dots and four of them have diacritics.

Kurdish characters have many shapes and sizes depending on their position in the word. For example, the character "ک" is written in the form of "ک" at the start, "ﻜ" in the middle, and "ﮏ" at the end of a word but the separated form of this character is "ﮏ." Kurdish characters vary with relevance to their position in the word, representing a great challenge for OCR [1].

Based on the nature of Kurdish fonts, characters may overlap vertically to produce certain compounds of letters at certain positions of the word segments, such as "لا, حمـ and نج," which can be represented by single atomic graphemes called ligatures [3], [4].

Access this article online

DOI: 10.21928/uhdjst.v5n1y2021.pp56-65 | E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Tofiq Ahmed Tofiq, Department of Computer Science, College of Science, University of Sulaimani, Sulaimani, Iraq. E-mail: Tofiq.ahmad@univsul.edu.iq

Received: 23-01-2021 Accepted: 12-05-2021 Published: 16-05-2021

Fig. 1. The character connectivity of Kurdish text.

Some Kurdish letters have single dots such as ن, ج and ب, other letters have double dots, such as ش and ق and other have triple dots, such as ش and ڤ. Furthermore, some letters have diacritics such as ۆ, ڵ, and ێ. Besides, the same shape with different dots and diacritics is used to represent different letters, such as, ح, ج, چ, ل,ڵ, ۆ, and و. The doted characters and letters with diacritics present a big challenge while being processed.

This paper presents a Kurdish text segmentation algorithm. The proposed algorithm uses the projection-based approach concepts to separate lines, words, and letters.

The rest of the paper is organized as follows: Section 2 segmentation-based methods, section 3 related works with different segmentation techniques. Section 4 presents the proposed algorithm. Section 5 demonstrates the results and performance analysis. Section 6 concludes this paper.

2. SEGMENTATION-BASED METHODS

In this part, the proposed algorithm for the segmentation of cursive text such as Arabic, Persian, and English handwriting text is discussed. The segmentation-based methods can be classified into:

a. Projection profile
b. Character skeleton based
c. Contour tracing based
d. Template matching based
e. Morphological operations based
f. Segmentation based on neural networks (NNs) [2], [5].

Projection profiles methods [9]-[11] are usually used to aim for lines, words, sub-words, and characters segmentation specifically when there is a certain gap between lines, words, sub-words, and characters. Indeed, horizontal projection is used for line segmentation, while vertical projection is usually used for word, sub-word, and character segmentation. In the skeleton method, different thinning techniques are engaged for this goal [7], [12]. In many cases, the shape of the characters is different from the main character after thinning, making the splitting process more difficult. In contour tracking [13]-[16] methods, pixels that represent the external shape of a character or word are extracted. Researchers used many methods to determine the cut points in the contour. In general, the contour-based technique avoids the issues that seem once applying to the thinning methods because they depend on extracting the structure of the word, which provides an obvious description of it. This type of method is sensitive to noise, which needs one to perform some preprocessing steps. Morphological methods [17]-[19] use a set of morphological operations for segmentation. In general, closing and opening operations are applied. These methods are dependent, meaning that other techniques must be used in addition to segmentation. Template matching methods [20], [21] often apply a sliding window over baselines. If any match is found, then the center pixel in the sliding window is considered as a cutting point. The main limitation of this method is when the cutting point locates under the baseline. Finally, in NNs segmentation, NNs are used to verify the valid segmentation points by training the NNs over manually classified valid segmentation points from the database of scanned images using features such as black pixel density and holes [22].

3. RELATED WORKS

Zheng *et al.* [10] proposed a machine-printed Arabic character segmentation algorithm that uses a vertical projection method and some rules or features, such as structural characteristics between background area and character components and the specification of isolated Arabic characters to find segmentation points.

Cheung *et al.* [6] proposed a segmentation algorithm that uses a technique wherein the overlapping Arabic words/sub-words are horizontally separated, extensively utilizing a feedback loop among the character segmentation and final recognition stages. In the segmentation stage, a series of experimental lines have been produced in two processes, the first process uses Amin's character segmentation algorithm [21] and the second procedure use the convex dominant points detection algorithm developed by Bennamoun and Boashash [23].

Shaikh *et al.* [7] propound an algorithm for Sindhi text segmentation. The height profile vector (HPV) was used for the character extraction. More analysis was done over HPV to detect the locations of the possible segmentation points (PSPs), in some cases, the algorithm failed by performing under or over-segmentation.

Yeganeh *et al.* [13] introduced a segmentation algorithm for up and down contours based on qualified labeling, and the algorithm was developed for multifont Farsi/Arabic

texts. The contour of the sub-word is measured using a convolution kernel with a Laplacian edge recognition-based segmentation detection method. The algorithm goes through several stages including contour labeling of each sub-word, contour curvature grouping to improve the segmentation results, character segmentation, adaptive local baseline, and post-processing, the results showed that 97% of characters of the printed Farsi texts were segmented correctly.

Mostafa [24] proposed a segmentation method for printed Arabic text, especially for "simplified Arabic" font with different sizes. Most characters start with and end before a T-junction on the baseline, that is, the main rule used in this. This rule was fine for most characters, except for some special characters such as "س" and "ش," which had a special solution. The algorithm was tested and achieved a 96.5% of good segmentation accuracy.

Alipour [8] presents an improved segmentation technique for Persian text where a few structural features were used to regulate the relevant segment to increase the quality of segmentation. The vertical projection was used to bring out the word segment over the baseline dots and diacritics were not checkout then the segment was regulated in an additional step by merging the small fragments, this step was needful

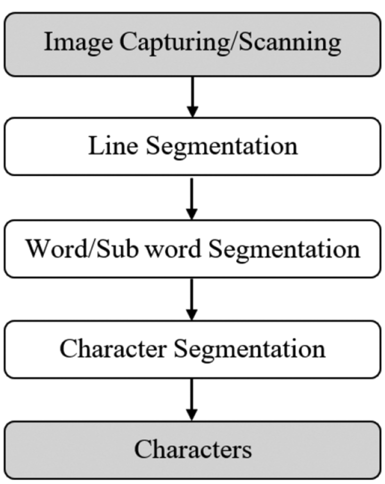

Fig. 2. The major steps of the proposed segmentation approach.

in the cases where one character is isolated into more than one segment such as "س" and "ش."

4. THE PROPOSED ALGORITHM

Our technique is a segmentation-based approach, which contains three main segmentation stages, as shown in Fig. 2. The proposed method takes a binary image that has multiple lines of text and executes several image processing methods to finally segment characters in the image. In this method, the segmentation is performed at three levels: Line segmentation, word/sub-word segmentation, and character segmentation. This work focused only on the line, word/sub-word, and character segmentation steps, considering that the input image has been preprocessed (by applying operations such as binarization, noise removal, and separating text from non-text regions). The image binarization used the below equation.

$$g(x,y)=\begin{cases} \text{üü}|\text{if} \; x \; y < T \\ \text{üü}|\text{if} \; x \; y \geq T \end{cases}$$

Where, (x,y) is the coordinate of the pixels, T represents the threshold value, g(x,y) represents the binary image pixels, and f(x,y) represents gray level image pixels.

4.1. Line Segmentation

Line segmentation is achieved by image horizontal projection that calculates the horizontal axis for the binarized image. The horizontal projection matrix I_j is calculated by summing up the pixel values P(i, j) along the X-axis for each y value, as shown in Equation (1):

For each j ∈ 0.m-1

$$I_j = \sum_{i-0}^{n-1} P(i,j) \quad (1)$$

Where, *n* and *m* are, respectively, the width and height of the image and P(*i*,*j*) is the pixel's value at the index (*i*,*j*).

This projection has information about the text lines that are indicated by areas of white intensity, as shown in Fig. 3. White intensities indicate the text area that contains text

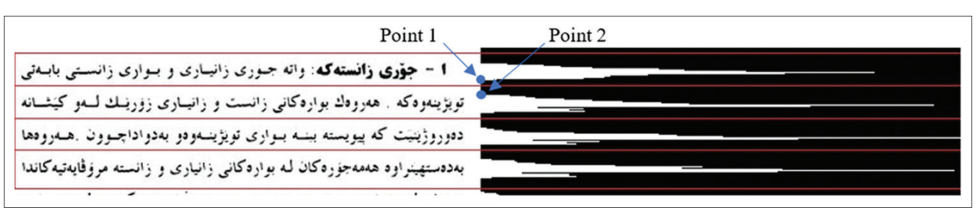

Fig. 3. Horizontal projection of input image that contains text line.

and the black intensities show the gap between the text lines.

Fig. 3 shows the line segmentation method that accepts an image of text written in the Kurdish language and separates its lines. The horizontal projection technique does this in two stages. The first one is to locate each group of white intensities in the horizontal projection and the second is to indicate the line position to separate lines from each other. To find the indicator line's position, we used the index of last white intensities (point 1) and the first index of next white intensities(point 2) and calculated the distance between these two points. The line position is in the middle of these two points (point 1 and point 2).

4.2. Word/Sub-word Segmentation

After the line segmentation stage, the subsequent stage is word segmentation. The method that is used for word segmentation is based on the connected component method. The algorithm takes a binary text line image of Kurdish text without dots and diacritics as input, and the result is a word/sub-word segmented image as output. The steps of word/sub-word segmentation are described in detail below.

4.2.1. Find the connected components

In this step, the text line image from the previous section is used to find the connected components. A connected component is every component that has adjacent pixels that are connected. Fig. 4(a) shows an example of connected components with boundaries. In the first version of the connected component result, all components are selected but for better word/sub-word extraction dots and diacritics must be ignored. To do so, the baseline of text lines must be found. A baseline is a fictitious line that follows and joins the lower and upper parts of the character bodies [35]. The baseline is the maximum value from the horizontal projection. The index of the max value determines the location of the baseline in the text line image. Fig. 4(a) shows an example of baseline detection that shows by the horizontal line that crosses the whole word. In continuation, using the baseline, the connected components are filtered based on whether these components are intersected with the baseline. Usually, dots and diacritics location are above or below the baseline, so we can ignore connected component that is not on the baseline. Fig. (4a) shows the original image after determining the connected components and the baseline. In Fig. (4b), the dots above the baseline are ignored.

4.2.2. Word/sub-words extraction

For the Kurdish script, a connected component either represents a word or a sub-word. This means that a single word may consist of one or more connected components. For extraction, we applied vertical projection to find the space between the words/sub-words (places that the projection is zero). Projection along the vertical axis is called the vertical projection. Vertical projection is calculated for every column as the sum of all row pixel values inside the column, as shown in Equation 2.

For each i ∈ 0.n-1

$$I_i = \sum_{j-0}^{m-1} P(i, j) \tag{2}$$

After finding gaps between components, the task is to decide if these gaps are spaces between words or between sub-words. In other words, what is the correct threshold to decide whether the separation space between two sub-words is a gap inside the same word or space between two different words? Although these gaps are not constant and can be vary based on the font sizes, font type, or style. To deal with this issue, first, we find the pen size, which is the pen thickness used for the writing of the adjacent two sequential words/sub-words, and evaluate it with the distance between the current consecutive words/sub-words. Calculating the pen size can handle by taking the most frequent value in the vertical projection applied for each sub-word. However, taking the maximum common value from the vertical projection of some single characters like "ا" gives a wrong conjecture of the pen size. Therefore, the pen size is calculated by considering the most common value calculated from the horizontal projection. Hence, if the maximum frequent value computed from the horizontal projection is

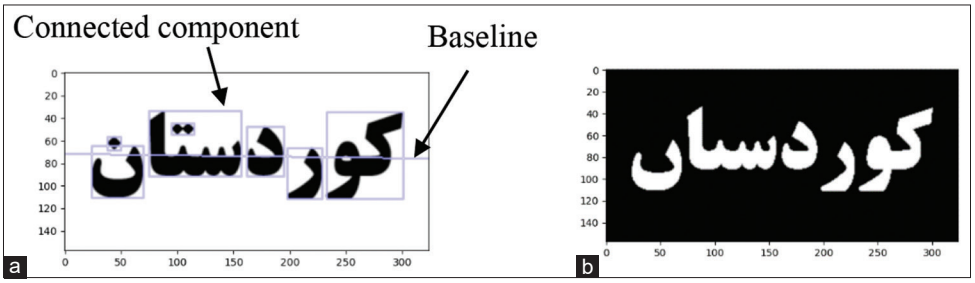

Fig. 4. (a) All connected component. (b) Binary version of ignored dots and diacritics.

greater than the maximum frequent value calculated from the vertical projection, then the pen size is equal to the maximum frequent value calculated from the vertical projection.

Pen size calculation is formally defined as:

$$PS = \begin{cases} \text{üüüüüüüü}(\text{üüüü}), & \text{[( )]} > \text{[( )]} \\ MFV(HP), & otherwise \end{cases}$$

Where, PS is the pen size, SW indicates sub-word, HP shows horizontal projection, VP shows vertical projection, and MFV represents the most frequent value. Fig. 5 shows an example of a pen size calculation for two cases.

After finding the pen size for each sub-word, it is compared with the spaces between the components. If the gap between two adjacent components (SS) is greater than the mean of the PS value of these two adjacent components, then this gap is a space between two separate words (WS), otherwise, this gap is a space between two sub-words (SWs) that belong to the same word. Figure 6 shows an example that determining the type of gap between the two words/sub-words in the same line. It is defined formally as:

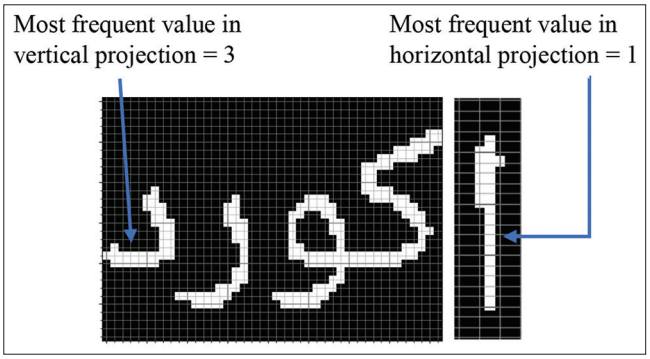

Fig. 5. Finding the most frequent value in both VP and HP, this value is the sub-word pen size.

$$SR = \begin{cases} WS, & SS[SW(i), SW(i+1)] > \dfrac{PS(i) + PS(i+1)}{2} \\ SWS, & otherwise \end{cases}$$

Where, SR, WS, SWS, SS, and SW are the separation region, the word separation, the sub-word separation, the separation space, and the sub-word, respectively.

4.3. Character Segmentation

The proposed algorithm for character segmentation is based on the vertical projection. The algorithm consists of two stages. The algorithm takes a binary image of word/sub-word and returns a binary image of segmented characters. Each step is explained in detail below:

4.3.1. Word/sub-word vertical projection

Vertical projection can provide a better definition of a letter's shape. This method can give us an accurate result. At this stage, we will once again find a vertical projection for the word. This technique reveals all the convexity and dents in the word. Fig. 7 shows an example of a vertical projection stage.

4.3.2. Segmentation areas identification

In this step, the vertical projection as shown in Figure 8 is examined to identify the segmentation (splitting) areas between letters. The segmentation area between the two letters is an area that ended one letter and started another letter or ended the word and we know that the baseline shared between all letter in a word, this means, letters join by the baseline, and if we find the pixel of baseline, we can find the start and the end of a letter or the area between letters. In the different font sizes and font styles, the baseline height or the pen size can be found by the most frequent data (MFD) in the vertical projection that we discussed previously. The process starts with finding the MFD in the vertical projection array. After this, we compare the other data in the VP array with the MFD. If the

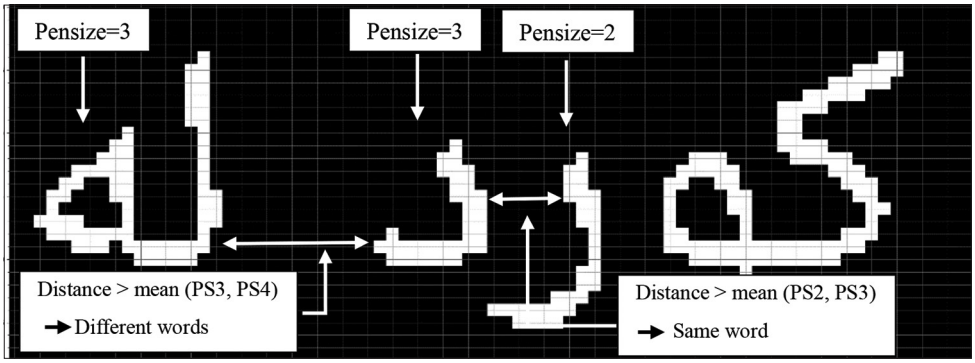

Fig. 6. Calculate the distance between two sub-words in the same word and different words.

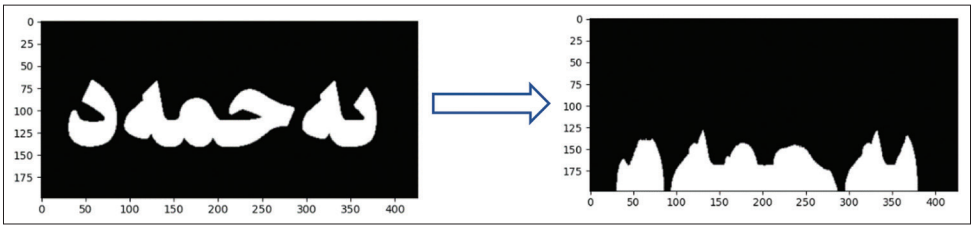

Fig. 7. Vertical projection example.

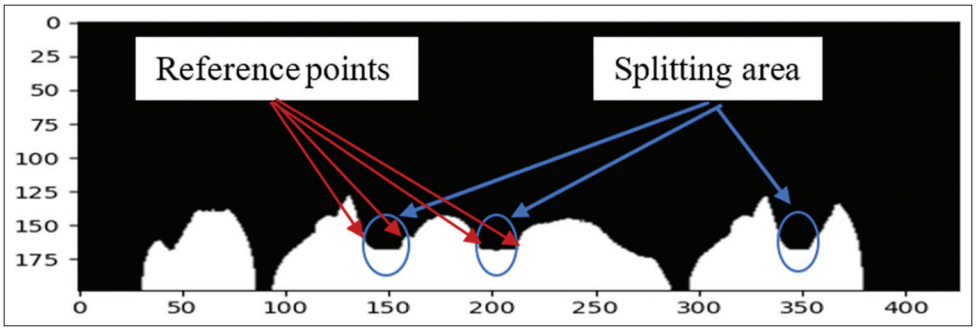

Fig. 8. Splitting areas with reference points.

difference of these data is less than 1, we will change the data of this index to the MFD, but if this difference is more than 1, we change the value of this cell to zero because this index is part of the character, not the splitting area. Now, we have an array consisting of MFD and zero values. Therefore, the MFD data in the array represent the splitting area between the two letters. By finding the start and end index of this group of data in the array, the beginning and end of the splitting area can be specified. After that, the first and last points in the region are considered as splitting area reference points. The start and endpoints can be used to separate the letters. By adding the middle line between these two points (separator line), as well as adding the start and end lines of connected components that are found in previous steps, we now have a separator line between the characters. Fig. 8 shows an example of finding the splitting area and reference points.

After identifying the splitting areas, each character is located between two consecutive splitting area. Fig. 9 shows the separator line over splitting areas for some characters.

However, there are some special cases where the splitting area locates within a character such as the letters "س" and "ش" in all forms. Besides, some splitting areas locate within a character when the position of the character is at the end of the word or exists independently in the text such as "ب" and "ی." Hence, a post-processing step for character segmentation is necessary to ignore these spatial splitting areas. However, in this paper, we do not work on the "س"

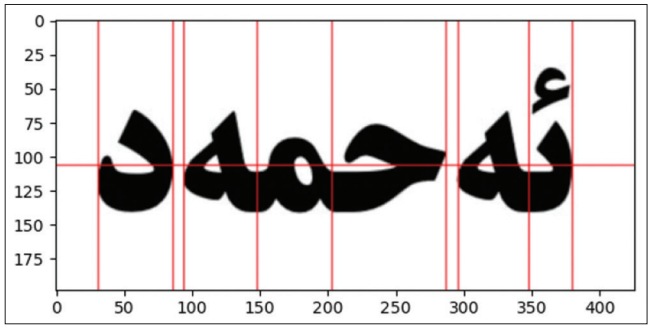

Fig. 9. Splitting regions for some regular characters.

Fig. 10. Example of special cases that must ignored.

and "ش," we only worked on the "ب" and "ی." Figure 10 shows some examples of these cases.

4.3.3. Post-processing

In the post-processing phase, we will take a step. In this main step, we can eliminate some of these special cases using the baseline that we found earlier, as well as the separator line that we found in the previous step. To do this, for any separation lines find the intersection point with the baseline. After that, check the value of this point (intersection point) in the image

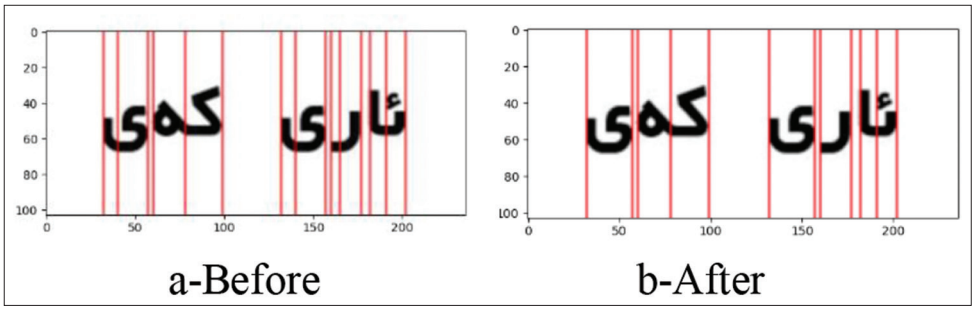

Fig. 11. Example of post-processing before and after applying.

binary data and if this value is zero its means this point not on the letter and we know that the separation line between the adjacent latter must be on the letters. Finally, these separation lines can be removed from the list. In this way, we can solve some of these situations using this technique. Fig. 11 shows some example after applying the post-processing step.

5. PERFORMANCE ANALYSIS

In this section, the results of testing our approach on a collection of images that contain Kurdish text are shown. We use the Python programming language to implement and then test our proposed character segmentation algorithm since it is a commonly known high-level programming language that provides well-implemented packages for image processing. The performance of line segmentation is measured by computing the ratio of the number of lines that are correctly segmented to the total number of inputted lines. The same measurement is used for each of word and character segmentations.

$$Accauracy = \frac{Num.of\ corrected\ segment}{Total\ of\ segments}$$

The proposed algorithms (line, word, and character segmentations) were experimented and evaluated using a manually created dataset. We develop a software to generate a dataset with the ground truth from the random Kurdish text that we collected. To make the dataset generic and comprehensive, the collected dataset includes text content from different sources (e.g. books, magazines, reports, and papers) and topics (e.g. religious, sport, and poetry texts), in addition to a considerable variation at font type, size, and style levels. These texts are converted to image word by word and add some noise to every image and saved all images. The proposed line segmentation methods were tested on 6099 lines and reported excellent results in terms of line segmentation ratio, which computed with an average of 99.9%. Table 1 shows the results generated through the testing process using different font types, styles, and sizes.

TABLE 1: Line segmentation results for different font styles and types on text

Font	Font type	Total number of input lines	No. of correctly segmented lines	Accuracy (%)
Plain	UniQAIDAR_ Blawkrawe 004	497	497	100
	Unikurd Web	507	506	99.9
	Noto Naskh Arabic UI	525	524	99.9
	UniQAIDAR_ JWNEYD	504	504	100
Bold	UniQAIDAR_ Blawkrawe 004	497	497	100
	Unikurd Web	507	506	99.9
	Noto Naskh Arabic UI	525	524	99.9
	UniQAIDAR_ JWNEYD	504	504	100
Italic	UniQAIDAR_ Blawkrawe 004	497	497	100
	Unikurd Web	507	506	99.9
	Noto Naskh Arabic UI	525	524	99.9
	UniQAIDAR_ JWNEYD	504	504	100
	Total	6099	6093	99.9

TABLE 2: Word segmentation results for different font types and size between 24 and 48

Font	Font type	Total number of input word	No. of correctly segmented word	Accuracy (%)
Plain	UniQAIDAR_ Blawkrawe 004	2218	2180	98.28
	Unikurd Web	2218	2112	95.22
	Noto Naskh Arabic UI	2218	2150	96.93
	UniQAIDAR_ JWNEYD	2218	2119	95.53
	Total	8872	8561	96.5

TABLE 3: Character segmentation results for different font types and font size between 24 and 48

Font	Font type	Total number of input character	No. of correctly segmented character	Accuracy (%)
Plain	UniQAIDAR_ Blawkrawe 004	15,887	15,842	99.7
	Unikurd Web	15,887	16,101	98.7
	Noto Naskh Arabic UI	15,887	15,636	98.4
	UniQAIDAR_ JWNEYD	15,887	15,495	97.5
	Total	63,548	63,074	98.6

The results of the word segmentation stage in terms of word segmentation ratio are reported in Table 2. The proposed word segmentation methods are experimented on about 8872 words with four font types (Noto Naskh Arabic UI, Unikurd Web, UniQAIDAR_JWNEYD and UniQAIDAR_Blawkrawe 004) and five font sizes (24, 26, 28, 36, and 48 points), with an average accuracy of 96.5%. The results show that the algorithm has almost the same performance when changing the font size. Furthermore, we experimented with the character segmentation stage on different font types and sizes on about 63,548 characters. Table 3 shows the performance of the proposed algorithm with an average accuracy of 98.6%. The results show that the algorithm has almost the same performance regardless of the font type, style, and size.

TABLE 4: Comparing with other related work

Articles	Year	Segmentation method	Dataset	Font type	Font size	Font style	Average accuracy (%)
Zheng *et al*. [10]	2004	Vertical histogram and some structural characteristics rules	500 samples of Arabic text	Simplified Arabic and Arabic transparent	12, 14, 16, 18, 20, and 22	Plain	94.8
Javed *et al*. [27]	2010	Pattern matching techniques	A total of 1282 unique ligatures are extracted from the 5000 high-frequency words in a corpus-based dictionary	Noori Nastalique font	36	Plain	92
Saabni [26]	2014	Partial segmentation and Hausdorff distance	APTI	Different fonts to cover different complexity of shapes of Arabic printed characters	10 different sizes	Plain	96.8
Anwar *et al*. [28]	2015	Projection-based	127 sentences composed of 1061 letters	Traditional Arabic	70	Plain	97.5
Amara *et al*. [29]	2016	Histogram and contextual properties	APTI	Different font types	Different sizes	Plain, italic, and bold	85.6
Radwan *et al*. [32]	2016	Multichannel neural networks	APTI	Arial, Tahoma, Thuluth, and Damas	18	Plain	95.5
Qomariyah *et al*. [33]	2017	Interests points, contour-based	10 lines of 30 sub-words	Not reported	Not reported	Plain	86.5
Fakhry [30]	2017	Connected component	5 lines 15 words	Not reported	Not reported	Plain	80.2
Amara *et al*. [31]	2017	Projection profile, SVM	APTI	Advertising bold	6,8,10, 12	Plain, italic, and bold	98.24
Zoizou *et al*. [34]	2018	Contour-based and template matching	83 lines of 984 words	34 different fonts	Different font sizes	Plain	94.7
Mohammad *et al*. [25]	2019	Contour-based method	1500 lines of (23,350 words)	Advertising bold, simplified Arabic, Arial, traditional Arabic, and Times New Roman	8, 9, 10, 12, 14, 16, 18, and 24	Plain, italic, and bold	98.5
Our approach	2020	Projection based	6099 line of (63,548) letters	UniQAIDAR_004, Unikurd_ Web, Noto_Naskh Arabic UI, UniQAIDAR_JWNEYD	24, 26, 28, 38, and 48	Plain	98.6

Table 4 shows our results compared with some previous related works. As shown in the table, the proposed algorithm performs better in comparison with other related works in: (i) Using more font types, sizes, and styles than the other approaches (ii) and recording higher average accuracies.

6. CONCLUSION

In this paper, line, word, and character segmentation algorithms are proposed for Kurdish printed text based on projection-based segmentation methods. The proposed algorithm can segment the characters of words. The algorithm can also handle certain complex cases that occur due to over-segmentation problems. We tested the algorithm on the manually created dataset by creating different versions of the same text using different font types, styles, and sizes. Experimental results show the reliability of our algorithm in performing a correct segmentation of more than 63,074 out of 63,548 words without the س and ش letter.

The segmentation of the Kurdish text is prone to errors, which leads to classification errors. The proposed segmentation algorithms are capable of minimize errors and maximize the classification rate. An advanced method is proposed for word/sub-word segmentation. Horizontal and vertical segmentations are used to distinguish between words and sub-words based on the size of the gaps that separate the connected components in comparison to the pen size.

For the character segmentation step, an advanced projection-based algorithm is proposed. The proposed algorithm is built easily and reliably that can fit a variety of fonts and styles, the character segmentation algorithm shows good results up to 98.6%.

For future work, we plan to find the correct segmentation for characters, such as "س" and "ش," by ignoring the over-segmentation part that occurs in these two special characters cases. Furthermore, we want to extend the work to extract all characters more accurately to facilitate the recognition stage.

REFERENCES

[1] H. Althobaiti and C. Lu. "*A survey on Arabic Optical Character Recognition and an Isolated Handwritten Arabic Character Recognition Algorithm Using Encoded Freeman Chain Code*". 2017 51st Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, pp. 1-6, 2017.

[2] A. Lawgali. "A survey on Arabic character recognition". *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 8, no. 2, pp. 401-426, 2015.

[3] S. Elaiwat and M. A. Abu-Zanona. "A three stages segmentation model for a higher accurate off-line arabic handwriting recognition. *World of Computer Science and Information Technology Journal*, vol. 2, no. 3, pp. 98-104, 2012.

[4] M. A. Abdullah, L. M. Al-Harigy and H. H. Al-Fraidi. "Off-line arabic handwriting character recognition using word segmentation". *Journal of Computing*, vol. 4, pp. 40-44, 2012.

[5] Y. M. Alginahi. "A survey on Arabic character segmentation". *International Journal on Document Analysis and Recognition*, vol. 16, no. 2, pp. 105-126, 2013.

[6] A. Cheung, M. Bennamoun and N. W. Bergmann. "An Arabic optical character recognition system using recognition-based segmentation". *Pattern Recognition*, vol. 34, no. 2, pp. 215-233, 2001.

[7] N. A. Shaikh, G. A. Mallah and Z. A. Shaikh. "Character segmentation of Sindhi, an Arabic style scripting language, using height profile vector". *Australian Journal of Basic and Applied Sciences*, vol. 3, no. 4, pp. 4160-4169, 2009.

[8] M. M. Alipour. "A new approach to segmentation of Persian cursive script based on adjustment the fragments". *International Journal of Computers and Applications*, vol. 64, no. 11, pp. 21-26, 2013.

[9] S. N. Nawaz, M. Sarfraz, A. Zidouri and. W. G. Al-Khatib. "*An Approach to Offline Arabic Character Recognition Using Neural Networks*". In: 10th IEEE The IEEE International Conference on Electronics, Circuits, and Systems, IEEE, vol. 3, pp. 1328-1331, 2003.

[10] L. Zheng, A. H. Hassin and X. Tang. "A new algorithm for machine printed Arabic character segmentation". *Pattern Recognition Letters*, vol. 25, no. 15, pp. 1723-1729, 2004.

[11] A. Zidouri and K. Nayebi. "*Adaptive Dissection Based Subword Segmentation of Printed Arabic Text*". In: 9th International Conference on Information Visualisation (IV), IEEE, pp. 239-243, 2005.

[12] J. Ahmad. "Optical character recognition system for Arabic text using cursive multi-directional approach". *Journal of Computational Science*, vol. 3, pp. 549-555, 2007.

[13] M. Omidyeganeh, K. Nayebi. "*A New Segmentation Technique for Multi font Farsi/Arabic Texts*". In: IEEE International Conference on Acoustics Speech, and Signal Process., IEEE, vol. 2, 2005.

[14] T. Sari, L. Souici, and M. Sellami. "*Off-line Handwritten Arabic Character Segmentation Algorithm: ACSA*". In: Proceeding 8th International Workshop Front Handwriting Recognit., IEEE, pp. 452-457, 2002.

[15] R. Mehran, H. Pirsiavash and F. Razzazi. "*A Front-end OCR for Omni-font Persian/Arabic Cursive Printed Documents*". In: Digital Image Computing: Techniques and Applications (DICTA), IEEE, pp. 56-56, 2005.

[16] A. Al-Nassiri, S. Abdulla and R. Salam. "The segmentation of off-line arabic characters, categorization and review". *International Journal on Media Technology*, vol. 1, no. 1, pp. 25-34, 2017.

[17] M. M. Altuwaijri and M. A. Bayoumi. "A thinning algorithm for Arabic characters using art2 neural network". *IEEE Transactions on Circuits and Systems*, vol. 45, no. 2, pp. 260-264, 1998.

[18] A. A. A. Ali and M. Suresha. Survey on segmentation and recognition of handwritten arabic script. *SN Computer Science*, vol. 1, p. 192, 2020.

[19] I. Aljarrah, O. Al-Khaleel, K. Mhaidat, M. Alrefai, A. Alzu'bi and M. Rabab'ah. 2012. *Automated System for Arabic Optical Character*

Recognition. In: Proceedings of the 3rd International Conference on Information and Communication Systems(ICICS'12).

[20] Y. Alginahi. "A survey on Arabic character segmentation". *International Journal on Document Analysis and Recognition*, vol. 16, pp. 105-126, 2013.

[21] Y. Zhang, Z. Q. Zha and L. F. Bai. "A license plate character segmentation method based on character contour and template matching". *Applied Mechanics and Materials*, vol. 333, pp. 974-979, 2013.

[22] I. Ahmed, M. Sabri and P. Mohammad. *Printed Arabic Text Recognition*. Guide to OCR for Arabic Scripts, 2012.

[23] M. Bennamoun and B. Boashash. "A structural-description-based vision system for automatic object recognition". *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 27, no. 6, pp. 893-906, 1997.

[24] M. Mostafa."*An Adaptive Algorithm for the Automatic Segmentation of Printed Arabic Text*". In: 17th National Computer Conference, International Society for Optics and Photonics, Saudi Arabia, pp. 437-444, 2004.

[25] K. Mohammad, A. Qaroush, M. Ayesh, M. Washha, A. Alsadeh and S. Agaian. Contour-based character segmentation for printed Arabic text with diacritics. *Journal of Electronic Imaging*, vol. 28, no. 4, p. 1, 2019.

[26] R. Saabni. "*Efficient Recognition of Machine Printed Arabic Text Using Partial Segmentation and Hausdorff Distance*". In: 6th International Conference Soft Computing and Pattern Recognition (SoCPaR), pp. 284-289, 2014.

[27] S. T. Javed, S. Hussain, A. Maqbool, S. Asloob, S. Jamil and H. Moin. "Segmentation free nastalique urdu OCR". *World Academy of Science, Engineering and Technology*, vol. 4, no. 10, pp. 456-461, 2010.

[28] K. Anwar, Adiwijaya and H. Nugroho. "*A Segmentation Scheme of Arabic Words with Harakat*". In: IEEE International Conference on Communications, Networks and Satellite (COMNESTAT), pp. 111-114, 2015.

[29] M. Amaram, K. Zidi, G. Ghedira and S. Zidi. "*New Rules to Enhance the Performances of Histogram Projection for Segmenting Small-sized Arabic Words*," In: International Conference on Hybrid Intelligent Systems, 2016.

[30] F. I. Firdaus, A. Khumaini and F. Utaminingrum. "*Arabic Letter Segmentation Using Modified Connected Component Labeling*". In: International Conference on Sustainable Information Engineering and Technology (SIET), pp. 392-397, 2017.

[31] M. Amara, K. Zidi and K. Ghedira. "An efficient and flexible knowledge- based Arabic text segmentation approach". *The International Journal of Computer Science and Information Security*, vol. 15, no. 7, pp. 25-35, 2017.

[32] M. A. Radwan, M. I. Khalil and H. M. Abbas. "Predictive segmentation using multichannel neural networks in Arabic OCR system". *Lecture Notes in Computer Science*, vol. 9896, pp. 233-245, 2016.

[33] F. Qomariyah, F. Utaminingrum and W. F. Mahmudy. "The segmentation of printed Arabic characters based on interest point". *The Journal of Telecommunication, Electronic and Computer Engineering*, vol. 9, no. 2-8, pp. 19-24, 2017.

[34] A. Zoizou, A. Zarghili and I. Chaker. "A new hybrid method for Arabic multi-font text segmentation, and a reference corpus construction". *Journal of King Saud University Computer and Information Sciences*, vol. 32, no. 5, pp. 576-582, 2018.

[35] A. Fawzi, M. Pastor and C. D. Martínez-Hinarejos. "*Baseline Detection on Arabic Handwritten Documents*". P Proceedings of the 2017 ACM Symposium on Document Engineering, pp. 193-196, 2017.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Prediction of CoVid-19 mortality in Iraq-Kurdistan by using Machine learning

Brzu T. Muhammed[1], Ardalan H. Awlla[2], Sherko H. Murad[3], Sabah N. Ahmad[4]

[1]Department of Computer Science, Kurdistan Technical Institute, Sulaymaniyah, Iraq, [2]Department Information Technology, University of Human Development, Sulaymaniyah 0778-6, Iraq, [3]Department of Computer Science, Kurdistan Technical Institute, Sulaymaniyah, Iraq, [4]General Manager of Health in Sulaymaniyah, Iraq

ABSTRACT

This research analyzed different aspects of coronavirus disease (COVID-19) for patients who have coronavirus, for find out which aspects have an effect to patient death. First, a literature has been made with the previous research that has been done on the analysis dataset of coronavirus using Machine learning (ML) algorithm. Second, data analytics is applied on a dataset of Sulaymaniyah, Iraq, to find factors that affect the mortality rate of coronavirus patients. Third, classification algorithms are used on a dataset of 1365 samples provided by hospitals in Sulaymaniyah, Iraq to diagnose COVID-19. Using ML algorithm provided us to find mortality rate of this disease, and detect which factor has major effect to patient death. It is shown here that support vector machine (SVM), decision tree (DT), and naive Bayes algorithms can classify COVID-19 patients, and DT is best one among them at an accuracy (96.7 %).

Index Terms: Coronavirus disease, Coronavirus, Forecasting, Machine learning, Kurdistan-IRAQ

1. INTRODUCTION

The coronavirus disease (COVID-19) is the family of viruses including SARS, ARDS. W.H.O declared this outbreak as a public health emergency [1] and mentioned the following; the virus is being transmitted through the respiratory tract when a healthy person comes in contact with the infected person.

In December 2019, Wuhan, Hubei region, China, has been accounted for as the focal point of the COVID-19 episode [2]. A quarter of a year later, that outbreak was pronounced as a worldwide pandemic by the World Health Organization (WHO) [3]. More than 54.40 million confirmed COVID-19 cases and more than 1.32 deaths worldwide have been officially reported in 16 November, 2020. Therefore, it has been considered as the most critical universal crisis since the World War-II [4]. The coronavirus has spread in Kurdistan – Iraq like all the country in the world, and it has expanded fast in Sulaymaniyah city. The mortality of this disease expands day by day and this infection becomes as a major danger to the mankind of whole world. Alongside the clinical explores, the examination of related information will support the humanity. Recent studies identified that machine learning (ML) and artificial intelligence (AI) are promising technology employed by various health-care providers as they result in better scale-up, speed-up processing power, reliable, and even outperform human in specific health-care tasks [5]. In this paper, we established three ML algorithm for the prediction of coronaviruses' diseased patients' mortality. The models forecast when COVID-19 infected patients would be death or recovered. The proposed algorithms are designed with the dataset found from Sulaymaniyah city for coronavirus and dataset cases of the death and recovery records of the infected coronavirus's pandemic. ML algorithm which includes decision tree (DT), support vector

Access this article online

DOI: 10.21928/uhdjst.v5n1y2021.pp66-70 **E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: Brzu T. Muhammed, Department of Computer Science, Kurdistan Technical Institute, Sulaymaniyah, Iraq. E-mail: brzu.tahir@kti.edu.krd

Received: 22-11-2020 Accepted: 19-05-2021 Published: 23-05-2021

machine (SVM), and naive Bayes (NB) was implemented directly on the dataset using Weka Tool which is a data mining tool.

2. LITERATURE REVIEW

Development of AI changed the world in all fields. ML a subset of AI causes the human to discover answers for exceptionally complex issues and furthermore assumes an imperative part in making human life refined. The application zones of ML incorporate business applications, clever robots, medical services, atmosphere demonstrating, picture handling, natural language preparing, and gaming [6]. According to Al Sadig *et al.* [7], depend on the dataset as given by the various site developed by digital science in cooperation with over 100 leading research organizations all over the world. Create a model using J48 algorithm to predict the most common symptoms causing death is acute kidney injury and coronary heart disease.

Arun and Iyer [8] propose some of the ML techniques such as rough set (SVM), Bayesian Ridge and Polynomial Regression, SIR model, and RNN to examination of the transmission of COVID 19 disease and predict the scale of the pandemic, the recovery rate as well as the fatality rate.

According to Bullock *et al.* [9], ML and deep learning can replace humans by giving an accurate diagnosis. The perfect diagnosis can save radiologists' time and can be cost-effective than standard tests for COVID-19. X-rays and computed tomography scans can be used for training the ML model.

Wang and Wong [10] created COVID-Net, which is a profound convolutional neural network, which can analyze COVID-19 from chest radiography pictures.

Alibaba Cloud 2020 [11] exploit ML to set up an adjusted Susceptible - Exposed - Infectious - Recovered model to anticipate the commonness of COVID-19 and evaluate the expanded danger of defilement in a particular territory.

Kemenkes [12] finding diabetes utilizing AI and ML methods result demonstrated that ensemble technique guaranteed exactness of 98.60%. These reasons can be advantageous to analyze and foresee COVID-19. According to Muhammad *et al.* [13] use several ML algorithms which includes DT, SVM, NB, LR, RF, and K-NN are applied directly on the dataset which include COVID-19 infected patients' recovery, the model invented with DT algorithm was discovered to be the most precise with 99.85% exactness which has all the earmarks of being the most noteworthy among others.

3. DATA PREPARATION

Collection of data is the vital step to induce data over corona virus. The information was collected from the distinction health care center in Sulaymaniyah City in the Kurdistan Region of IRAQ. The dataset comprises 1376 patients which have appeared side effects of crown infection. The data collection comprises seven factors (Gender, Age, status, O_2, ventilate, Day of hospitalization, and death patient). The informational index contained data about hospitalized patients with COVID-19. After informational index start another stage data preprocessing. Information preparing is a significant cycle being developed of ML model. The information gathered is frequently approximately controlled with out-of-range esteems, missing values; and so, on such information can deceive the consequence of the examination. Weka, one of the expansively utilized data mining computer program, is utilized for the classification. The processing of data preparation illustrated in Figure 1.

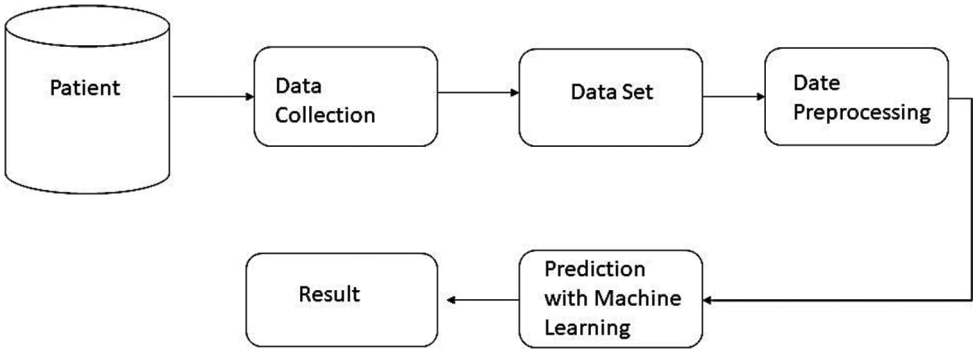

Fig. 1. The workflow diagram for coronavirus disease.

After preprocessing stage of the dataset, the collected variables are divided into two classes "Death" assigned as "Yes" and "No Death" assigned as "No." The selected data samples are transferred to a spreadsheet file for further processing to be suitable for data mining approaches. The dataset were normalized to minimize the effect of scaling on the data and saved as a commas separated value file format. In Table 1, all the attributes explained with their description.

4. METHODOLOGY

Recently, ML techniques have been used to medical prediction; there are different types of ML algorithms that can be applied to different types of applications in various fields [14]. Many different types of research have demonstrated that algorithms of ML had given better help to clinical backings moreover for decision-making on the basis of the patient information. In the medical services field, illness predictive examination is one of the valuable and strong uses of ML forecast algorithms. In this paper proposed a machine-learning algorithm to analyze unusual COVIDE-19 disease datasets. In this paper, rate of death across the region analyzed based on the factors explained in Table 1.

4.1. SVM

SVM is one supervised classification algorithm which is commonly utilized for linear classification and regression problem. It means SVM can solve both linear and nonlinear problems. SVM provides unique and optimal solution, the kernel function is selected based on the points of the variables in the hyperplane. The best separating hyper plane can be written as, W.X + b = 0, Where w is a weight vector, the value of the attributes is referred as x, and b is scalar often referred as bias [15].

TABLE 1: Attribute's description used for predication does the patient recovered or died

Variables	Description	Possible values
Gender	It is a social definition of men and women.	Male, Female
Age	Patient age	Date
Status	Situation of the patients' status.	Bad, Severe, Good, Critical
O_2	It indicates does the patient need oxygen or not.	Yes, No
Ventilate	A ventilator uses pressure to blow air or air with extra oxygen	Yes, No
Date of hospital admission	Day of hospitalization	Date
Death	Does the patient died or recovered?	Yes, No

4.2. DT

DT is a supervised learning algorithm that can be utilized for both classification and regression issues, however generally it is ideal for attempting Classification issues. It is a DT classifier; the structure of this algorithm is divided by three parts: Internal node which is features of dataset, branches are demonstrating rules, and leaf is represent outcome for each leaf.

4.3. Naïve Bias

NB classifier is the simple and powerful supervised machine-learning algorithm used for predictive modeling. It considers all variables contribute in the direction of arrangement and they are equally connected [16]. The algorithm is based on a theorem called Bayesian Theorem and used when the coordination of the inputs is high, which assumes that features are statistically independent.

5. EXPERIMENT RESULTS

For the experiment Weka tool have been used, the dataset was collected used to train the above algorithms using the Weka tool. In this paper, the dataset is divided for two parts, for the classification algorithms first part which is 80% used for training the classification algorithms and the second part which is 20% used as a test set and the results are illustrated in Table 2. The achievement of every algorithm was assessed at phases of the training set. Every algorithm was trained with the record sets having 1100 records. This examination is carried out to achieve which algorithm can be the most appropriate for the prediction of COVID-19.

The accuracy of the forecast algorithm in almost all of the research work has exploited like one of the regular measurability while working on the forecast algorithm. In

TABLE 2: Accuracy classification algorithms

S. No.	Classification algorithms	Accuracy
1	Decision tree	96.07
2	Support vector machines	95.27
3	Naive Bayes	94.47

this paper, the accuracy forecast is whether the patient is recovered or deceased while the patient infected by the COVID-19. Base on the above algorithms mentioned in Table 2 and in Figure 2. Each classification algorithm has an alternate expectation precision dependent on its hyper parameters.

TABLE 3: Error metrics for the classification algorithms

S. No.	Algorithm	Kappa statistics	Mean absolute	Root mean square
1	Decision Tree	0.29	0.07	0.19
2	Support Vector Machine s	0.42	0.10	0.21
3	Naive Bayes	0.32	0.12	0.22

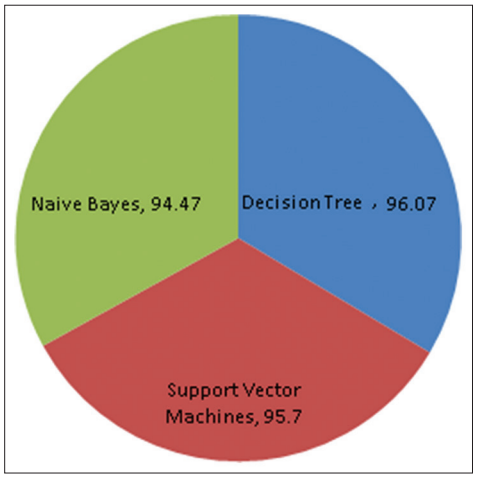

Fig. 2. Accuracy classification algorithms.

Table 3. Describe the performance error measurement for each algorithm; the error metrics which are kappa statistics, mean absolute error, and root mean square error for each algorithm is assessed.

As shown in Table 3. The decision tree has lowest error rates compared to other algorithms.

According to Figure 3, which is the visualization tree for the DT algorithm and Figure 4, the main factor which is the ventilator has the maximum effect on patients, which made it the beginning tree and this cause has the most effect on patients to recover or not. If the patient is not recovered depend on the most second-factor attribute which is status and the rest of the other attributes showed in the DT has a type of impact on the patient is recovery or died. However, the main factor attributes are ventilator, status as shown in Figure 3. This means that if a patient attribute ventilator is yes and the status are bad the patient died otherwise status factors Severe and Critical mostly recovered. In addition, the interesting Status attribute is good which is depending on the patient's date of hospital admission, as shown in Figure 3. It means some patients are in a good status, but they are dead. Because epidemic COVID-19 has more influence in cold weather as shown in Figure 5, it means weather conditions can increase cause death because of COVID-19 in the winter season.

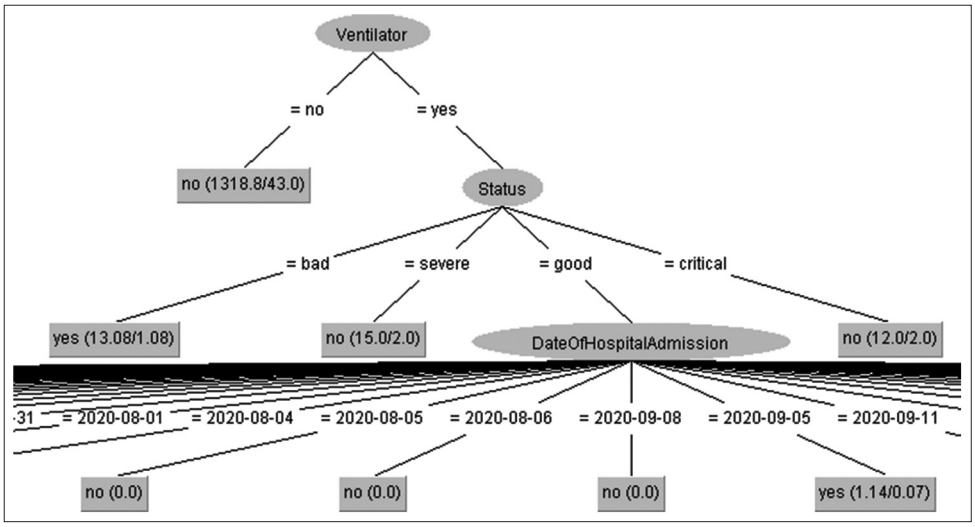

Fig. 3. A decision tree generated by the C4.5 algorithm for predicting COVID-19.

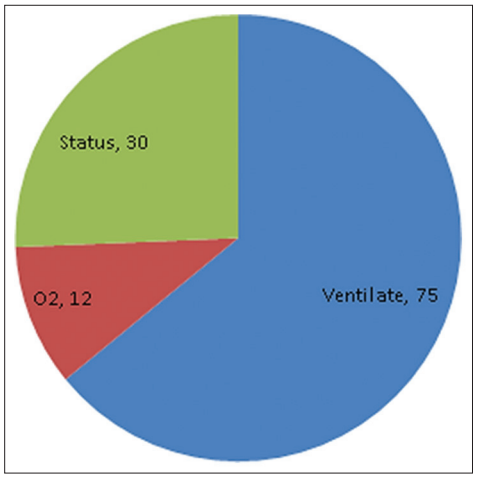

Fig. 4.: Factors with a significant effect on patient mortality.

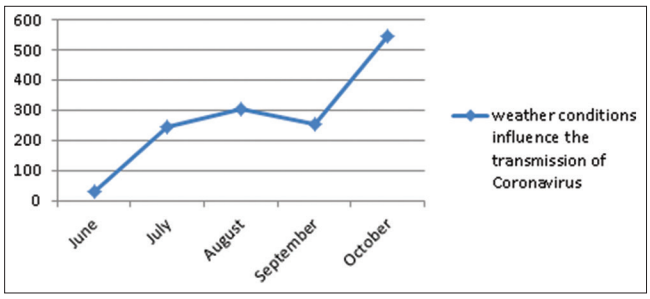

Fig. 5. The role of weather condition on transmission rates of the coronavirus.

6. CONCLUSION

The COVID-19 pandemic lay-down medical care systems in the entire world into a difficult situation. Computer algorithms and ML can help humanity to finding best solution to overcome the coronavirus epidemic. In this paper, data mining technique was used for the predication of coronaviruses infected patient's using dataset of coronaviruses patients of Iraqi-Kurdistan region. DT, support vector machine and NB were used directly on the dataset using Weka ML tool. To identify the accuracy suggested algorithms, the accuracy of the algorithms has been calculated based on the dataset features that have been used. The experiment result showed that the DT has the highest percentage of accuracy which is 96.7% followed by Support Vector Machine which is 95.27 accuracy and Naïve Bayes which is 94.47% accuracy. The experiment result showed that the most effective reason for the patient to recover or not is ventilator and other factors have their effect on the patients to recover or not. In addition, the weather condition means with the coming of the cold weather the virus's effects will increase.

REFERENCES

[1] Medscape Medical News. *The WHO Declares Public Health Emergency for Novel Coronavirus*, 2020. Available from: https://www.medscape.com/viewarticle/924596.

[2] M. C. Collivignarelli, C. Collivignarelli, M. Carnevale Miino, A. Abbà, R. Pedrazzani and G. Bertanza. "SARS-CoV-2 in sewer systems and connected facilities". *Process Safety and Environmental Protection*, vol. 143, pp. 196-203, 2020.

[3] P. Shi, Y. Dong, H. Yan, C. Zhao, X. Li, W. Liu, M. He, S. Tang and S. Xi. "Impact of temperature on the dynamics of the COVID-19 outbreak in China". *Science of the Total Environment*, vol. 728, p. 138890, 2020.

[4] S. Boccaletti, W. Ditto, G. Mindlin and A. Atangana, A. "Modeling and forecasting of epidemic spreading: The case of Covid-19 and beyond". *Chaos Solitons Fractals*, vol. 135, p. 109794, 2020.

[5] T. Davenport and R. Kalakota. "The potential for artificial intelligence in healthcare". *Future Healthcare Journal*, vol. 6, no. 2, pp. 94-98, 2019.

[6] P. Theerthagiri, I. J. Jacob, A. U. Ruby and Y. Vamsidhar. *An Investigation of Machine Learning Algorithms on COVID-19 Dataset*, 2020.

[7] M. Al Sadig and K. N. Abdul Sattar. "Developing a prediction model using j48 algorithm to predict symptoms of COVID-19 causing death". *International Journal of Computer Science and Network Security*, vol. 20, no. 8, p. 80, 2020.

[8] S. S. Arun and G. N. Iyer. "*On the Analysis of COVID19-Novel Corona Viral Disease Pandemic Spread Data Using Machine Learning Techniques*. 4th International Conference on Intelligent Computing and Control Systems, pp. 1222-1227, 2020.

[9] J. Bullock, A. Luccioni, K. H. Pham, C. S. N. Lam and M. Luengo-Oroz. Mapping the landscape of artificial intelligence applications against COVID-19". *Journal of Artificial Intelligence Research*, vol. 69, pp. 807-845, 2020.

[10] L. Wang and A. Wong. "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 Cases from chest radiography images". *Scientific Reports*, vol. 10, p. 19549, 2020.

[11] Alibaba Cloud e-Magazine. "*Alibaba Cloud Helps Fight COVID-19 through Technology*". Alibaba Cloud, 2020.

[12] Kementerian Kesehatan RI. Pedoman Pencegahan dan Pengendalian Coronavirus Disease (COVID-19). In: L. Aziza, A. Aqmarina and M. Ihsan (Eds.), Revisi Ke4. Kementerian Kesehatan RI, Direktorat Jenderal Pencegahan dan Pengendalian Penyakit (P2P), 2020. Available from: https://www.infeksiemerging.kemkes.go.id.

[13] L. J. Muhammad, M. M. Islam, U. S. Sharif and S. I. Ayon. "Predictive data mining models for novel coronavirus (COVID-19) infected patients recovery". *SN Computer Science*, vol. 1, no. 4, p. 206, 2020.

[14] Sirwan. M. Aziz and Ardalan. H. Awlla. "Performance to build effective student using data mining techniques". *UHD Journal of Science and Technology*, vol. 3, no. 2, p. 10, 2019.

[15] R. Sukanya and K. Prabha. "Comparative analysis for prediction of rainfall using data mining techniques with artificial neural network". *International Journal of Computational Science and Engineering*, vol. 5, pp. 1-5, 2017.

[16] S. D. Jadhav and H. Channe. "Comparative study of K-NN, naive Bayes and decision tree classification techniques". *International Journal of Science and Research*, vol. 5, pp. 1842-1845, 2016.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Slantlet based Statistical Features Extraction for Classification of Normal, Arrhythmia, and Congestive Heart Failure in Electrocardiogram

Sawza Saadi Saeed, Raghad Zuhair Yousif

Department of Physic, College of Science, Salahaddin University-Erbil, Erbil, Iraq

ABSTRACT

Intelligent and automated systems for diagnosing heart disease play a key role in treatment of heart disease and hence mitigating the possibility of heart disease, heart failure or sudden death. Thus, a Computer-Aided Design CAD system can provide a doctors with a clue about the category of patient heart disease, which might be Normal Sinus Rhythm, Abnormal Arrhythmia (ARR), and Congestive Heart Failure (CHF) electrocardiogram (ECG) signal. In this work a novel Slantlet transform (SLT) statistical features have been extracted and selected for 900 ECG segments taken from MIT-BIH ARR Database equally from three classes mentioned above for heart dieses classification through ECG signals. Based on the superiority of SLT in time localization as compared to the traditional wavelet transform, 12 out of 14 statistical features have been successfully passed the ANOVA test with *P*-value of 10^{-3}. Then after, the relevant features are provided to three well-known classifiers (Support Vector Machine [SVM], K-nearest neighbor, and Naive Bayes). The performance tests show that Attaining 99.254% classification average AUC it can be achieved using SLT transform based features along with SVM classifier, which is a set of related supervised machine learning algorithm used for regression and classification with high generalization ability. It performs classification on two group problems. SVM classifier determines the best hyperplane which distinguishes between each positive and negative training sample.

Index Terms: *Electrocardiogram, Slantlet, Abnormal arrhythmia, Congestive heart failure, Normal sinus rhythm*

1. INTRODUCTION

An electrocardiogram (ECG) is a non-invasive diagnostic method that detects variations in the electrical activity of the heart over time by graphically measuring the heart's rhythm and electrical activity [1]. Hence, it is vital to obtain and track ECG signals for early detection of diseases such as arrhythmia (ARR) and CHF [2]. Therefore, the automatic ECG signal classification of the latter ARR is worth studying field. The four main stages in a standard Computer-Aided Design system diagnosis involves: Preprocessing of signals, extraction of specific features, collection of significant features, and classification [3]. For classification problems, significant feature vectors, for example, continue to be the main and appropriate means of signal depiction. Many researchers from various fields who are involved in data modeling and classification are engaging to solve feature extraction problems [4]. The discrete wavelet transform (DWT) is especially useful in the fields of signal/image processing, such as denoizing, compression, and estimation [5]. However, in terms of time localization, it is unable to produce an ideal discrete-time basis [6], [7]. The Slantlet transform (SLT) has commonly been suggested as a better alternative to the classical DWT in terms of time localization [6]. Thus, in this article, SLT transform, has been suggested to extract a statistical feature, from ECG signals. The standard ECG signal is depicted in Fig. 1.

Access this article online

DOI: 10.21928/uhdjst.v5n2y2021.pp71-80

E-ISSN: 2521-4217

P-ISSN: 2521-4209

Corresponding author's e-mail: Sawza Saadi Saeed, Department of Physic, College of Science, Salahaddin University-Erbil, Erbil, Iraq. E-mail:sawza.saeed@su.edu.krd

Received: 29-03-2021 Accepted: 07-06-2021 Published: 10-06-2021

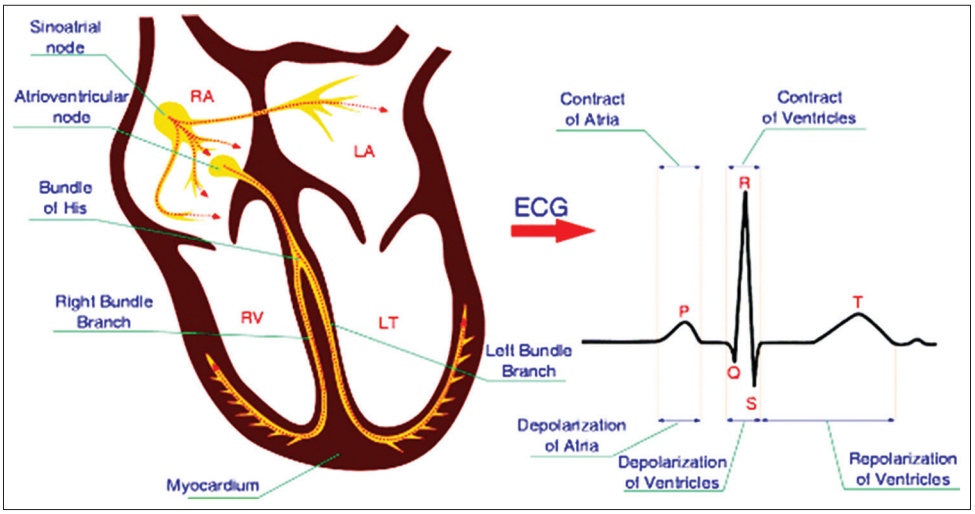

Fig. 1. Electrocardiogram signal parameters [1].

Electrocardiography is the recording of the electrical activity of the heart. The waveform is used to assess the rate and regularity of heartbeats, as well as the existence of any heart damage and the effects of medications or devices used to control the heart, such as a pacemaker. The ECG signals are weak (in mV) and have a broad frequency range (0.05–100 Hz), with the bulk of the useful information found in the 0.5–45 Hz range [8], [9]. P wave, which states with atrial depolarization, is one of the numerous waveforms and features of ECG. P waves have a typical amplitude of 0.1–0.2 mV and a normal length of 60–80 ms. The QRS complex is ventricular depolarization with a typical amplitude of about 1 mv and a duration of 0.06–0.12 s [10], [11]. The aim of this paper is to extract features from ECG signals using SLT transforms-based statistical features. This task uses an ECG dataset to identify people into three groups: those with cardiac ARR, congestive heart failure (CHF), and those with normal sinus rhythm (NSR). The aim of this paper is to extract features from ECG signals using SLT transforms-based statistical features. This task uses an ECG dataset to identify people into three groups: those with cardiac ARR, CHF, and those with NSR.

Our approach is consisting of preprocessing, feature extraction, feature selection based on ANOVA test, and then training the input to three different types of NN-based classifiers which are support vector machine [SVM], Naive Bayes [NB], and K-Nearest Neighbor [KNN]. Chami *et al.* proposed a system for five classes of heartbeat categories classification for ECG ARR diagnosis based on a combination of DWT and higher order statistics feature extraction and entropy based feature selection methods along with SVM classifier [12]. Nahak *et al.* proposed a method for analyzing and classifying the three types of ECGs (namely ARR, CHF, and NSR). Feature representations from the ECG signal's heart rate variability (HRV) were derived based on wavelet-based functions, as well as auto-regressive coefficients. After feature fusion with SVM, the highest accuracy of 93.33% for three-class classification was obtained (SVM) [13]. Daqrouq *et al.* proposed the employment of wavelet energy to characterize ECG signals and ARR. The percentage energy (PE) of terminal wavelet packet transform (WPT) sub signals was used in the analysis to derive wavelet-based features for CHF [14]. Singh *et al.* suggested a model for cardiac ARR diagnosis. Three filter-based feature selection methods were applied to the cardiac ARR dataset using three separate machine learning methods, and the best features were picked. Feature selection is a crucial preprocessing phase in identifying effective factors in the diagnosis of ARR patients. As a consequence, the underlying health causes for heart-related deaths may be established. The output of feature selection methods was evaluated using SVM and random forest. The random forest classifier obtained the highest accuracy of 85.58% percent [15]. Mütevelli *et al.* proposed a method for extracting features from ECG signals based on the frequency domain DWT method. To extract DWT features, wavelet packet analysis was used. Wavelet packet analysis' benefit has been highlighted in that it decomposes all approximations and information at all levels to achieve complete sub-band decomposition. Each signal was subjected to a 4-level wavelet packet decomposition, yielding 16 sub-bands. However, since the approximation coefficients reflect the key characteristic of each heart signal, the approximation coefficients from the low frequency variable are preferred, and eight of these sub-

bands were used. Then there are some statistical features are extracted from different wavelet sub-bands [8]. Haoren Wang *et al.* in this article, the effect of using a dual fully-connected neural network model for accurate heartbeat classification was investigated. The classes were normal beats (N), supraventricular ectopic beats (S), ventricular ectopic beats (V), fusion beats (F), and unknown beats (Q). The tests show that the proposed approach is effective at detecting ARRs [16]. Çınar and Tuncer *et al.* This study proposed a deep learning system with high precision and popularity for the classification of ECG signals with Regular Sinus Rhythm (NSR), Pathological ARR, and CHF. The proposed architecture was designed using a hybrid Alexnet-SVM. There are 192 ECG signals in total, with 96 ARR, 30 CHF, and 36 NSR signals available. The SVM and KNN algorithms were used to classify the classification efficiency of deep learning techniques, ARR, CHR, and NSR signals, and afterward the signals were classified in their raw form using the LSTM algorithm (Long Short Time Memory). The spectrograms of the signals are obtained using the Hybrid Alexnet-SVM algorithm. The Hybrid Alexnet-SVM algorithm is applied to the images after obtaining the spectrograms of the signals. The performance results revealed that their proposed deep learning architecture outperformed traditional machine learning classifiers [2]. Wady *et al.* point out the improvement in replacing traditional DWT for feature extraction with SLT calculated from neutrosophic set for images of brain tumor. Thus, a new composite NS-SLT model was proposed as a source for obtaining statistical texture features which was efficiently used for binary classifies a brain tumor malignancy [7].

2. METHODS

The main stages of the proposed system are depicted in the Fig. 2. The first stage is preprocessing, then features extraction, features normalization, features selection ECG signal classification and finally, performance evaluation.

2.1. Preprocessing

Three ECG signs, ARR, Normal Sinus, and CHF are investigated in this article. There are a total of 162 ECG signals of ARR data have taken from the MIT-BIH ARR Database on Physio.Net. 96 of them are used for ARR, 30 are regular sinus signals, and 36 are CHF signals. Fig. 2 provides an illustration of ECG signals for ARR, CHF, and normal sinus rate. The previously described database is preprocessed prior to introduce it to the classifiers; it is originally made up of 162 records with 65536 samples per record, so to increase

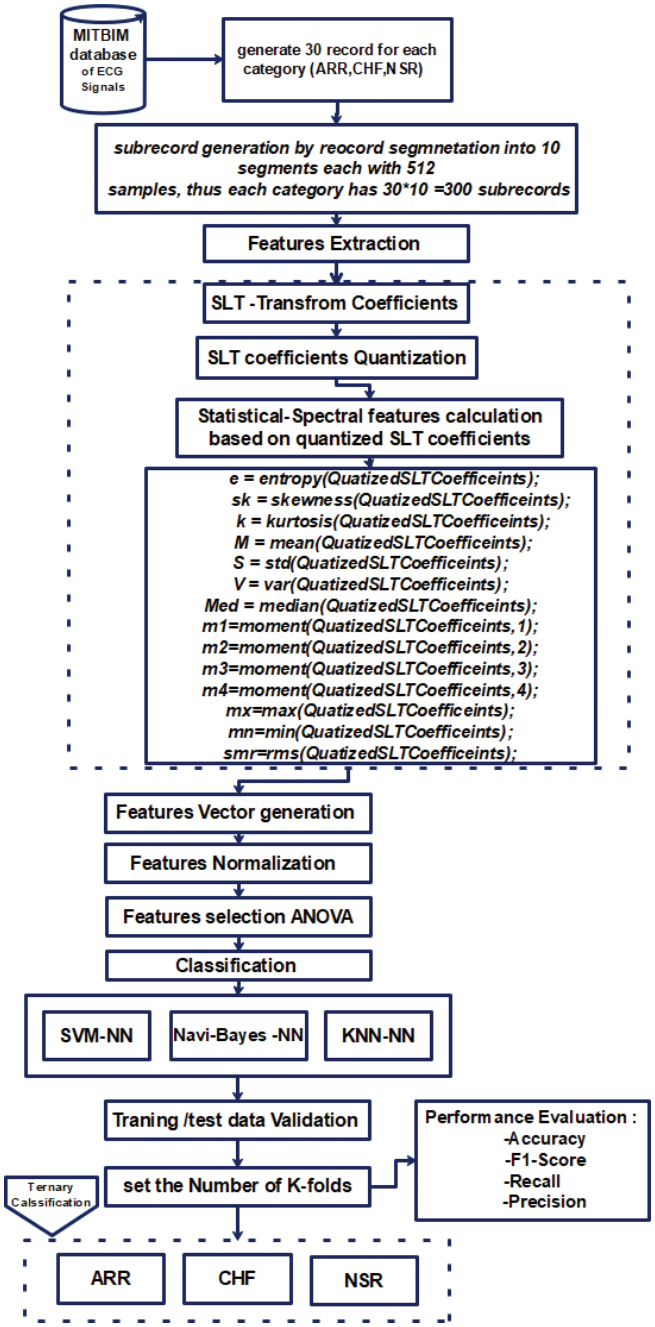

Fig. 2. Proposed electrocardiogram classifier system.

the data and to reduce the processing time. Each record has been chucked into small segments each of 512 samples.

To ensure fair comparison, a 30 records (each with 10 sub-records of 512 sample) have been extracted from each of ECG signal the (ARR, CHF, and NSR). Eventually each category has 300 sub-records, leads to totally 900 sub-records which has been provided to the training stage (750 sub-

records) and (250 sub-records) were used by testing stage the next subsections will describe the details of each block as depicted in Fig 2. However, Fig 3. displays signals samples for each category (ARR, NSR, CHF) proposed ARR, CHF, NSR, and selected randomly from the database signals.

2.2. Features Extraction and Selection

The classification techniques usually start by the stage of extraction of relevant features [4]. For instance, SLT transform coefficients based statistical features extraction which defines the distribution signal energy in time and the frequency domains have been investigated in this work. Following DWT implementation, the SLT transform filter banks have a parallel structure. In several of these parallel divisions, DWT uses a product form of simple filters, and the filter bank "Slantlet" uses a similar structure in parallel. The part filter branches, on the other hand, do not have a product form of implementation, giving SLT an advantage. SLT will create a filter bank with each filter's length in power of two, for a mathematical perspective of SLT transformation, consider a simplified representation of Fig. 3 for (l) scales. This results in a periodic output for the analysis filter bank and reduces the samples (2^i-2) which

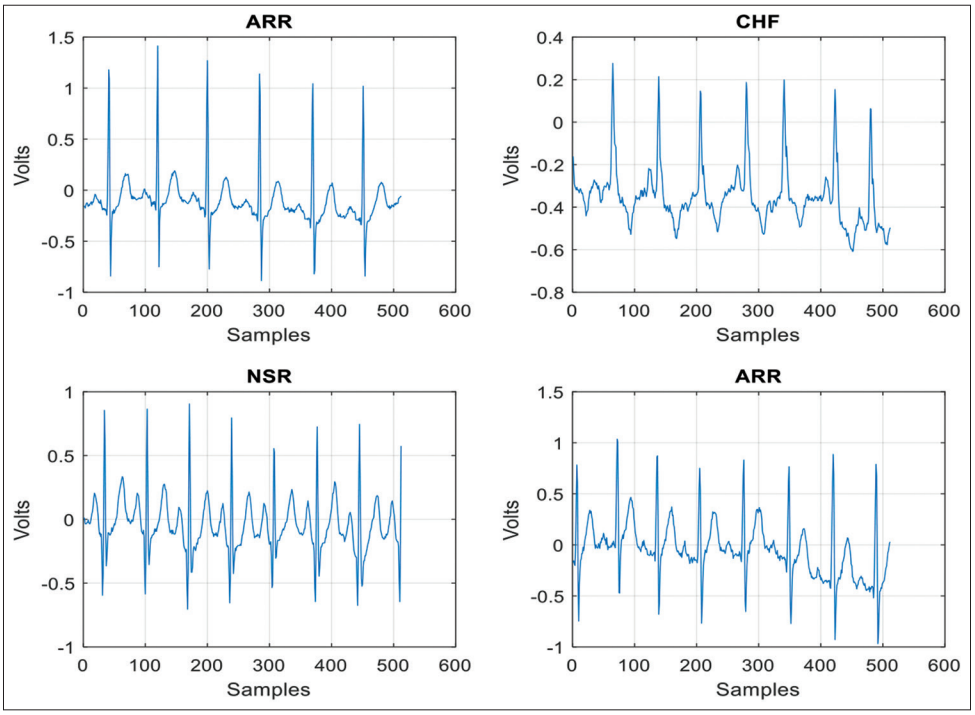

Fig. 3. Electrocardiogram sample from each class (Arrhythmia, normal sinus rhythm, and congestive heart failure).

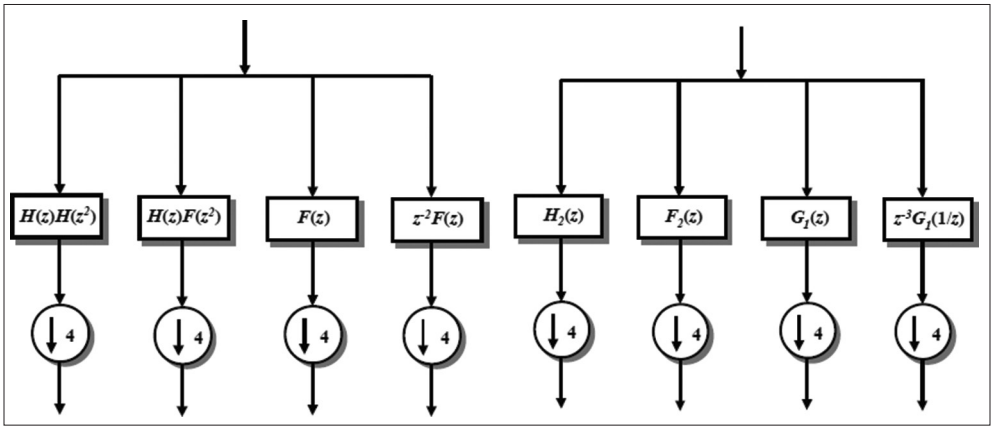

Fig. 4. The two-scale iterated D2 filter bank (on the left) and the two-scale Slantlet transform filter bank (on the right) (right-hand side) [17].

support approaches one thirds, as (i) increases. The filters in scale (i) must be $g_i(n)$, $f_i(n)$, and $h_i(n)$ to analyze the signal where each filter has an appropriate 2^{i+1} support. For (l), the SLT filter bank uses (l) number of pairs of channels, that is, ($2l$) channels in total. The low pass $h_i(n)$ filter is then combined with its adjacent $f_i(n)$ filter, where a down sampling of 2^i is followed by any filter. The channel pairs of each ($l-1$) constitute a $g_i(n)$, followed by a down sampling by 2^{i+1} and the down sample by a reversed time version $i=1,2,3\ldots,l-1$. The following expressions are represented by: as the filters $g_i(n)$, $f_i(n)$, and $h_i(n)$ implement linear forms in pieces [7], [17]:

$$g_i(n)=\begin{Bmatrix} a_{0,0}+a_{0,1}n,\ for\ \ n=0,\ldots\ldots 2^i-1 \\ a_{1,0}+a_{1,1}n,\ for\ \ n=2^i,\ldots\ldots 2^{i+1}-1 \end{Bmatrix} \quad (1)$$

$$h_i(n)=\begin{Bmatrix} b_{0,0}+b_{0,1}n,\ for\ \ n=0,\ldots\ldots 2^i-1 \\ b_{1,0}+b_{1,1}n,\ for\ \ n=2^i,\ldots\ldots 2^{i+1}-1 \end{Bmatrix} \quad (2)$$

$$f_i(n)=\begin{Bmatrix} c_{0,0}+c_{0,1}n,\ for\ \ n=0,\ldots\ldots 2^i-1 \\ c_{1,0}+c_{1,1}n,\ for\ \ n=2^i,\ldots\ldots 2^{i+1}-1 \end{Bmatrix} \quad (3)$$

To correctly classify ECG signals requires generation of the feature vector which contains features both in the time domain and the frequency domain.

The trace of amplitude for more than 550 SLT coefficients from each ECG category are illustrated by Fig.5

The SLT quantized coefficients are used to extract the statistical-spectral features such as mean, standard deviation, variance, entropy, maximum, minimum, kurtosis, momentum, median, skewness, and root mean square error values of each ECG signal. This technique results in reduction in the length of the feature vector used as an input to a classifier.

The original database for each class is a vector of 300 ECG signal sub-record, thus the initial dimension is {y1, y2., y300}. Then after by calculating statistical features from SLT coefficient a features vector of 14 dimensions (14-features) {S1, S2., S14}is generated for each element y_n (for each of the three class) [Table 1]. Each feature vector corresponds to a single point in the feature space. Points of the same class should be closer together, and points of different classes should be apart. A normalization has been applied on the input features sets before the feature selection stage. The features selection stage is considered as a final stage in feature extraction and processing procedure [18]. It is aimed to improve the performance of a classifier and achieve a minimum classification error. The analysis of variance (ANOVA) methodology has been used in this study to

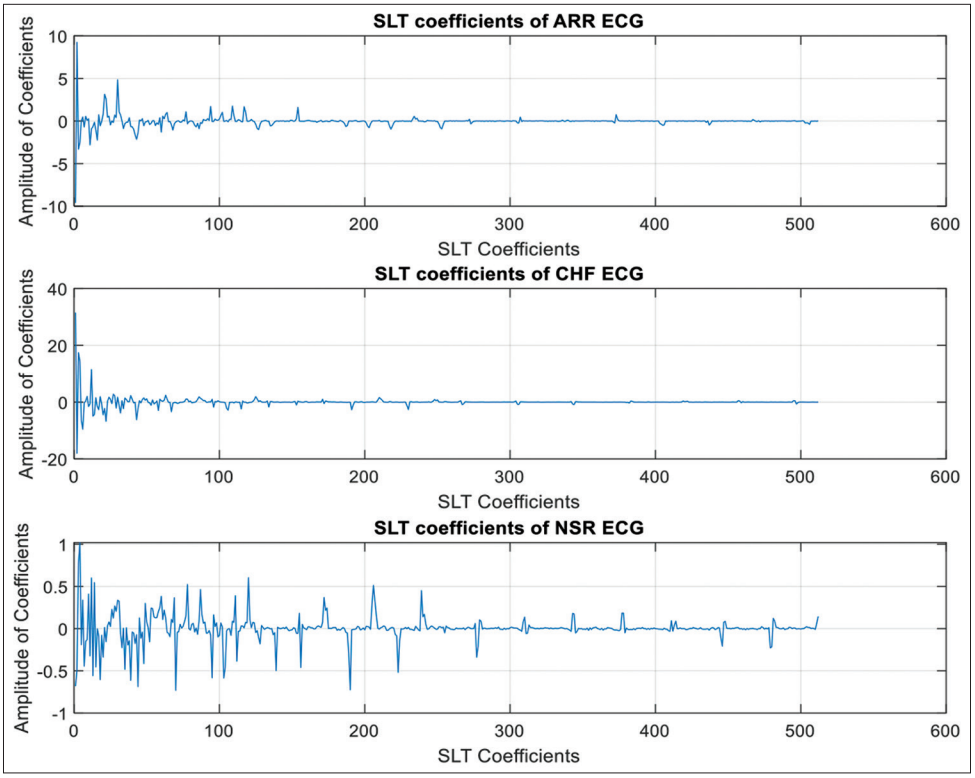

Fig. 5. Slantlet transform coefficients for classes (Arrhythmia, congestive heart failure, and normal sinus rhythm).

TABLE 1: Statistical features extracted from SLT coefficients.

Statistical feature	Formula/Description
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
Standard deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$
Maximum	$Maximum = \max(x_i, 1 \le i \le N)$
Minimum	$Minimum = min(x_i, 1 \le i \le N)$
Skewness	$S = \frac{1}{N}\frac{\sum_{i=1}^{N}(x_i - \mu)}{\sigma^3}$
Kurtosis	$K = \frac{1}{N}\frac{\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4}$
Variance	$va_r(x) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
Tropy	$H(x) = \sum_{i=1}^{N} p(x_i) log_2(\frac{1}{p(x_i)})$
Momentum	$momentum(k) = E(\boldsymbol{x} - \mu)^k$
Root Mean Square	$x_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i)^2}$
Median	Middle value separating the greater and the lesser halves of the set of data

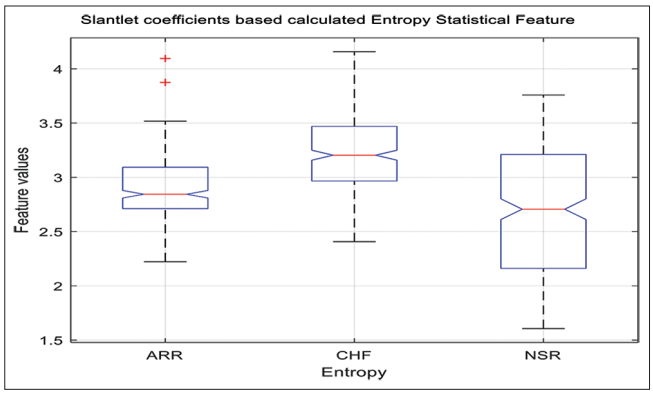

Fig. 6. Distribution of sample of relevant feature.

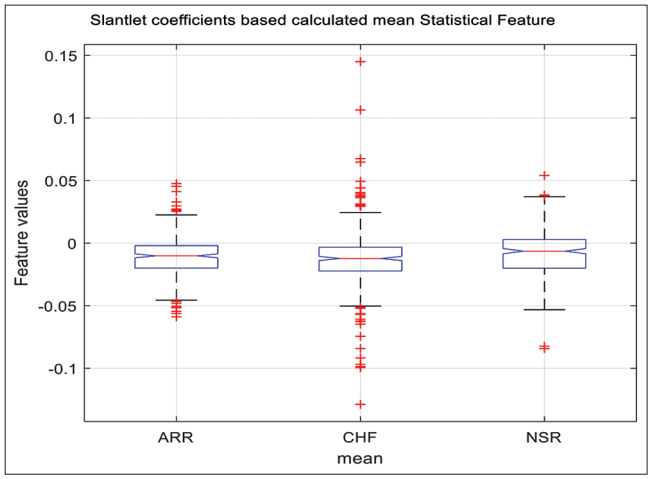

Fig. 7. Distribution of sample of irrelevant feature.

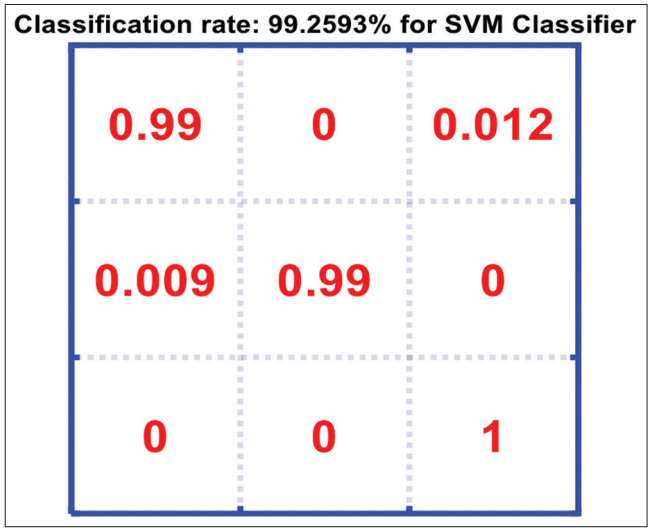

Fig. 8. Confusion matrix and accuracy for support vector Machine-NN.

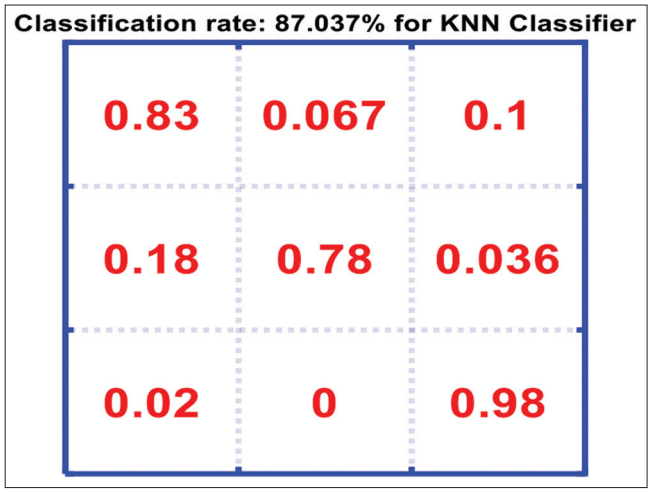

Fig. 9. Confusion matrix and accuracy for k-nearest neighbor-NN

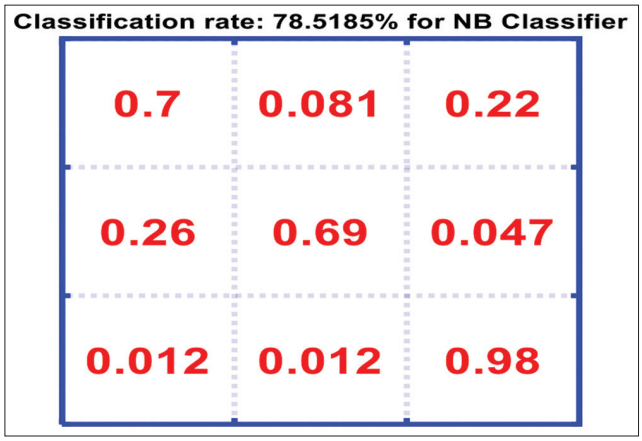

Fig. 10. Confusion matrix and accuracy for Naive Bayes-NN.

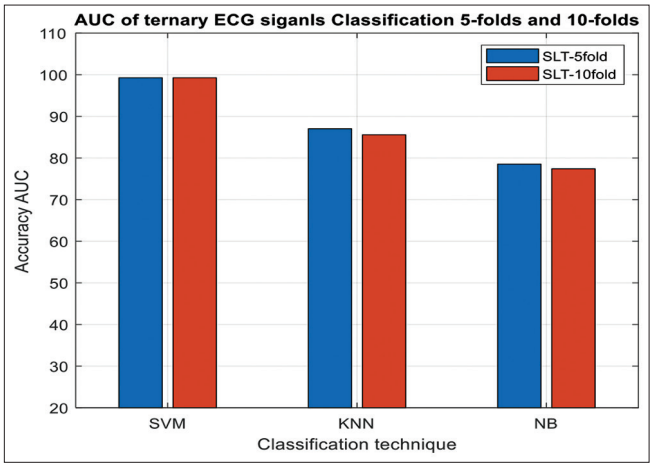

Fig. 11. Area under the curve of three classes two cross-validations.

minimize the dimension of data based on its importance and variance while preserving as much information as possible. ANOVA is a useful technique for deciding if two or more sets of data vary statistically [7]. The ANOVA test with *P*-value of 10^{-3} selects 12 out of 14 input features vector dimension. Hence, two features are cancelled based on their *P*-value $>10^{-3}$. The two features omitted are SLT based mean, and SLT based first momentum Fig. 6 shows the distributions of the sample of relevant feature form relevant feature space of dimension 12 (SLT entropy), whereas Fig. 7 shows on of the irrelevant features (SLT mean), from the figures, the difference between relevant and irrelevant features can be noticed easily

2.3. ECG Classification

The SLT transform has been used to improve the performance of SVM, KNN with, K=3 [13] [13], NB classifiers [19]. For instance, each classifier has been tested with different number of fold (5-fold or 10-folds) cross validation to measure the performance of proposed features space.

2.4. Performance Tests

The proposed Ternary ECG classification system is assessed for classification task performance using a variety of metrics. These figures come from the confusion matrix that describes the classes. The confusion matrix is a table that is often used to calculate the performance of a class predictor or classification model on a set of test data for which the true/actual values are unknown Accuracy is the proportion of correctly classified predictions (i.e., true

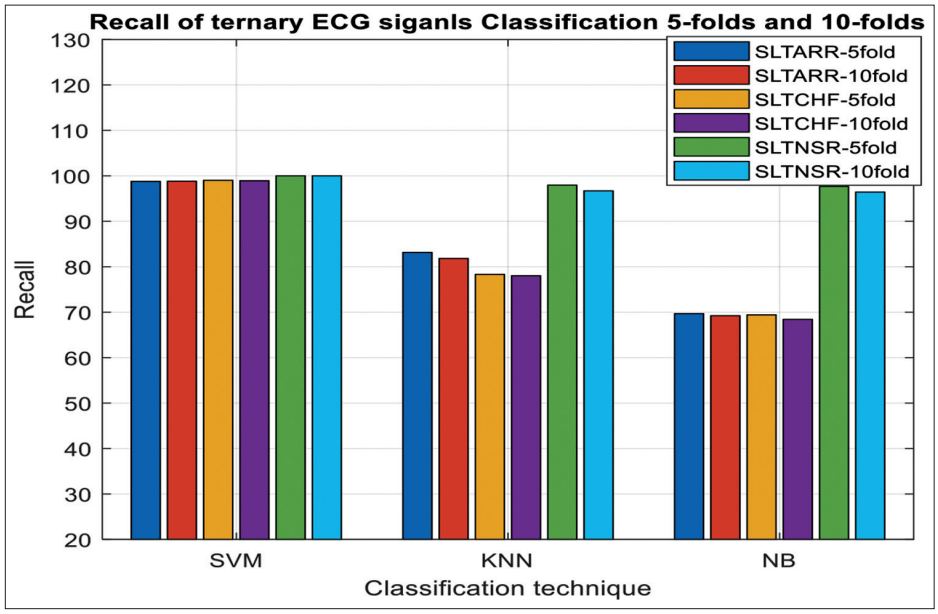

Fig. 12. Precision of three classes two cross-validations.

positive and true negative) over the total number of cases examined.

$$\text{Accuracy} = (TP+TN)/(TP+FN+TN+FP) \quad (4)$$

Where TP=true positive, TN=true negative, FP=false positive and FN= false negative. The accuracy is defined as the percentage of positive class predictions that are actually positive class predictions. Precision is a measure of how accurately target areas are extracted when compared to the ground truth.

$$\text{Precision} = TP/(TP+FP) \quad (5)$$

The recall (also known as sensitivity or true positive rate, TPR) is the proportion of positive class forecasts to the total number of positively classified units. The memory is a measure of how well the extracted target represents the ground truth.

$$\text{Recall} = TP/(TP+FN) \quad (6)$$

The F-score (also known as the F-measure) is a metric for evaluating the performance of problems with binary labels and different classes. The harmonic mean of macro-precision and macro-recall is the macro F-score. High macro F-scores indicate that the system performs well across all classes, while low macro F-scores indicate that classes are poorly predicted [20].

$$\text{F-Score} = 2*(\text{Precision}*\text{Recall})/(\text{Precision Recall}) \quad (7)$$

3. RESULTS AND DISCUSSION

The performance was calculated based on proposed statistical-features derived from SLT transform coefficients from more than 900 ECG signals to identify automatically the class of the ECG signal. NB, SVM, and KNN were used as a classification method. The Confusion Matrix for is shown in the diagram below. Fig. 8 shows the confusion matrix and the accuracy for the SVM classifier. It is obvious that the SVM classifier along with SLT transform features outperform the other proposed classifiers KNN and NB. For instance, a recognition accuracy of 99.2593% has been attained.

Fig. 9 depicts the confusion matrix and the accuracy for the SLT transform based features with KNN classifier.

It is clear that there is a big gap between the performances of this classifier which is 87.037% as compared to the SVM classifier along with SLT transform features. Fig. 10 gives the confusion matrix of the SLT features combined with NB classifier which gives very poor classification accuracy result of 78.5185%, which results on unrecommend it as a good choice with the proposed feature extraction scheme.

Fig. 11 gives the accuracy (AUC) of the proposed classifiers integrated with the proposed scenario of feature extraction with two cross–validations (5-folds and 10-folds), its also clear that increasing cross-validation from 5 to 10 folds has little bad impact on the accuracy results for the three proposed classifiers.

The same fact has been concluded by investigating other performance measures, for example, precision depicted in Fig. 12, recall (sensitivity) showed in Fig. 13, and finally F1-score performance results depicted in Fig. 14 below. Thus, increasing the number of folds has mitigated the performance of classification for all proposed classifiers.

As a final tool for proposed system performance evaluation a comparison has been made with the proposed system and some state of art schemes illustrated by Table 2. Thus, Singh *et al.* [15] who proposed a model for cardiac ARR diagnosis three separate machine learning approaches were used in this model to pick features (filter-based feature selection) from a cardiac ARR dataset. The highest accuracy of 85.58% percent was obtained with the random forest classifier using the gain ratio feature selection approach with a subset of 30 features, according to the experimental study. Hussain *et al.* [21] proposed a classification system based on SNN, KNN, and decision tree classification had achieved accuracy up to 97%. Nahak *et al.* [13] used wavelet transform fused features with auto-regression model was able finally to attain accuracy up to 93.3%.

TABLE 2: Literature comparison.

Research	Technique	AUC	Classes
Singh and Pradeep (2018)	Feature selection, SVM, Random forest, JRIP	A=85.5%	Arrhythmia
Hussain *et al.* (2020)	SNN, KNN, Decision Tree	A=97%	CHF
Nahak and Saha (2020	Wavelet Feature Fusion, SVM, KNN, DT, and NB	A=93.3%	ARR, CHF, NSR
Current Article	SLT Transform based Statistical features with (SVM, KNN, and NB)	A=99.25%	ARR, CHF, NSR

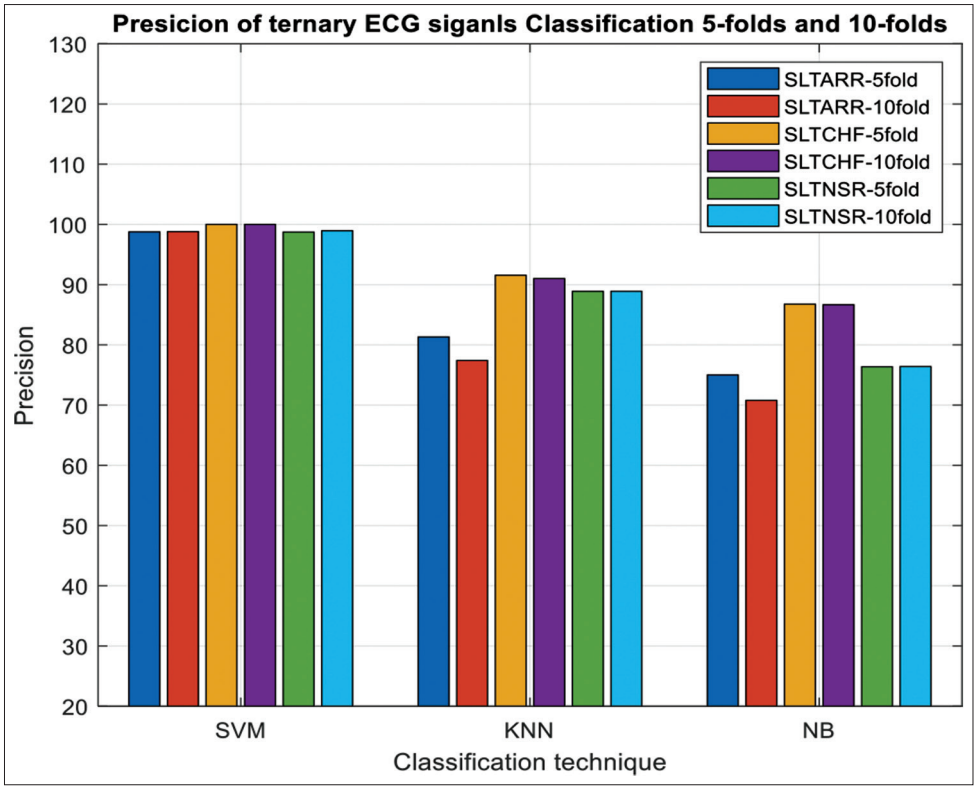

Fig. 13. Recall of three classes two cross-validations.

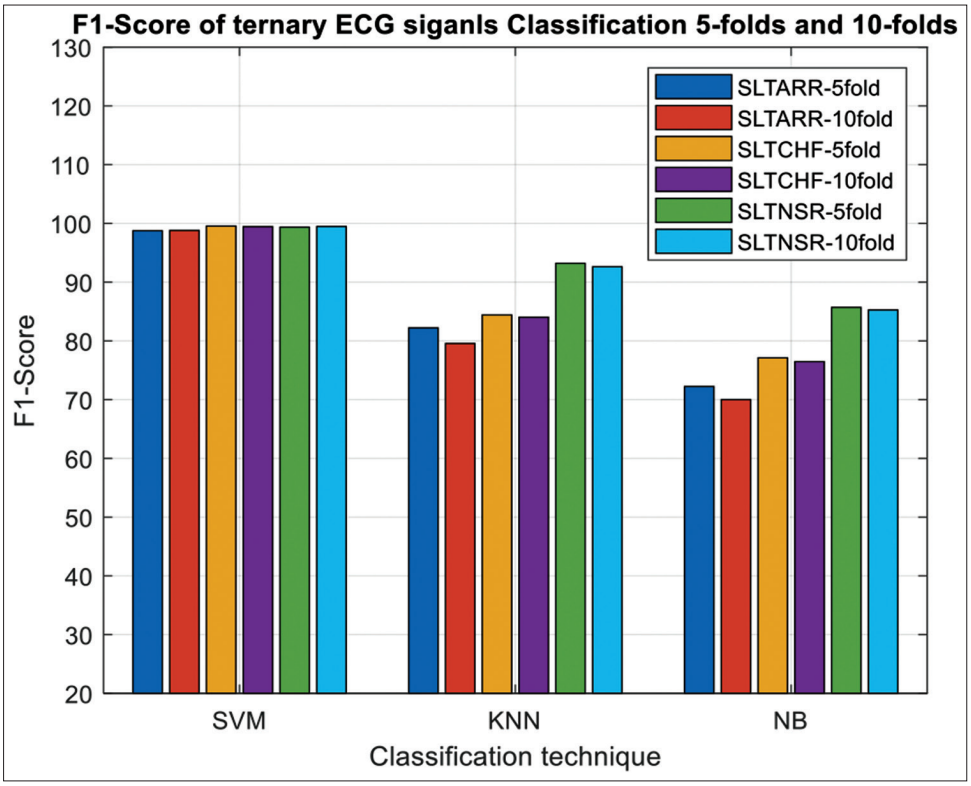

Fig. 14. F1-score of three classes two cross-validations.

4. CONCLUSION

The main goal of this article was to create a combined features extraction and machine learning intelligent system that could automatically distinguish three different types of ECG signals: ARR, CHF, and regular sinus rhythm (NSR). The experiments were carried out on 90 ECG signal recordings (900 sub-records or segments) collected from a publicly accessible database. Fusion with wavelet and AR features improved the performance. Three classifiers were investigated (SVM, KNN, and NB) in this study, and ultimate accuracy of 99.256% was obtained from SVM classifier. The simulation results highly recommended SLT transform based statistical features extraction and showed that increasing the cross-validation folds from 5 to 10 has bad impact on the performance results from different metrics. Furthermore, the NB–NN classifier gave very poor results as compared to other two classifiers. Eventually the proposed technique outperform the accuracy attained by other similar literatures

5. ACKNOWLEDGMENT

The authors would like to thank Salahaddin University–Erbil for supporting this article.

REFERENCES

[1] 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.

[2] A. Çınar and S. A. Tuncer. "Classification of normal sinus rhythm, abnormal arrhythmia and congestive heart failure ECG signals using LSTM and hybrid CNN-SVM deep neural networks". *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 1, no. 1, pp. 1-12, 2020.

[3] V. Jahmunah, S. L. Oh, J. K. En Wei, E. J. Ciaccio, K. Chua, T. Ru San, U. R. Acharya. "Computer-aided diagnosis of congestive heart failure using ECG signals a review". *Physica Medica*, vol. 62, pp. 95-104, 2019.

[4] A. Subasi. "*Practical Guide for Biomedical Signals Analysis Using Machine Learning Techniques*". Academic Press, Cambridge, Massachusetts, 2019.

[5] 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.

[6] M. Maitra, A. Chatterjee and F. Matsuno. "A Novel Scheme for Feature Extraction and Classification of Magnetic Resonance Brain Images Based on Slantlet Transform and Support Vector Machine". *SICE Annual Conference*, pp. 1130-1134, 2008.

[7] S. H. Wady, R. Z. Yousif and H. R. Hasan. "A novel intelligent system for brain tumor diagnosis based on a composite neutrosophic-slantlet transform domain for statistical texture feature extraction". *BioMed Research International*, vol. 2020, p. 8125392, 2020.

[8] M. H. Mütevelli and S. Ergin. "The usage of statistical features in the approximation components of wavelet decomposition for ecg classification: A case study for standing, walking and single jump conditions". *Electronic Journal of Vocational Colleges*, vol. 8, pp. 178-182, 2018.

[9] M. R. Diniari and S. M. Isa. "Electrocardiogram classification for arrhythmia using convolutional neural network 2D and adabound optimizer". *The International Journal of Recent Technology and Engineering*, vol. 8, no. 5, pp. 1277-1284, 2020.

[10] S. Nibhanupudi, R. Youssif and C. Purdy. "Data-specific Signal Denoising Using Wavelets, with Applications to ECG Data". *International Midwest Symposium on Circuits and Systems.* vol. 3, pp. 20-23, 2004.

[11] S. K. Sahoo, A. K. Subudhi, B. Kanungo and S. K. Sabut. "Feature extraction of ECG signal based on wavelet transform for arrhythmia detection. *International Conference on Electrical, Electronics, Signals, Communication and Optimization EESCO 2015*, no. December 2018, 2015.

[12] A. J. Chashmi and M. C. Amirani. "An efficient and automatic ECG arrhythmia diagnosis system using DWT and HOS features and entropy-based feature selection procedure". *The Journal of Electrical Bioimpedance*, vol. 10, no. 1, pp. 47-54, 2019.

[13] S. Nahak and G. Saha. "A Fusion Based Classification of Normal, Arrhythmia and Congestive Heart Failure in ECG". *26th The National Conference on Communications*, pp. 1-6, 2020.

[14] K. Daqrouq and A. Dobaie. "Wavelet based method for congestive heart failure recognition by three confirmation functions". In: *Computational and Mathematical Methods in Medicine*. Taylor Francis Online, Milton Park, 2016.

[15] N. Singh and P. Singh. "Cardiac arrhythmia classification using machine learning techniques". In: *Engineering Vibration, Communication and Information Processing*. Springer, Berlin, Germany, 2019, pp. 469-480.

[16] H. Wang, K. Lin, H. Shi and C. Qin. "A high-precision arrhythmia classification method based on dual fully connected neural network". *Biomedical Signal Processing and Control*, vol. 58, p. 101874, 2020.

[17] M. Maitra and A. Chatterjee. "A Slantlet transform based intelligent system for magnetic resonance brain image classification". *Biomedical Signal Processing and Control*, vol. 1, no. 4, pp. 299-306, 2006.

[18] S. O. Haji and R. Z. Yousif. "A novel run-length based wavelet features for screening thyroid nodule malignancy". *The Brazilian Archives of Biology and Technology*, vol. 62, pp. 1-17, 2019.

[19] H. Zhang. "The Optimality of Naive Bayes". In: *Proceedings of the 7th International Florida Artificial Intelligence Research Society Conference*, vol. 2, pp. 562-567, 2004.

[20] M. Sokolova, N. Japkowicz and S. Szpakowicz. "Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation". *AAAI Workshop Technical Reports*, vol. 6, pp. 24-29, 2006.

[21] L. Hussain, I. A. Awan, W. Aziz, S. Saeed, A, Ali, F. Zeeshan and K. S. Kwak, *et al.* Detecting congestive heart failure by extracting multimodal features and employing machine learning techniques. *Biomed Research International*, vol. 2020, p. 4281243, 2020.

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

جامعة التنمية البشرية
UNIVERSITY OF HUMAN DEVELOPMENT

UHD Journal of Science and Technology

A Scientific periodical issued by University of Human Developement

Vol.5 No.(1) June 2021

2021 2721

e.mail:jst@uhd.edu.iq