Neuroscience Informatics 2 (2022) 100097

Contents lists available at ScienceDirect

Neuroscience Informatics

www.elsevier.com/locate/neuri

Original article

A new feature extraction approach of medical image based on data distribution skew

Farag Hamed Kuwil [a,b,*]

[a] Department of Computer Engineering, Karabuk University, Karabuk, Turkey
[b] Department of Computer Engineering, Tripoli University, Tripoli, Libya

ARTICLE INFO

Article history:
Received 15 May 2022
Received in revised form 26 July 2022
Accepted 1 August 2022

Keywords:
Feature reduction
Feature extraction
Medical images
Classification

ABSTRACT

Building a highly efficient machine learning model requires sufficient data to allow robust feature extraction capable of recognizing patterns in each class; thus, the model can distinguish among different classes. It is important to extract effective features from the available amount of data without the need for more real data or improve them using an augmentation technique. The matter gets more complicated if the data is of the image type. In this paper, a new approach for feature extraction called Feature Extraction Based on Region of Mines (FE_mines) is presented that includes three versions to deal with different medical images; this approach obtains multiple formulas for each image using the signal and image processing, then data distribution skew is used to calculate three statistical measurements that include the hidden features, which leads to increased discrimination among classes to build powerful models with better performance and high efficiency. Three experiments were conducted using three types of medical image datasets, namely: Diabetic Retinopathy (Color Fundus photography); Brain Tumor (MRI); and COVID-19 chest (X-ray). The results proved that the FE_mines approach achieved higher accuracy ranges (1 to 13)% within the three experiments than the two traditional methods (RGB and ASPS approaches). In addition, an augmentation technique to increase the size of the dataset is not required which has negative effects on performance. Furthermore, the approach simultaneously included three preprocessing techniques: feature selection, reduction, and extraction.

1. Introduction

Machine Learning (ML) models are strong tools that may be utilized to achieve vital jobs and solve complicated issues quickly and efficaciously. As data is the fuel or oxygen that ML lives on, data has grown at an exponential rate in the modern world, meaning that more methods and approaches are needed to organize and process them to make them effective when used to build models. Data size, randomness, hidden patterns, and weak attributes are great challenges that lead to the necessity of increasing the development and improvement of ML algorithms to overcome these challenges within the data, or in other words, providing a minimum set of organized data that contains clear patterns to allow algorithms build strong and effective models.

The diversity of data is called data type, which is a classification system that defines which type of values a variable may store, and which type of mathematical, relational, and logical operations can be performed on it without causing an error or resulting outputs that are difficult to interpret. It is critical to understand the appropriate data types for labels and features. Quantitative and qualitative data types are the main categories of data [1]. They have a big impact on data mining and machine learning applications, such as Natural Language Processing (NLP). In decision-making problems, topics connected to the normalization of qualitative and quantitative variables were studied, with the introduction of a property that allows the order of alternative evaluations to be preserved. Another domain is urban soundscape evaluation [2]. Other data types that are referred to as data measures are nominal, ordinal, interval, and ratio. The branches of ML can be identified according to the data distribution. When the label is available with the data, it is called supervised learning, but if the label is absent, it is called an unsupervised learning problem. Furthermore, these continuous and discrete label values can determine whether it is a regression or a classification problem.

Data mining techniques are used to process data to build a model; it comprises data cleaning, data selection, data transformation, ... etc. There are other techniques, such as feature reduction and feature extraction, which are used to prepare a dataset to the minimum size with strong patterns. Many algorithms and methods

[*] Correspondence to: Department of Computer Engineering, Karabuk University, Karabuk, Turkey.
E-mail address: kuwil73@gmail.com.

https://doi.org/10.1016/j.neuri.2022.100097

with their application within different domains have been presented, such as Principal Component Analysis (PCA) depending on linear algebra [3], [4]. In addition, the hybridization and combination of diverse methods have been used to enhance features that represent the object effectively. The multi-strategy feature selection and grouped feature extraction were combined to develop a novel method based on fast hybrid dimension reduction via incorporating their advantages of removing irrelevant and redundant information. The reduction of the data attributes improves the efficiency compared to the contrastive methods [5]. In addition to using hybridization to identify and reduce features, optimization techniques are also used to do so. The optimization of feature reduction was used in the economic domain to build a predictive model of the stock market. The model performance was compared with Principle Component Analysis (PCA), Factor Analysis (FA), Genetic Algorithm (GA), and Firefly-based prediction [4]. Another literature related to the Features or Attribute collects the most important methods which are based on a variety of approaches, such as statistical and geometric [6], [7].

Medical image classification is one of the specialties that combine machine learning and Computer Vision (cv), where medical images use special equipment, while cv uses public or surveillance cameras in the public area such the streets and airports, and others. Computer vision and machine learning are complementary; cv employs ML approaches to automate the acquisition of visual models, translate signals into symbols, and construct trainable image processing systems. In other words, it can be said that medical image classification is a special branch of Computer Vision, as dealing with a specific type of image to discover a specific disease. It is easy to reduce the size of the image to 0.1% to classify. Several studies have been published on image classifications within various domains using different algorithms. Image classification is widely used in different domains in ML, such as engineering, botany, landscape, fonts or calligraphy, and others. Mary and Dharma have introduced an improvement for feature extraction of Coral reef images using ANN, KNN, and SVM algorithm introducedoduced approach minimized the bin size of the histogram, thus, the time complexity is reduced, and the rate of recognition is improved [8]. The spectral-spatial information fusion methods are summarized into three sets; 1) segmentation map is used for the relaxation of the pixel-wise classification, 2) feature fusion approach classifiers, 3-D spectral-spatial feature extraction, and classifiers based on deep learning, 3) decision fusion-based methods [9]. New automatic defect detection and classification platform for solar cells images called Deep Featureweresed (DFB) method were introduced in which features extracted from these images through deep neural networks are classified with ML methods such as SVM, K-NN, DT, RF, and Naive Bayes [10]. A novel local feature descriptor was introduced to locate various patterns and double directional relation patterns for face recognition of age invariant. The approach achieved high performance and outperformed the existing age invariant face recognition such as FGNET and MORPH datasets [11]. A novel approach to feature extraction is based on the attitude of the edge pixels between a black pattern and a white background within the image. Experiments on an Arabic calligraphic script image for optical font recognition showed that a decision tree classifier can boost the overall performance of optical font recognition [12]. Sachar and Anuj provided a thorough overview of the various strategies utilized in computer vision for automated plant identification using leaf images [13]. In-plant leaf disease classification, the EfficientNet deep learning architecture was suggested, and its performance was compared to that of existing state-of-the-art deep learning models [14].

In terms of the classification of medical images and their characteristics, which is the subject of this paper, many types of research in diagnosing different parts of the human body have been presented, such as the comprehensive and accurate study of the applications of artificial intelligence in ophthalmopathy, and the fundamental imaging techniques used for early diagnosis and therapy of eye illnesses. The most significant are the photography of Fundus digital photography, tomography of optical coherence (OCT), and the ultra-widefield (UWF) [15]. The Coding Network with Multilayer Perceptron (CNMP) is a new approach based on a deep learning model that integrates and combines high-level features that are extracted from a CNN and some traditional selected features. Two medical image datasets (HIS2828 and ISIC2017) were used to evaluate the CNMP method, where the accuracy reached 90.1% and 90.2%, respectively, which are better than peers [16]. A medical image processing method based on multi-features fusion was introduced by [17] which has a high impact on feature extraction within the following medical images: chest, brain, liver, and lung, and can best express the composition of these images. In a detailed study of the classification of COVID-19 disease using X-ray images, CNN was used to build a model referring to different types of pneumonia, with feature optimization transferable multi-receptive. The study resulted in an accuracy range of 90% to 97% [18]. To train Deep Learning models that produce radiologist text from pictures, a comprehensive literature study was conducted on multimodal datasets by [19]. This topic is critical because these approaches can give new diagnostic criteria rapidly and accurately by providing unobservable data from pictures and text [20]. The goal of the study was to examine how the CNN model's segmentation influences magnetic resonance imaging (MR) for Alzheimer's diagnosis. The segmentation results of the fully CNN and SVM algorithms were compared with those of the CNN model, which was used to segment the MR imaging of Alzheimer's patients. The experimental group had a higher clinical dementia rating (CDR) score and a lower mini-mental state examination (MMSE) score, and it was discovered that the CNN model had greater segmentation precision [21]. The introduction of robots was identified and proposed to take up the challenge of how can Robots undertake human-like activities, and solve the problem in fighting the pandemic of COVID 19. The available literature was reviewed through some search engines. A comprehensive review of the literature identified different types of robots being used in the medical field. Therefore, several vital applications for Robots in the management of the COVID-19 pandemic are used gainfully to deliver medicine, food, and other essential items to COVID-19 patients who are under quarantine [22]. This article discusses recent studies that use ML and AI to support researchers in various ways. It also discusses a few mistakes and difficulties that might occur when applying such methods to actual scenarios. In addition, recommendations are presented for politicians, medical professionals, and researchers on model design in the current context of combating the Covid-19 epidemic and in the future [23]. The typical ANFIS is inappropriate for complicated human jobs that call for the careful manipulation of computers and systems, according to this paper's conclusion. The application of ANFIS in the expanding field of engineering sciences has been the main emphasis of the state-of-the-art and practical research problems that have been explored. When combined with metaheuristic methods and further regulated using nature-inspired algorithms through calibration and parameter adjustment, the basic ANFIS architecture was significantly enhanced.

Using computer techniques, a brain-computer interface (BCI) framework may control external devices and identify patterns in mental activity. An EEG-based BCI system's output is evaluated through the classification of EEG signals for various purposes. Researchers used machine learning (ML) and deep learning (DL) methods to categorize EEG-based BCI as a result of the development of artificial intelligence technologies. As a result the brain-computer interface is able to gain knowledge from the subject's brain with each new session. By adjusting the rules that were established for categorizing ideas, the system's effectiveness is increased [24]. S. Aggarwal and N. Chugh provide a focused overview on the use of several ML/DL algorithms in EEG-based BCI. A study aimed at quantifying neurological EEG-biomarkers was presented by I. Hussain et al., where five classes of sleep phases were predicted using sleep EEG data. For the purpose of classifying the various stages of sleep into many categories, the C5.0, Neural Network, and CHAID machine-learning models' accuracy was 91 percent, 89 percent, and 84 percent, respectively. A wearable sleep monitoring system is anticipated to use the EEG-based sleep stage prediction technique [25].

The purpose of this paper is to look at the concerns of quality of extraction of the features from medical images, and to develop a model with a limited number of attributes to conduct the classification in less time and with better accuracy. Skewness is used to locate mines of features using the statistical concept of data distribution. The mines could be within negative or positive skewness areas that contain hidden features; thus, distinguishing images is easier through three statistical measures: the mean (μ), the standard deviation sigma (σ), and the coefficient of variance (cv). Feature selection, feature reduction, and feature extraction are three technical concepts that are combined to produce the proposed approach. The method called Feature Extraction is based on the Region of Mines (FE_mines) andconsists of four stages: reading images and converting them to extra formal and determining the mines of hidden features and calculating features within the mines, which correspond to the following techniques: Image processing, Signal processing, Skewness regions, and three statistical measures, respectively. Three versions were developed to be compatible with different types of medical images. Three healthcare datasets have been selected as a domain, namely: Diabetic Retinopathy; Brain Tumor; and COVID-19 chest to illustrate the characteristics of the proposed approach and compare them with previous traditional methods such as RGB and ASPS. The Random Forest algorithm was also chosen to build the models.

The contribution of this paper can be summarized in the following points

1) Introducing a new method based on the skewness concept for extracting medical image features.
2) This method based on qualitative, not quantitative, to increase both performance and efficiency of the model.
3) Using the trait mine term to extract hidden features in ML.
4) The amount of color can determine the appropriate technique to represent the image and then extract its features.

The rest of this paper is ordered as follows: Section 2 presents the proposed approach. Analysis of the methodology was presented in Section 3. Then, the experiments and results were presented in Section 4, and the conclusion and future work were presented in Section 5.

2. The methodology

Medical image diagnosis in Machine Learning (ML) has a high priority within the healthcare domain because of the low cost and high accuracy. The diversity of medical images results from different types of medical imaging equipment; therefore, it is necessary to find various methods and techniques compatible with each type of image (species of image), such as CT-scan, MRI, Ultrasound, X-rays, Fundus photography... etc. This diversity, besides size and image types, represents a challenge in medical image classification. Each image is represented by a 3-D matrix (three channels), and each element within represents a pixel. Each pixel is encoded as an integer from 1 (white) to 255 (black): the bigger this value, the dusky the color. Although images can be read in a unique formula on a computer, however, they can be represented in different formulas, such as Grayscale, and Binary. Furthermore, they can be split into their base component (R, G, B) or converted to other formulas using Signal processing techniques. Many methods of feature extraction were introduced in the ML field based on different strategies; four of the most significant have been addressed, namely Geometric, Statistical, Texture, and Color features (Mutlag et al., 2020). Statistical strategy is selected in this paper to overcome the challenges of dealing with various medical images by developing three methods of feature extraction from images.

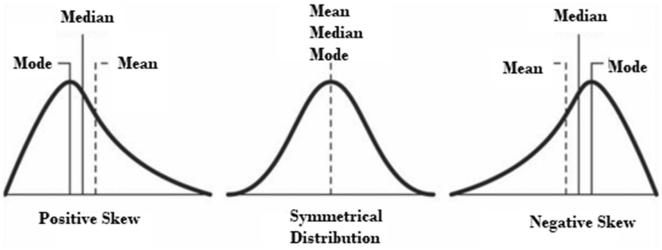

Fig. 1. Three scenarios of skewness.

This study concentrates on images in the healthcare domain where medical image diagnosis is one of the difficult classification problems since the discrimination among medical images is tiny in most cases. The statistical methodology depends on data distribution into different formulas (matrices) of image representation, where Measures of Central Tendency (MCT) and Dispersion are the basis of investigating data distribution. Depending on that, the disease is concentrated in a specific region of the image called Mines of Attributes represented by the contrast or intensity of the color. Therefore, calculating statistical coefficients within these mines has a strong indication of the presence of disease. This is done by partitioning each matrix by the median into two regions, namely upper median, and lower median, where two additional virtual channels are obtained within each main channel. To complete the idea, another concept to be illustrated beside the median is Skewness and its types, and it can be defined as a measure of the asymmetry of the data distribution assuming a unimodal distribution. Because the normal distribution has a skew of 0, it reflects how much data is skewed to one of the side edges. From here, the idea came to create virtual channels (mines) that provide an area to extract the hidden features.

It is known that the mean is a strong indicator if it is calculated within a range of data that are close to each other, or in other words when its standard deviation is small. Three statistic measurements depend on each other; the mean (μ) is the best value that can represent a set of data values, while the standard deviation sigma (σ) is a measure of the average distance between the data set's values and the mean. The third measure is the Coefficient of Variation (cv) which can be defined as the sigma divided by the mean, and it is a standardized measure of the dispersion of data distribution. More precisely, cv is used to compare the relative dispersion or homogeneity of different datasets. The set of data with the largest value of cv is more relatively dispersed with less homogeneity and vice versa. Through the three measures mentioned above, the reason for the combined use of each representation or formula of the image is obvious. Whereas the mean is higher in the uninfected cases, the sigma is less, and the cv is lower, and vice versa with the uninfected situation. The disease is distinguished in medical images based on color contrast, thus, the mean of each color is relatively smaller, while the sigma for each color is bigger due to the difference in the color of the disease within the image, and therefore the cv is large. Fig. 1 shows that three forms of skewness may generally be generated according to MCT as follows:

1) **Symmetric**: when the skewness is near zero and the mean is almost equal to the median.
2) **Positive skew:** when the right tail of a distribution's histogram is longer than the left tail, and the bulk of observations are focused on the left tail.
3) **Negative skew**: when the left tail of the distribution's histogram is longer than the right tail, and the bulk of the observations are focused on the right tail. Moreover, the median is higher than the mean.

In the case of a skew to the right, the arithmetic mean is greater than the median, so the value of the skew coefficient becomes positive (+), but in the case of the opposite and the skew to the left, the arithmetic mean is smaller than the median, and thus the value of the skew coefficient becomes negative (−). The simplest way to calculate the coefficient of skew is presented in Eq. (1):

$$\mathrm{Sk} = \frac{mean - median}{sigma} \tag{1}$$

The proposed approach is called Feature Extraction based on the Region of Mines (FE_mines) and it consists of 3 methods: Feature Extraction for All Color Images Based on Median (FE_AM), Feature Extraction for Color Images Based on Median (FE_CM), and Feature Extraction for Uncolored Images Based on Median (FE_UM).

2.1. The feature extraction method for images

The FE_AM is the first method in this paper that can deal with any species of the image, such as fundus, MRI, and X-ray. Although three channels represented by the 3-D matrix are obtained when reading an image on a computer, however, the values within the second and third channels decrease as the colors in the image decrease, thence, this method is more effective in the case of color images. The image is read into a matrix called Image Matrix (I_M) with size (n.n.3), then split into its basic components to obtain three extra channels or (matrix) namely: red, green, and blue matrixes of the same size (n.n) as (R_M), (G_M), and (B_M) respectively, meaning that 4 matrices were obtained (I_M) with size (n.n.3) and three channels (R_M), (G_M), and (B_M) with size (n.n). Then, each matrix is divided by the median into two regions, namely upper median, and lower median, where two additional virtual channels are obtained in each of the four channels. Each region represents a matrix upper matrix (um) and lower matrix (lm) in addition to the original matrix which created these two matrices (im); then, mean (μ), sigma (σ), and cv are calculated within the um, im, and lm matrixes. Thence, nine features are acquired in each of them. Fig. 2 shows the steps of obtaining the features for each channel.

Following the same steps for the rest of the matrices I_M, R_M, and G_M, as shown in Fig. 3, additional 27 features are extracted such that the final sum of features per image is (9+27) = 36 according to the FE_AM method.

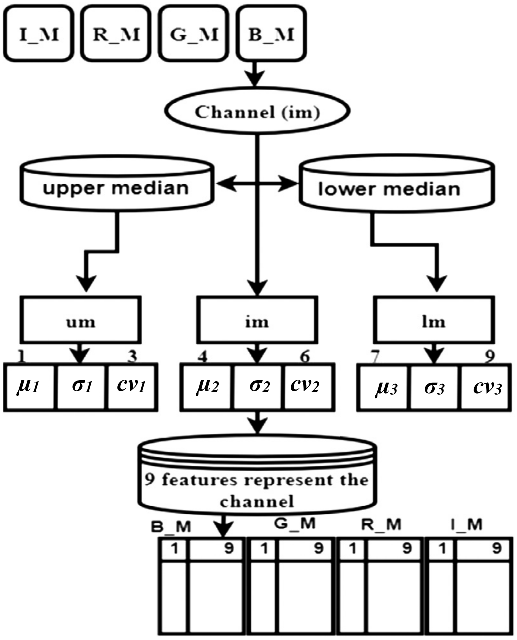
I_M
R_M
G_M
B_M
Channel (im)
upper median
lower median
um
im
lm
1
3
4
6
7
9
μ_1 σ_1 cv_1
μ_2 σ_2 cv_2
μ_3 σ_3 cv_3
9 features represent the channel
B_M
G_M
R_M
I_M
1 9
1 9
1 9
1 9

Fig. 2. Extracting the features from the channel.

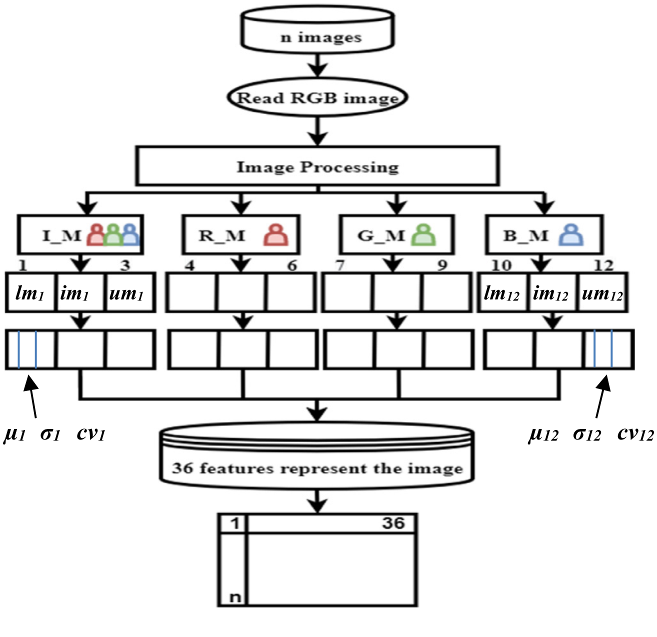
n images
Read RGB image
Image Processing
I_M
R_M
G_M
B_M
1
3
4
6
7
9
10
12
lm_1 im_1 um_1
lm_{12} im_{12} um_{12}
μ_1 σ_1 cv_1
μ_{12} σ_{12} cv_{12}
36 features represent the image
1
36
n

Fig. 3. The FE_AM method for extracting features of images.

2.2. The Features extraction method for colored images

The second method is FE_CM which uses Original, Grayscale, and Binary formulas based on Median. Fig. 4 illustrates the steps of the FE_CM method to extract features from images, where four sequential stages conclude the approach namely: 1) Image processing 2) Signal processing 3) Skewness 4) statistical measures. Since discrete-wave transform is used to obtain some formulas, it is necessary to briefly discuss the topic. The Wavelet transform is similar to the Fourier transform, where Fourier transform converts the signal into sines and cosines, or functions localized in Fourier space while the Wavelet transform employs functions localized in both real and Fourier space. In general, the Wavelet transform is represented by Eq. (2):

$$\mathrm{F(a,b)} = \int_{-\infty}^{\infty} f(x)^{*}_{(a,b)}(dx) \tag{2}$$

The image is read in the original form (Img), then converted into two additional formulas, namely grayscale (G), and binary scale (B), followed by the application of discrete-wave transform from signal processing on grayscale to give (cA1, cD1) and on the binary scale to give (cA3, cD3), on top of the original form to give (cA2, cD2). Thence, six different formats were obtained, three of them using image processing and the rest using signal processing, leading to the extraction of hidden attributes and enabling the model to identify the presence of diseases within the images. Al-

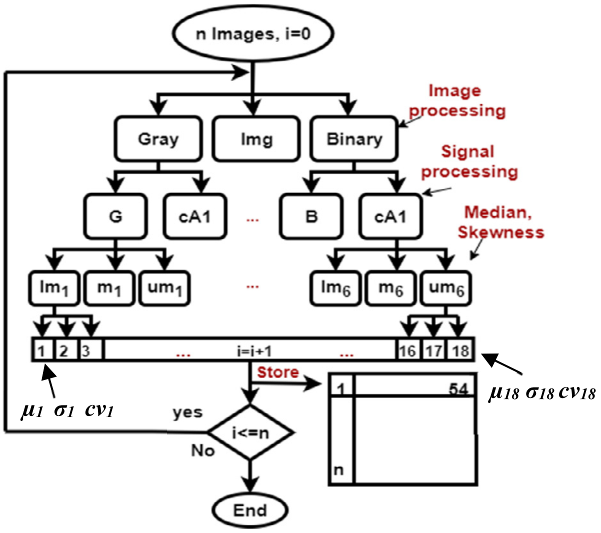

Fig. 4. The FE_CM method for medical color images.

though, two matrixes are obtained when converting any form of image using Signal processing techniques, however, only one of them is used (cA) because the (cD) matrix mostly consists of zeros, resulting in ting zeros for each mean, sigma, and cv; thence, the redundant features to be eliminated. Given that the matrix of the lower median region (lm5) that was calculated in the binary matrix where all values of mean, sigma, and cv (μ, σ, cv) were zeros, they were excluded with their resulting features; so, the number of features becomes 51 instead of 54 since the proposed method does not require any other feature techniques such as the selection or reduction features.

2.3. The Features extraction method for uncolored images

The third method is FE_UM which uses less amount of data to be represented within a computer; therefore, a few matrices will contain zeros when applying transformation of signal processing. Thus, the previous method needs to be reconsidered in terms of the matrices used to extract the features. The methods introduced in this paper involve all feature techniques simultaneously, namely (reduction, extraction, and selection). To adapt the previous method of dealing with uncolored medical images, the matrices eliminate cA2 and cA3 to avoid extracting redundant features which result from calculating (μ, σ, cv). Several features become $51-18 = 33$ features.

The idea can be summarized in general in one form through which the proposed method can be simplified. Fig. 5 represents the three main stages of FE_mines which consist of Image representation, Data distribution, and Feature extraction.

2.4. Mathematical formula

Suppose that I_M is the matrix of a medical image of size (n, n, 3) where n expresses the resolution of the image, and when the image is split into the primary components, the result will be three rank matrices R_M, G_M, B_M with size (n × n) as three colors red, green, and blue respectively, p is the number of elements within R_M, G_M, and B_M matrix while 3*p number of elements within I_M matrix.

$$I\_M \in \mathbb{R}^{\boldsymbol{n}.\boldsymbol{n}.3} \quad \text{and} \quad R\_M, G\_M, B\_M \in \mathbb{R}^{\boldsymbol{n}.\boldsymbol{n}}$$

Each matrix will divide according to the Median; thus, two extra matrices derived from each one are obtained as

$$I\_M_{up}, I\_M_{dn}, R\_M_{up}, R\_M_{dn}, G\_M_{up}, G\_M_{dn}, B\_M_{up}, B\_M_{dn}$$

Three measures are calculated for the original matrix I_M as:

$$\mu_{(I\_M)} = (\frac{1}{3p})\sum_{p=1}^{p} I\_M(3p)$$

$$\sigma_{(I\_M)} = (\frac{1}{3p})\sum_{p=1}^{p}(I\_M(3p) - \mu_{(I\_M)})^2$$

$$CV_{(I\_M)} = \frac{\sigma_{(I\_M)}}{\mu_{(I\_M)}}$$

The following equations are used to find the tree features of I_M_{up}

$$\mu_{(I\_M_{up})} = (\frac{1}{p})\sum_{p=1}^{p} I\_M_{up}\,(p)$$

$$\sigma_{(I\_M_{up})} = (\frac{1}{p})\sum_{p=1}^{p}(I\_M_{up}(p) - \mu_{(I\_M_{up})})^2$$

$$CV_{(I\_M_{up})} = \frac{\sigma_{(I\_M_{up})}}{\mu_{(I\_M_{up})}}$$

Then all other features are calculated, as a result 36 features are extracted for I_M, R_M, G_M, B_M and their subset matrixes (mines) which all of them represent the image. In the same way all features are extracted for FE_CM and FE_UM methods with different matrixes (binary and gray), All features are calculated from themt, where the number of features is 51 in FE_CM method while is 33 in the FE_UM method.

3. Analysis of the methodology

Since the proposed approach is built on a simple idea via obtaining different formulas for each medical image and determining the areas of discriminative in each of them (mines of attributes) using the median, then extracting the effective features within the mines using three statistical measurements. Therefore, many equations and mathematical formulas are not feasible to make it understandable but little analysis will be better.

3.1. Analysis of FE_mines approach

This is a novel technique for representing the medical images and others that combine main three concepts of feature, namely selection, reduction, and extraction, which is characterized by the following:

1) The method depends on the hybridization of four different concepts which are image processing, signal processing, data distribution, and measures of central tendency and dispersion.
2) The normal or symmetrical distribution of the data images is a weak scenario for this approach since there is no data density at one of the edges where data can be concentrated.
3) It minimized the image data size down to 0.0007%; thus, the traditional machine learning algorithms can be used instead of the deep learning environment.
4) Since the statistical measurements (μ, σ, cv) work sequential depending on where each measure interprets the previous one, it means that the strength of feature increases and the correlation of features shows the hidden patterns and strengthens the weak features.
5) Although the proposed approach has not been implemented in other domains, it will be successful since medical image diagnosis is one of the difficult domains in the image classification problem.

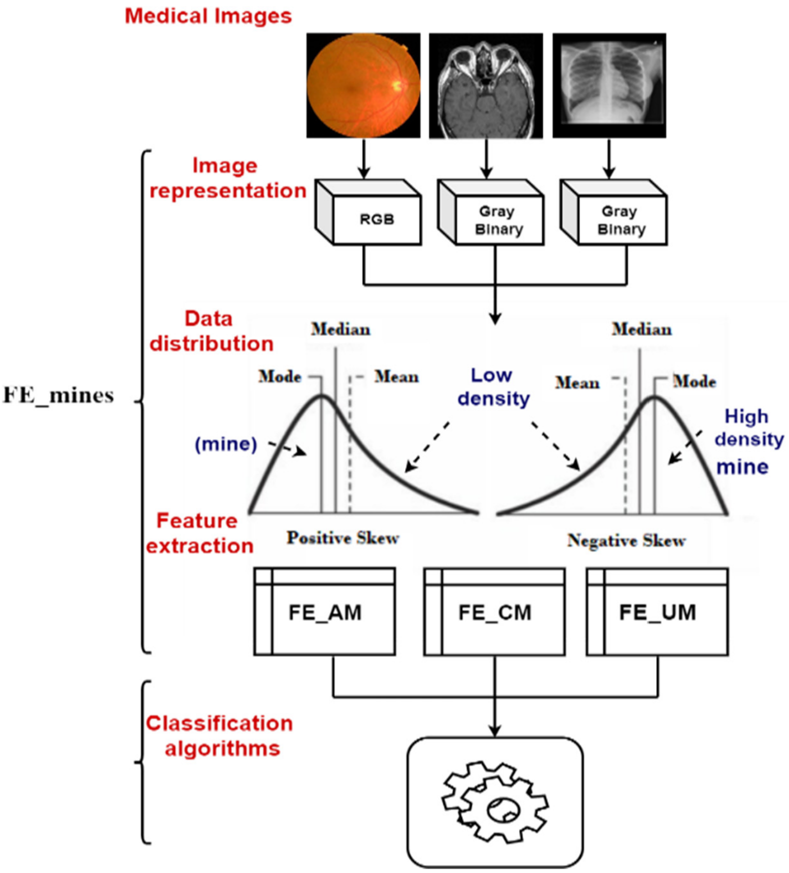

Fig. 5. The FE_mines approach for medical images.

6) FE_mines approach is not label-dependent; therefore, all its methods succeed in both supervised learning and unsupervised learning.
7) The variety of formulas that represent an image is additional resources that help extract hidden features, thus reducing the need to raise the number of real objects or even through a dataset augmentation technique.

3.2. Analysis of One-way-ANOVA

The mechanism of the three different methods can be delved into to strengthen the argument and make it understandable. Using an inclusive statistical concept which is One-way ANOVA or analysis of variance (ANOVA) is a statistical method for testing for differences in the means of three or more sets. The analysis of data distribution will be clear via the view of the graph; many statistical indicators can be recognized, such as mean, median, nature of data distribution, upper limit outlier, lower limit outlier, skewness, kurtosis ... etc. Fig. 6 shows an ANOVA graph of the RGB data of three medical images (Diabetic Retinopathy (Color Fundus photography); Brain Tumor (MRI); and COVID-19 chest (X-ray)). The amount of color is large in the first image, while it decreases in the second and decreases more in the third. Most of the data is focused on the bottom edge of Covid-19 ANOVA which indicates to strong Negative Skew, whilst is less in the brain ANOVA, and it tends to be symmetric in the case of Diabetic Retinopathy.

Fig. 7 illustrates the ANOVA graph of Diabetic Retinopathy, where the discriminative is tiny between the infected and uninfected images. Therefore, the three statistical measurements (μ, σ, cv) that belonged to different classes are very similar.

Fig. 8 is the ANOVA graph for both images mentioned above after separating the images into base components, which are R_M,

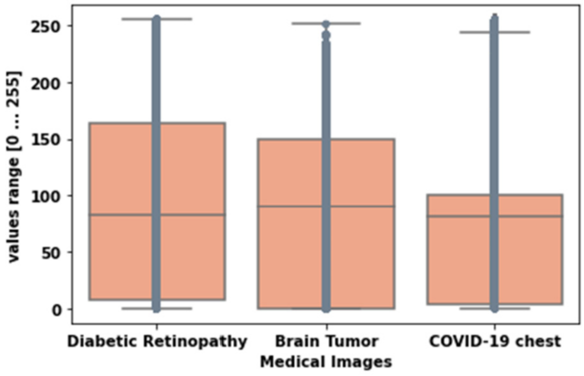

Fig. 6. ANOVA graph of three types of medical images.

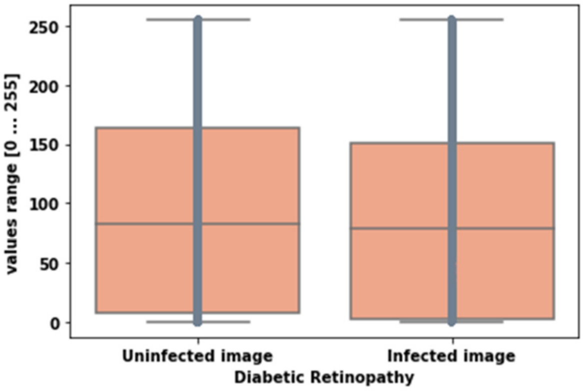

Fig. 7. ANOVA graph of two Diabetic Retinopathy images.

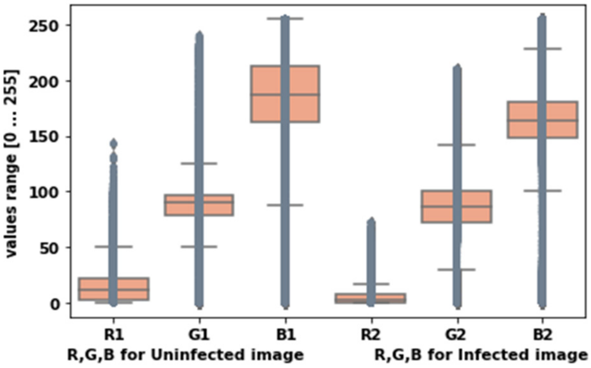

Fig. 8. ANOVA of RGB of Diabetic Retinopathy images.

Table 1
The description of the three datasets.

S.	Type of disease	No. of images	Medical image type	Image type
1	Diabetic retinopathy	5541/3663	Color Fundus	jpg
2	Brain Tumor	1427	MRI	png
3	COVID-19 chest	13808	X-ray	png

G_M, and B_M, where the distinction is large and clear, as the size of the body (orange color) in every two symmetrical candles is double, such as (R1, R2) and (G1, G2) and (B1, B2). Thence, calculating the three coefficients (μ, σ, cv) in each channel has a great indication of the presence of the disease or not. Moreover, in the uninfected image, the means in the unaffected images are larger than those in the infected images, which is reflected in the means of their classes.

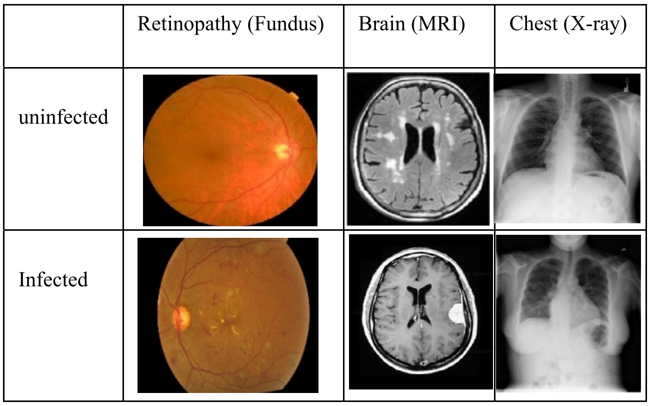

Fig. 9. Samples of the three medical images datasets.

Table 2
The models' accuracy results of the three experiments.

		Previous methods		FE_mines approach		
Exp.	DS size	RGB	ASPS	FE_AM	FE_CM	FE_UM
1.a	5000	**97.11%**	95.84%	**98.02%**	97.83%	–
1.b	3500	**93.86%**	90.42%	**95.63%**	94.10%	–
2	1427	89.51%	**94.67%**	93.00%	–	**96.85%**
3	13808	**80.59%**	–	**93.36%**	–	91.23%

4. Experimental analysis and results

The proposed approach was implemented using python version 5. The experiments were conducted using three healthcare datasets namely: Diabetic Retinopathy [26] (Color Fundus photography); Brain Tumor [27] (MRI); and COVID-19 chest [28] (X-ray) to demonstrate the advantages of FE_AM, FE_CM, and FE_UM methods. The Automated Sensory and Signal Processing Selection System (ASPS) [29] and RGB statistical method are used to compare the performance of the proposed approach under the same condition, Random Forest model hyperparameters tuning was done manually for data extracted by the proposed method and the two other methods. Table 1 represents the healthcare datasets used in this paper.

The FE_AM method is used to conduct three experiments as it is compatible with color and uncolored images, while the FE_CM is used for the first experiment (Diabetic Retinopathy) as it can deal with color images; the FE_UM is used to conduct the second and third experiments (Brain Tumor and COVID-19 chest) because the images are uncolored. Fig. 9 shows a few samples of each dataset, where Diabetic Retinopathy images are colored and images of the brain and chest are uncolored.

Table 2 shows the results obtained using the Random Forest algorithm, where the first and second rows represent the results of the first experiment. The first row's number of images was added using the augmentation technique, while the second row represents the dataset without any increase as they are. Images of the brain were represented in the third row, and the fourth row belongs to X-ray chest images. The result shows the improvement in the accuracy of the models using the three proposed methods. In the first experiment with increasing data size using augmentation, the proposed method achieved an improvement in accuracy by about **1**%, while it achieved **2**% in the case of the data without an increase for the same experiment. The second experiment achieved a 2% and **7**% accuracy improvement over the previous methods. Finally, a **13**% improvement in the model accuracy was achieved in the case of X-ray images using the FE_AM method.

X-ray images are an old technology in which the image is represented by a small amount of data and is not clear enough. The other methods face challenges in extracting the features, while FE_AM and FE_UM methods fulfilled higher accuracy, reaching 13%. This is due to its mechanism of deep search to extract hidden features using the concept of statistical skewness to locate mines of attributes. Through the results of the first experiment, which consists of two sub experiments 1(a) and 1(b), where 1(a) involved a larger number of images using augmentation technology while the second did not, the difference in accuracy between 1(a) and 1(b) was less by 2.39 using the FE_AM method, while the difference was higher by 5.42% and 3.25% for RGB and ASPS methods, respectively. Therefore, the FE_AM method can be deemed as an alternative to using an augmentation technique, especially if it is considered that unreal data is added to the dataset. To more result in analysis detail and evaluate the models efficiency. Table 3 shows the comparison utilizing three significant measures namely: sensitivity, specificity, and precision, which provides a concise caption for the confusion matrix. Confusion Matrix is a diagram that shows the differences between actual and predicted values. It gauges how well the machine learning classification model is doing. The best result obtained using the proposed approach was compared with the best result obtained using the previous methods, as shown in Table 2 in bold. The three indicators confirmed the superiority of the model using the proposed method over the traditional methods of extracting features, where the percentage of sensitivity, specificity, and precision were 2.61%, 0.58%, and 11.63% respectively.

Experiment (1.b) was selected to compare RGB and FE_AM methods using ROC curves, In terms of the expected probability, ROC indicates how well the model separates the specified classes and as an indicator of a test's discriminative power, and it is the area under the ROC curve (AUROC). Fig. 10 illustrates training and testing ROC curves, where the ROC curve shows that the features extracted from the medical image by the FE_AM method have no overlap for the different categories, while the area under the

Table 3
The confusion matrix accuracy of the three experiments.

	Previous method			FE_AM approach		
Exp.	sensitivity	specificity	precision	sensitivity	specificity	specificity
1.a	97.48	96.78	97.45	98.03	98.03	97.99
1.b	94.07	93.82	93.91	95.21	96.24	95.01
2	92.13	92.86	95.67	95.05	96.00	97.31
3	85.09	.90.00	51.91	90.91	90.00	95.16
Avg	92.19	94.49	84.74	94.80	95.07	96.37
%				**2.61%**	**0.58%**	**11.63%**

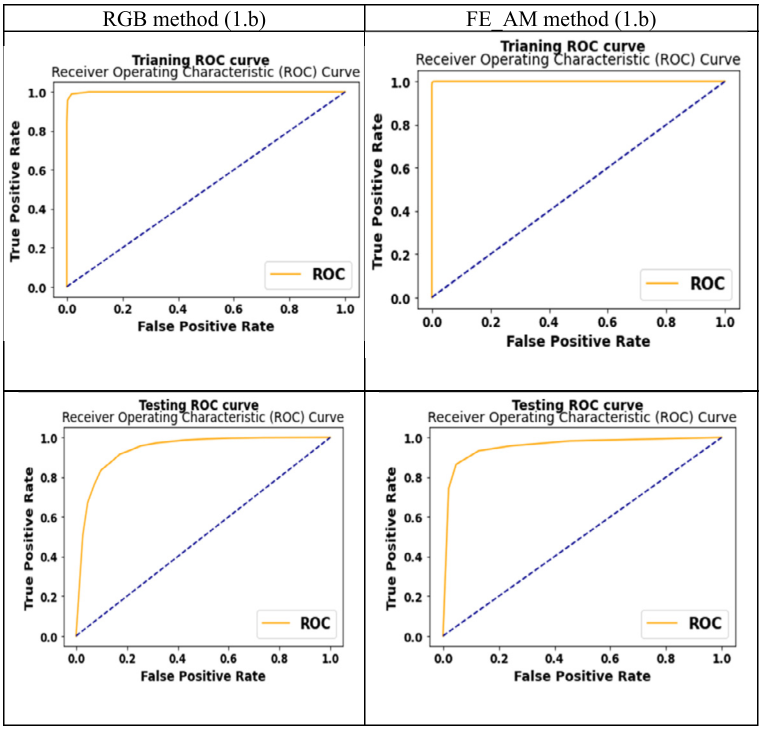

Fig. 10. Roc curve of the first experiment using RGB and FE_AM.

ROC curve for the RGB method is a little bit less, therefore, some overlap exists between both the classes, furthermore, this overlap increased in the case of testing ROC curves.

A validation Curve is a crucial diagnostic tool that demonstrates the sensitivity of a Machine Learning model's accuracy to changes in one or more model parameters. As long as the model is trained without ever seeing any testing data, it may be assessed perfectly. But if the model's hyperparameters need to be adjusted, a validation set must be provided, along with training data, so that correctness may be determined using testing data. Unseen data for many parameters are utilized to provide an accurate picture of the models' performance. Fig. 11 shows that the training accuracy reaches the optimum accuracy when the number of trees (hyperparameter) reaches about 25 in both models (using RGB and FE_AM). While in the case of cross-validation, it reached almost 95% and 93% using FE_AM, and RGB methods respectively.

It is possible to realize the effect of one or more factors in any study or research by changing the value of these factors with the fixation of the other factors simultaneously with analyzing the results each time. The strength of classification models can be increased either by factors related to the dataset such as feature extraction or feature selection or others that enable the model easily to identify patterns; or by factors related to selecting the appropriate algorithm and its hyperparameters used in building the model, therefore, This strategy was used to determine the effect of using the proposed approach instead of other to build the same model. As a result, three detailed results (confusion matrix, ROC curve, and validation accuracy) of the three experiments showed the superiority of the proposed method in building an effective model under the same conditions.

5. Conclusion and future work

A novel approach namely, FE_minis was introduced which includes three versions; FE_AM, FE_CM, and FE_UM. This approach relies on formed extra formulas of images and the nature of data distribution skew to determine the region of the mines, which helps to extract three statistical coefficients that represent the hidden features. The results of the three experiments showed that FE_mines was superior to the other two approaches (RGB and ASPS) in obtaining active features led to building models with higher accuracy, using confusion matrix, ROC curve, and validation accuracy for results analysis. thence, the FE_mines can be considered as an alternative to increase the size of the dataset to improve the efficiency of medical image models. The challenge of the FE_minis is that involving three versions may lead to confusing the user to choose an appropriate version compatible with its dataset. Furthermore, Two limitations can be summarized, the first is the proposed algorithm is less effective in extracting hidden features in the case of colored images, while the effectiveness increases with less color within the image, where the results showed an improvement in accuracy by 13% and 1% in the case of non-colored

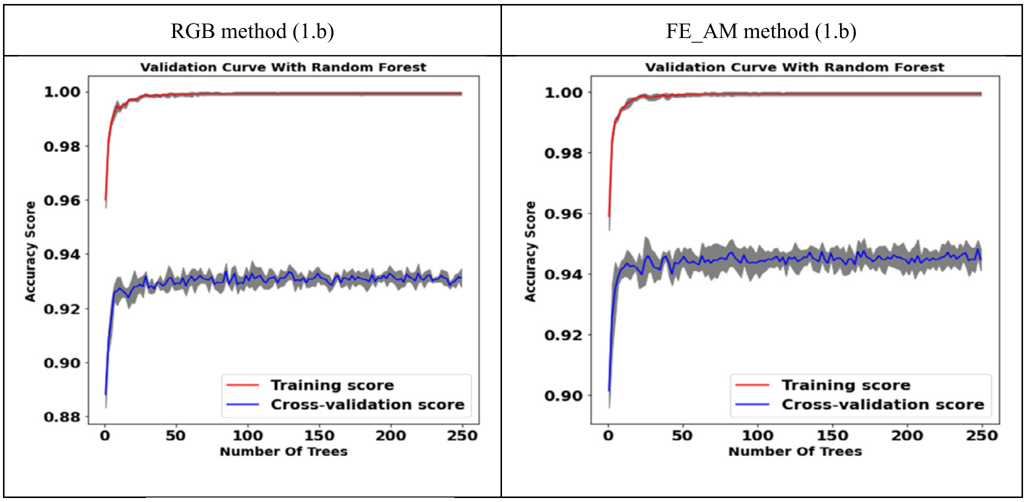

Fig. 11. Training and validation accuracy curve of the first experiment using RGB and FE_AM.

and colored images, respectively. And the second one is the absence of obvious disease in the images may lead to obtaining a normal distribution instead of a positive or negative skewed distribution, thence, the proposed method becomes ineffective in such rare cases. Understanding the distribution of data within medical images helps to comprehend the nature of data distribution for each class that these images belong to, as a result, obtaining some powerful attributes that express images allows adapting the dataset of images to the statistical concepts and data distribution, in addition to the algorithm that is based on similarities in ML, such as GCC [30], CDC [31], K-means, and KNN. A new methodology may be needed to calculate the similarity between images, especially in unsupervised learning where the multi-representation of each image reaches up to 18 sets that are represented by 6 formulas as illustrated in detail in the FE_mines approach. This leads to flexibility in finding the Euclidian distance objects that fulfill the similarity and help to perform optimal clusters. On the other hand, there are wide open horizons that the proposed approach could address, including diagnosing other diseases within the healthcare domain, furthermore, to several other domains, such as engineering, environments, security, botany, ... etc.

Human and animal rights

The authors declare that the work described has not involved experimentation on humans or animals.

Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

The authors attest that they fulfill the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

Declaration of competing interest

The authors declare that they have no known competing financial or personal relationships that could be viewed as influencing the work reported in this paper.

Acknowledgements

I would like to express my gratitude to my wife and children for providing the right environment to achieve this work.

References

[1] F. vom Scheidt, H. Medinová, N. Ludwig, B. Richter, P. Staudt, C. Weinhardt, Data analytics in the electricity sector – a quantitative and qualitative literature review, Energy AI 1 (2020) 100009, https://doi.org/10.1016/j.egyai.2020.100009.

[2] J.C. Pena, G. Nápoles, Y. Salgueiro, Normalization method for quantitative and qualitative attributes in multiple attribute decision-making problems, Expert Syst. Appl. 198 (March) (2022) 116821, https://doi.org/10.1016/j.eswa.2022.116821.

[3] A. Sengur, An expert system based on linear discriminant analysis and adaptive neuro-fuzzy inference system to diagnosis heart valve diseases, Expert Syst. Appl. 35 (1–2) (2008) 214–222, https://doi.org/10.1016/j.eswa.2007.06.012.

[4] S.R. Das, D. Mishra, M. Rout, Stock market prediction using Firefly algorithm with evolutionary framework optimized feature reduction for OSELM method, Expert Syst. Appl. X 4 (2019) 100016, https://doi.org/10.1016/j.eswax.2019.100016.

[5] M. Li, H. Wang, L. Yang, Y. Liang, Z. Shang, H. Wan, Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction, Expert Syst. Appl. 150 (2020) 113277, https://doi.org/10.1016/j.eswa.2020.113277.

[6] W.K. Mutlag, S.K. Ali, Z.M. Aydam, B.H. Taher, Feature extraction methods: a review, J. Phys. Conf. Ser. 1591 (1) (2020), https://doi.org/10.1088/1742-6596/1591/1/012028.

[7] B. Ghojogh, et al., Feature selection and feature extraction in pattern analysis: a literature review, 2019 [Online]. Available: http://arxiv.org/abs/1905.02845.

[8] N. Ani Brown Mary, D. Dharma, Coral reef image classification employing improved LDP for feature extraction, J. Vis. Commun. Image Represent. 49 (December 2016) (2017) 225–242, https://doi.org/10.1016/j.jvcir.2017.09.008.

[9] M. Imani, H. Ghassemian, An overview on spectral and spatial information fusion for hyperspectral image classification: current trends and challenges, Inf. Fusion 59 (October 2019) (2020) 59–83, https://doi.org/10.1016/j.inffus.2020.01.007.

[10] M.Y. Demirci, N. Beşli, A. Gümüşçü, Efficient deep feature extraction and classification for identifying defective photovoltaic module cells in electroluminescence images, Expert Syst. Appl. 175 (August 2020) (2021) 2021, https://doi.org/10.1016/j.eswa.2021.114810.

[11] R.K. Tripathi, A.S. Jalal, Novel local feature extraction for age invariant face recognition, Expert Syst. Appl. 175 (December 2019) (2021) 114786, https://doi.org/10.1016/j.eswa.2021.114786.

[12] B. Bataineh, S.N.H.S. Abdullah, K. Omar, A novel statistical feature extraction method for textual images: optical font recognition, Expert Syst. Appl. 39 (5) (2012) 5470–5477, https://doi.org/10.1016/j.eswa.2011.11.078.

[13] S. Sachar, A. Kumar, Survey of feature extraction and classification techniques to identify plant through leaves, Expert Syst. Appl. 167 (2021) 114181, https://doi.org/10.1016/j.eswa.2020.114181.

[14] Ü. Atila, M. Uçar, K. Akyol, E. Uçar, Plant leaf disease classification using EfficientNet deep learning model, Ecol. Inform. 61 (September 2020) (2021) 101182, https://doi.org/10.1016/j.ecoinf.2020.101182.

[15] J. Yang, et al., Artificial intelligence in ophthalmopathy and ultra-wide field image: a survey, Expert Syst. Appl. 182 (December 2020) (2021) 115068, https://doi.org/10.1016/j.eswa.2021.115068.

[16] Z. Lai, H. Deng, Medical image classification based on deep features extracted by deep model and statistic feature fusion with multilayer perceptron, Comput. Intell. Neurosci. 2018 (2018), https://doi.org/10.1155/2018/2061516.

[17] T. Song, X. Yu, S. Yu, Z. Ren, Y. Qu, Feature extraction processing method of medical image fusion based on neural network algorithm, Complexity 2021 (2021), https://doi.org/10.1155/2021/7523513.

[18] T. Mahmud, M.A. Rahman, S.A. Fattah, CovXNet: a multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization, Comput. Biol. Med. 122 (June) (2020) 103869, https://doi.org/10.1016/j.compbiomed.2020.103869.

[19] M.M.A. Monshi, J. Poon, V. Chung, Deep learning in generating radiology reports: a survey, Artif. Intell. Med. 106 (April 2019) (2020) 101878, https://doi.org/10.1016/j.artmed.2020.101878.

[20] X. Chen, L. Li, A. Sharma, G. Dhiman, S. Vimal, The application of convolutional neural network model in diagnosis and nursing of MR imaging in Alzheimer's disease, Interdiscip. Sci. – Comput. Life Sci. 14 (1) (2022) 34–44, https://doi.org/10.1007/s12539-021-00450-7.

[21] M. Javaid, A. Haleem, A. Vaish, R. Vaishya, K.P. Iyengar, Robotics applications in covid-19: a review, J. Ind. Integr. Manag. 5 (4) (2020) 441–451, https://doi.org/10.1142/S2424862220300033.

[22] S. Lalmuanawma, J. Hussain, L. Chhakchhuak, Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review, Chaos Solitons Fractals 139 (2020) 110059, https://doi.org/10.1016/j.chaos.2020.110059.

[23] S. Chopra, G. Dhiman, A. Sharma, M. Shabaz, P. Shukla, M. Arora, Taxonomy of adaptive neuro-fuzzy inference system in modern engineering sciences, Comput. Intell. Neurosci. 2021 (2021), https://doi.org/10.1155/2021/6455592.

[24] S. Aggarwal, N. Chugh, Review of machine learning techniques for EEG based brain computer interface, Arch. Comput. Methods Eng. (2022), https://doi.org/10.1007/s11831-021-09703-6.

[25] I. Hussain, et al., Quantitative evaluation of EEG-biomarkers for prediction of sleep stages, 2022 [Online]. Available: https://www.mdpi.com/1424-8220/22/8/3079/htm.

[26] Sovit Ranjan Rath, Diabetic retinopathy 224x224 Gaussian filtered, kaggle, 2020, https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-224x224-gaussian-filtered.

[27] Sartaj, Brain tumor classification (MRI), kaggle, 2020, https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri.

[28] T. Rahman, Covid-19 radiography database | kaggle, Kaggle, 2021, https://www.kaggle.com/tawsifurrahman/covid19-radiography-database/activity%0A https://www.kaggle.com/tawsifurrahman/covid19-radiography-database.

[29] A. Al-Habaibeh, N. Gindy, New approach for systematic design of condition monitoring systems for milling processes, J. Mater. Process. Technol. 107 (1–3) (2000) 243–251, https://doi.org/10.1016/S0924-0136(00)00718-4.

[30] F.H. Kuwil, Ü. Atila, R. Abu-Issa, F. Murtagh, A novel data clustering algorithm based on gravity center methodology, Expert Syst. Appl. 156 (2020), https://doi.org/10.1016/j.eswa.2020.113435.

[31] F.H. Kuwil, F. Shaar, A.E. Topcu, F. Murtagh, A new data clustering algorithm based on critical distance methodology, Expert Syst. Appl. 129 (2019), https://doi.org/10.1016/j.eswa.2019.03.051.