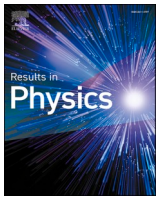
Results in Physics 31 (2021) 105045

Contents lists available at ScienceDirect

Results in Physics

journal homepage: www.elsevier.com/locate/rinp



COVID-19 anomaly detection and classification method based on supervised machine learning of chest X-ray images

ARTICLE INFO

Keywords
COVID-19 diagnosis
X-ray image
Local binary pattern
Haralick
Machine learning
K-nearest neighbor
Support vector machine

ABSTRACT

The term COVID-19 is an abbreviation of Coronavirus 2019, which is considered a global pandemic that threatens the lives of millions of people. Early detection of the disease offers ample opportunity of recovery and prevention of spreading. This paper proposes a method for classification and early detection of COVID-19 through image processing using X-ray images. A set of procedures are applied, including preprocessing (image noise removal, image thresholding, and morphological operation), Region of Interest (ROI) detection and segmentation, feature extraction, (Local binary pattern (LBP), Histogram of Gradient (HOG), and Haralick texture features) and classification (K-Nearest Neighbor (KNN) and Support Vector Machine (SVM)). The combinations of the feature extraction operators and classifiers results in six models, namely LBP-KNN, HOG-KNN, Haralick-KNN, LBP-SVM, HOG-SVM, and Haralick-SVM. The six models are tested based on test samples of 5,000 images with the percentage of training of 5-folds cross-validation. The evaluation results show high diagnosis accuracy from 89.2% up to 98.66%. The LBP-KNN model outperforms the other models in which it achieves an average accuracy of 98.66%, a sensitivity of 97.76%, specificity of 100%, and precision of 100%. The proposed method for early detection and classification of COVID-19 through image processing using X-ray images is proven to be usable in which it provides an end-to-end structure without the need for manual feature extraction and manual selection methods.

Introduction

The new Coronavirus 2019 (COVID-19) pandemic first appeared in Wuhan, China, in 2019, and started to spread rapidly, posing a critical public health problem to the entire world [1]. COVID-19 results in mild symptoms in about 82% of the cases, and other conditions are severe or critical [2,3]. The total number of COVID-19 confirmed cases throughout the world is 229,373,963, including 4,705,111 deaths reported by the World Health Organization (WHO) on 23 September 2021 [4]. Fig. 1 shows the distribution of COVID-19 diagnosed cases worldwide.

The COVID-19 pandemic virus is severe respiratory syndrome coronavirus 2, also called SARS-CoV-2. A high number of infected patients has survived the virus, while a smaller percentage has serious or critical conditions [5,6]. The increase in the number of people with the COVID-19 virus leads to an increased need for intensive care. This extension creates a workload on the healthcare system leading to the collapse of the health systems even in the best-developed countries. When intensive care units (ICUs) are full of patients, the health status of COVID-19 patients deteriorates, and the rate of death increases. Some researchers utilize medical images like X-rays or Computed Tomography (CT-scans) for the search of properties symptoms of the novel coronavirus [2,7].

The COVID-19 pandemic has lead to huge financial losses worldwide, posing a massive impact on world GDP growth [1]. Global recession has been very severe since the end of World War II resulting in the contraction of the global eonomy by 3.5% in 2020 based on the April 2021 World Economic Outlook Report published by the IMF, which states a 7% loss relative to the 3.4% growth forecast of October 2019. While virtually every country reported by the IMF posted negative growth in 2020, the downturn was more pronounced in the poorest parts of the world [8].

Researchers of some recent studies employ chest radiography in epidemiological regions for testing COVID-19 [1,3]. They found that the examination of radiographic images could be an alternative to the PCR scheme as it shows a higher sensitivity in some cases [9]. Xu et al. [10] introduced a new method based on a deep learning system to screen coronavirus COVID-19 pneumonia. The proposed method aims to build up an early examination model to recognize COVID-19 pneumonia from Influenza-A viral pneumonia and health conditions with lung section images based on deep learning methods [8]. The proposed algorithm is designed based on candidate infection areas divided using a three-dimensional deep learning technique from a set of pulmonary CT images [6]. The results of the experiments benchmark dataset shows that the inclusive accuracy is 86.7 % from the perspective of CT of the whole cases.

In the work of Sethy and Behera [9], they proposed an algorithm for the detection of COVID-19 based on deep features. Deep features are extracted from a pre-trained CNN model and fed to an SVM classifier in individual form. The proposed classification scheme for the detection of COVID-19 obtained an accuracy of 95.38%.

In the prior work of Ozturk et al. [11], they presented a new model based on deep learning techniques to detect and classify COVID-19 conditions from X-ray images. The proposed model is completely automated based on an end-to-end structure. In addition, the proposed method is able to perform binary and multi-class classifications with accuracy values of 98.08% and 87.02%, respectively. Subsequently,

https://doi.org/10.1016/j.rinp.2021.105045
Received 28 December 2020; Received in revised form 19 November 2021; Accepted 19 November 2021
Available online 22 November 2021

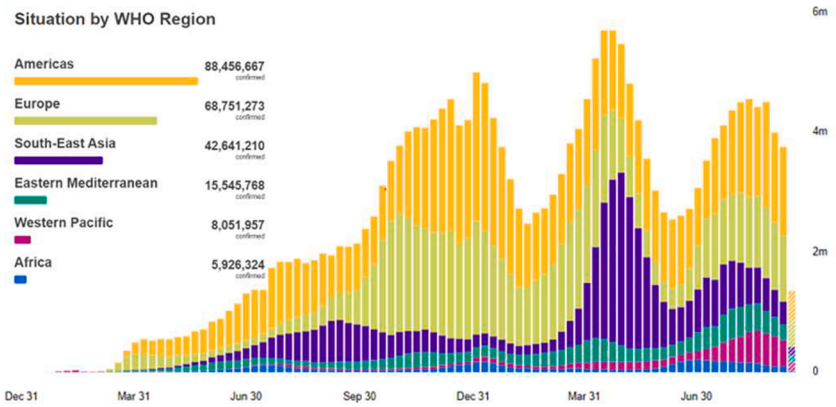


Fig. 1. Distribution of COVID-19 cases worldwide until Sep. 2021 [4].

Narin et al. [12] proposed utilizing three types of CNN-based models: Inception ResNetV2, InceptionV3, and ResNet50 for COVID-19 detection from chest X-ray images. Performance results of the proposed models illustrate that the pre-trained patterns of the ResNet50 obtained the highest accuracy of 98%. The major limitation is applying the proposed methods/models on a few numbers of COVID-19 X-ray images, which do not satisfy robust results.

This paper proposes a method for classification and early detection of COVID-19 through image processing using X-ray images. A set of procedures are applied in constructing the COVID-19 detection model, including preprocessing (image noise removal, image thresholding, and morphological operation), Region of Interest (ROI) detection, feature extraction using multiple methods such as Local binary pattern (LBP), Histogram of Gradient (HOG), and Haralick texture features. In the classification stage, the K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) are used with the percentage of training of 5-folds cross-validation for the region of interest. The contributions of our study are summarized as follows:

- We combine the feature extraction operators' and classifiers' outcome in six models, namely LBP-KNN, HOG-KNN, Haralick-KNN, LBP-SVM, HOG-SVM, and Haralick-SVM on 5,000 X-ray images. The combined six models are shown to output very high outcomes in a large dataset of 5,000 X-ray images. The results further show that chest X-ray images are one of the best means for the detection of COVID-19.
- We show that the LBP-KNN model is an effective model among other models. The LBP-KNN model outperforms the other models. It achieves an average accuracy of 98.66%, a sensitivity of 97.76%, a specificity of 100%, a precision of 100%, an error rate of 1.34%, and zero false positive.

The paper is organized as follows: Section 2 includes Methods and Materials of COVID-19 dataset, features extraction operators' and classifiers' outcome. Implementation and Results of COVID-19 detection and Classification are provided in Section 3. Finally, in Section 4, the conclusion and the future works are summarized.

Methods and Materials

The COVID-19 X-Ray dataset

This work utilizes a chest x-ray of 5,000 normal and pneumonia COVID-19 images that are obtained from the open-source GitHub warehouse shared by Cohen et al. [13], namely "Chest X-Ray Images (Pneumonia)". This warehouse provides chest X-ray/CT images of primary patients with COVID-19 along with other diseases. Samples of the selected chest X-ray images of normal and pneumonia COVID-19 sets are shown in Fig. 2 [5,13].

OTSU'S thresholding

Otsu's threshold method aims to convert a grayscale image to a binary image. This method employs various techniques of image processing to implement histogram-based image thresholding or to transform an image from grayscale to binary [11]. The Otsu thresholding method supposes that the image consists of the bi-modal histogram (foreground and background and the related optimal threshold).

Morphology operations

Mathematical morphology (MM) aims to extract components from the image that are useful in the depiction of region, shape and description like skeletons, and convex hull, boundaries. In addition, the morphological techniques are considered for pre-or post-processing, for example, morphological filtering by reconstruction, thinning, and pruning transform [9]. Generality morphological operations concentrate on binary images. Morphological operations are logical transformations dependent on a comparison between pixel neighborhoods with a predefined pattern.

Morphological Dilation: The dilation operation employed a structuring element also called "kernel" to check and extend the shapes [12]. When applied the structuring element S on image A, the result is a new image I,

$$I = A \oplus S = \bigcup_{s \in S} A_s \tag{1}$$

Opening: The image opening operation combines erosion and dilation operation using intersection and complementation [9]. The conditioned dilation (opening) begins by producing matrix X_0 of 0 s which its size equals A.

$$X_{i=}(X_{i-1} \oplus S) \cap A^C \tag{2}$$

wherein the final step:

$$X_i = X_{i+1} \tag{3}$$

where X(i) contains all the filled holes.

Morphological Closing: The morphological closing comes after dilation operation using the same structuring element [9]. The closing operation is,

$$A.S = (A \oplus S)\Theta S \tag{4}$$

Morphological Erosion: Morphological erosion operation aims to shrink the image. The output of erosion operation is an image I,

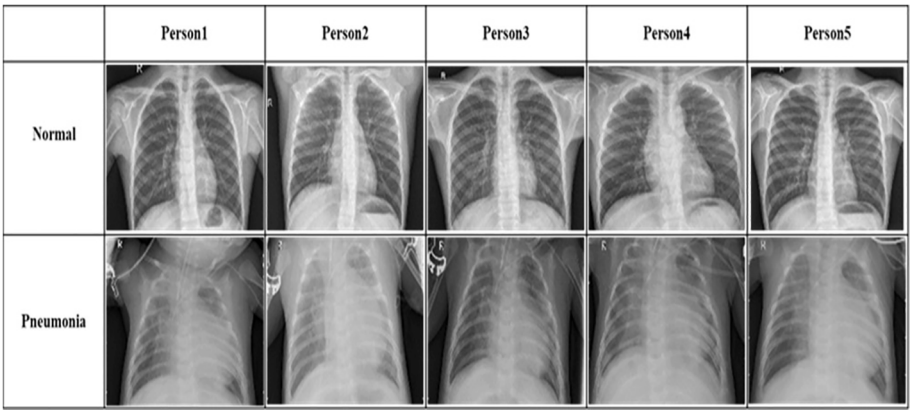


Fig. 2. Sample of the Chest X-Ray Images (Normal and Pneumonia) [5,13].

$$I = A\Theta S = \cap s \in S\ A - S \tag{5}$$

Midpoint ellipse drawing algorithm

The ellipse structure allows drawing using a circle scaling with a shorter radius in the direction of a longer radius. Several methods can be used to have midpoint ellipse as a drawing algorithm. The ellipse algorithm starts drawing at the origin and then moves straight towards the center point [14,15]. Fig. 3(a) illustrates the ellipse of a 4-way symmetry. It is similar to the scheme used to show a raster circle. The ellipse quadrant is split into two regions. Fig. 3(b) shows the section of the first quadrant that depends on the slope of an ellipse with Rx < Ry. As the ellipse is drawn from 90 to 0 degrees, x moves in the positive direction, and y moves in the negative direction, and the ellipse passes through two regions.

While the ellipse drawing algorithm preprocesses the first quadrant, then the algorithm moves towards x-direction (the magnitude of the curve slope < 1 for the first region) and towards the y-direction (the magnitude of the curve slope > 1 for the second region). Similar to the circle function, the ellipse function,

$$f_{ellipse}(x,y) = \left(r^2yx^2 + r^2xy^2 - r^2xr^2y\right) \tag{6}$$

Feature extraction

Local binary pattern

Local Binary Pattern (LBP) is one of the well-known image feature extraction operators adopted in many real-world applications [16]. The LBP is a simple, yet effective texture extraction operator. The LBP has a low computational complexity that enables it to work in complicated and real-time image processing applications. It is a unified approach to traditional structural and statistical models. It specifies the vicinity of each pixel of an image then labels these pixels with binary numbers. The LBP can be articulated in the decimal form given a pixel at (x_c, y_c) by Eq. (7):

$$LBP_{P,R}(X_c, Y_c) = \sum_{p=0}^{P-1} s(i_p - i_c)2^p \tag{7}$$

where i_c and i_p are respectively gray-level values of the central pixel and P surrounding pixels in the circle neighborhood with a radius R. The function s(x) is defined in Eq. (7) as follow:

$$s(x) = \begin{cases} 1 if x \geq 0 \\ 0 if x < 0 \end{cases} \tag{8}$$

HOG algorithm

With the aim to extract features, the HOG algorithm includes two main stages [15]. The first stage is histogram extraction of the oriented gradient. The gradient of the direction and magnitude are extracted from each pixel in the input image. These are employed to produce an angular histogram of gradients applied as an image texture feature vector. The vertical and horizontal components of the image I (i, j) are derivatives at pixel (i, j). They are respectively computed as below:

$$Gi(i,j) = I(i+1,j) - I(i-1,j) \tag{9}$$

where

$$Gj(i,j) = I(i,j+1) - I(i,j-1) \tag{10}$$

and,

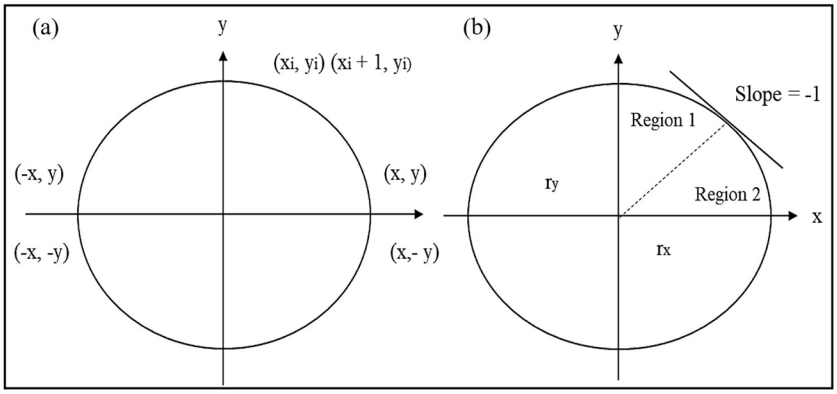


Fig. 3. Midpoint Ellipse Method [14].

$$G(i,j) = \sqrt{Gi(i,j)^2 + Gj(i,j)^2} \tag{11}$$

and,

$$\alpha_0(i,j) = \tan - 1\left[\frac{G_j(i,j)}{G_i(i,j)}\right], \alpha_0 \in \left[\frac{-\pi}{2}, \frac{\pi}{2}\right] \tag{12}$$

where G_i (i, j), G_j (i, j) are the derivative along a horizontal and vertical direction at pixel (i, j), respectively.

The second stage represents the construction of the HOG descriptor which is constructed based on the gradient of the image. Firstly, the whole image is split into blocks with size 8*8. The gradient direction range [-π/2, π/2] is calculated uniformly into nine intervals of direction (bins). To create a strong vector to brightness changes, the HOG feature results are normalized by segmenting each bin with the total of the histogram [15].

Haralick texture features (Second Order)

Gray Level Co-occurrence Matrix (GLCM) is a representation of interdependence levels and spatial distribution within a local area [17]. The Haralick operator calculates Harlic features according to the statistical distribution of the GLCM in which the peer of pixels is considered as second-order. It built relations between positions of pixels of an image [15,18]. The quantization process is applied to the image before calculating the co-occurrence matrix. Contrast is used to show the variation of the gray level of the neighbor pixels to the reference pixel as in Eq. (13):

$$contrast = \sum_i \sum_j (i-j)^2 p_d(i,j) \tag{13}$$

The homogeneity shows the relationships between the distribution of the elements and diagonal in the GLCM.

$$Homogeneity = \sum_i \sum_j \frac{1}{1+(i-j)^2} p_d(i,j) \tag{14}$$

The entropy shows the randomness of the image disorder which is formulated as in (15):

$$entropy = -\sum_i \sum_j p_d(i,j) 1np_d(i,j) \tag{15}$$

Classification methods

Different classification algorithms include naive Bayes, KNN, neural network, decision tree, SVM, etc. They are used to predict class labels of anonymous data. The KNN and SVM are selected in this work to construct the classification model.

K-Nearest Neighbor (KNN): The KNN algorithm utilizes the Euclidean distance standard to calculate the value of the variance between the training instance and the test instance [19]. The "K" indicates the number of closest neighbors that help predict the test pattern class [20]. The standard Euclidean distance d (x, y) is determined to follow as:

$$d(x_i, y_j) = \sqrt{(a_r(x_i) - a_r(x_j))^2} \tag{16}$$

Additionally, KNN computes the most popular category from the nearest neighbor K to estimate the test instance class for the test set. It is determined in Eq. (17):

$$c(x) = argmax_{c \in C} \sum_{i=1tok} \delta(c, c(y_i)) \tag{17}$$

The parameters y_1, y_2, y_3, …, y_k represents the k nearest neighbors of a specific instance of the test data set, k is the number of the neighbors, C represents the finite set of class labels, and δ (c, c(y_i)) = 1 if c = c(y_i) and δ(c, c(y_i)) = 0 otherwise [21].

Support Vector Machine (SVM): It is a type of supervised machine learning method that depends on the problem of maximum classification hyperplane interval linear separable. The Kernel function enables linear points to be less distant, relying on the region of the high dimensional feature. The selection of the model and the kernel function parameters directly affect the SVM learning results [22,23]. The parameter σ^2 determines the kernel function generalization and influences the kernel function generalization [9]. Gaussian Kernel (GK) is a sign function of the kernel in Kernel schemes. The feature space has an infinite dimension in which data that cannot be classified in a low linearly dimension can be classified in a higher dimension which the GK specifies as in (18):

$$k(x,y) = \exp(-\|x-y\|^2/2\sigma^2) \tag{18}$$

Implementation and results

The proposed method relies on image processing by performing a set of procedures that would give a preliminary diagnosis of COVID-19 patients through X-ray images [11,12,24,25]. The features operators of LBP, HOG, and Haralick and classifiers of SVM and K-NN made a combination of six models LBP-KNN, HOG-KNN, Haralick-KNN, LBP-SVM, HOG-SVM, and Haralick-SVM. Fig. 4 shows the overall proposed anomaly detection and classification model of COVID-19 based on chest X-ray images. The implementation of the proposed method is represented by applying several steps to the total image to specify ROI and then extracting features (multiple features based on chest X-ray images). Subsequently, building two classification models for detecting the abnormal case of COVID-19. The classification models consist of training and testing stages, and the model could be used for handling new cases. The key steps of the proposed COVID-19 diagnosis method are shown in Fig. 4.

The preprocessing of the images depends mainly on several steps for finding the region of interest (ROI). The first step is converting a color image into a gray image and applying the median filter on all dataset images, which removes the noise present in the image. Then, the gray images are converted into binary images based on the efficient common method called Otsu's thresholding which depends on the separation of the foreground from the background by reducing the intensity of the variance concerning the intra-class and increasing the intensity of the variance with the inter-class. The last preprocessing step is a morphological operation called opening operation, represented by two steps erosion followed by dilation. It performs the previous step by improving the binary image and removing a small region, which is considered unimportant areas or noise on the resulting image, and keeping the useful areas for processing in the next steps.

The method draws an ellipse to crop the ROI that represents the Midpoint Ellipse. The process is done by tracing the two points of the line in the lower area of the image and representing the right point X1, Y1, and the left point X2, Y2. Through these points, the middle point that represents X, Y is the center of the ellipse and is dependent on finding the distance between the two points X1, Y1 and X, Y or X2, Y2 and X, Y which are found as rx, the first radius. Then, ry is calculated from 0 to the height of the image and ry is chosen through the whitest percentage of blackness, which is calculated by taking the pixel value of points as Eq. (16, 17). This process ensures that the shape contains the lungs for testing, as shown in Fig. 5. The cropped region represents the lung area, and it is used later for feature extraction.

In the next step, the back-of-word process is used to standardize the length of the vectors generated for the classification process. From the previous step, all the 5,000 images are processed by morphological operation (closing) for producing enhanced binary images. These are inputted to the mid-point ellipse cropping algorithm in which all mid-points are extracted depending on the left and right coordinates. Then the direction changes towards the top, as shown in the red area of Fig. 5. Then a black and white mask (binary image) is applied to the original image to extract the ROI.

In the proposed method, three types of feature extraction operators extract the discriminatory properties of the ROI in which the size of the cell is 128 × 128. The LBP produces 59 features, HOG produces 104

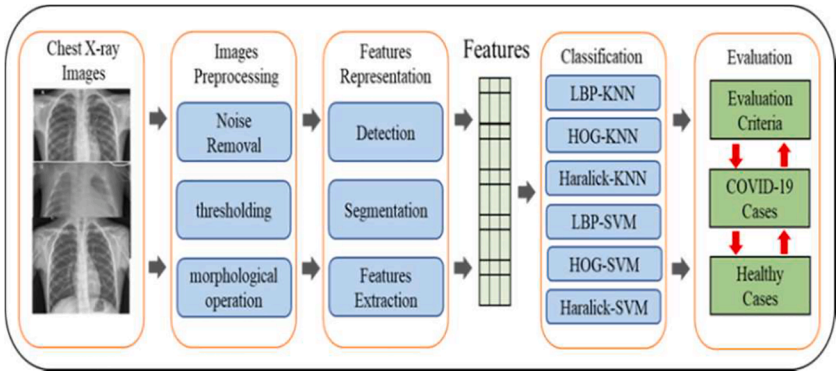


Fig. 4. The model of the proposed COVID-19 diagnosis method.

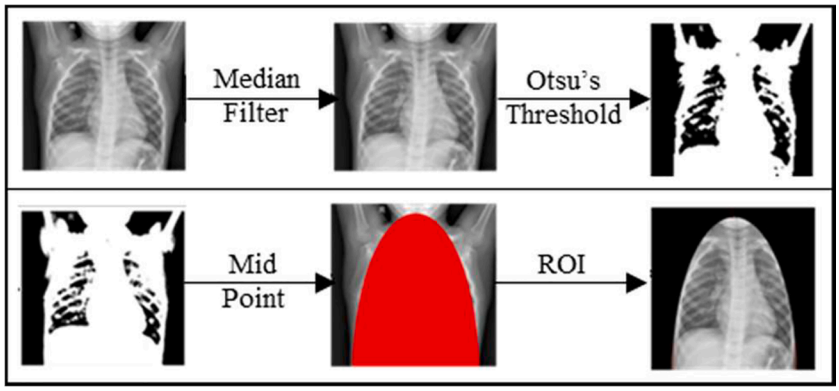


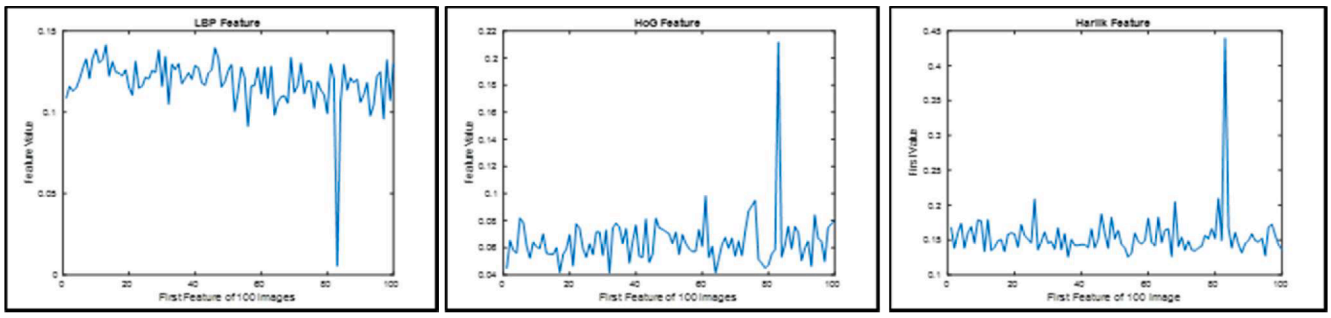
Fig. 5. ROI Cropping of an image.

features, and the Haralick produces a set of unspecified important points for each image, as shown in Fig. 6(a-c). The extracted features in the previous step are used in the classification process and to build the classification model for COVID-19 disease. All the features are used for training and testing the KNN and SVM classifiers [26,27,28].

Ultimately, the combinations of the classifiers and features extraction operators produce six models, namely LBP-KNN, HOG-KNN, Haralick-KNN, LBP-SVM, HOG-SVM, and Haralick-SVM. To produce robust results, the tests consist of multiple training conditions of 5-folds cross-validation (50%, 60%, 70%, 80%, and 90%). The evaluation considers comprehensive criteria of the confusion matrix, including accuracy, sensitivity, specificity, precision, prevalence, error rate, and false-positive rate, as shown in Table 1 and Table 2. Table 1 shows the 5-fold cross-validation evaluation results of the three KNN-based classification models, while Table 2 shows the evaluation results of the three SVM-based classification models.

Table 1 and Fig. 7(a) show the classification results of the KNN for the LBP, HOG, and Haralick features. As it can be observed from the results that the LBP-KNN model outperforms the other two models in which the average accuracy score of 5-folds is 98.66%. Moreover, it has the highest sensitivity of 97.76%, perfect specificity of 100%, perfect precision of 100%, the lowest error rate of 1.34%, and zero false positives. HOG-KNN and Haralick-KNN performance are relatively equal in which the average accuracy scores of 5-folds are 94.26% and 95.51%, respectively.

On the other hand, Table 2 and Fig. 7(b) show the 5-folds cross-validation evaluation results of the three SVM-based classification models. Generally, the LBP-SVM, HOG-SVM, and Haralick-SVM models show lower performance than the KNN-based classification models. It is mainly because each of these classification methods is performed based on different approaches. The hyperplane of the SVM separates the data points based on a kind of restrictive assumption, while the k-NN uses a non-parametric technique to approximate data distribution which is more suitable for the topology of the extracted features that has low



(a) LBP feature (b) HOG feature (c) Haralick feature

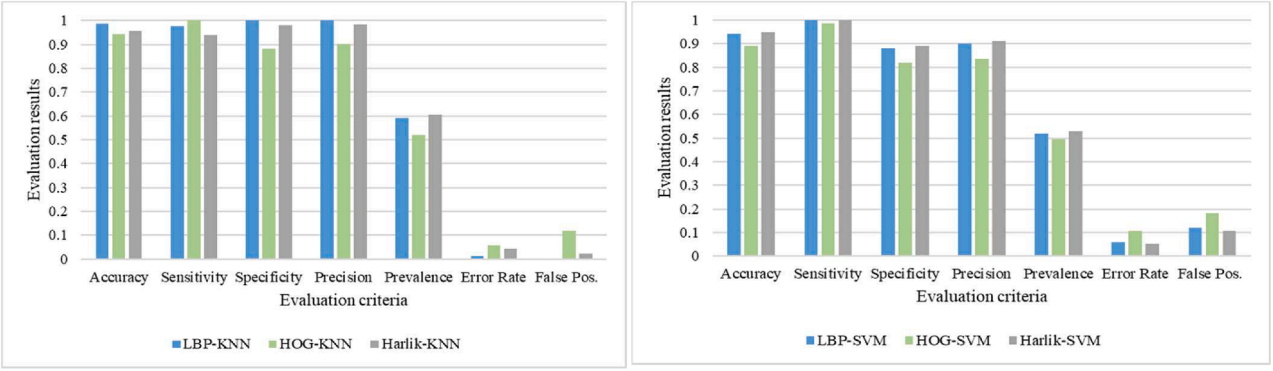
Fig. 6. Feature extraction samples of the three methods.

Table 1
The evaluation results of the KNN-based classification models.

Method	Training rate	Accuracy	Sensitivity	Specificity	Precision	Prevalence	Error Rate	False Pos.
LBP	50%	0.9830	0.9720	1.0000	1.0000	0.596	0.017	0.0000
	60%	0.9850	0.9750	1.0000	1.0000	0.594	0.015	0.0000
	70%	0.9880	0.9800	1.0000	1.0000	0.595	0.012	0.0000
	80%	0.9870	0.9780	1.0000	1.0000	0.597	0.013	0.0000
	90%	0.9900	0.9830	1.0000	1.0000	0.585	0.01	0.0000
HOG	50%	0.9350	0.9993	0.8677	0.8879	0.5129	0.065	0.1323
	60%	0.9380	0.9996	0.8724	0.8932	0.5163	0.062	0.1276
	70%	0.9447	0.9992	0.8850	0.9053	0.5239	0.0553	0.1150
	80%	0.9454	0.9995	0.8844	0.9066	0.5293	0.0547	0.1156
	90%	0.9498	0.9998	0.8956	0.9122	0.5200	0.0502	0.1044
Haralick	50%	0.9522	0.9378	0.9748	0.9825	0.6053	0.0478	0.0252
	60%	0.9539	0.9392	0.9769	0.9841	0.6067	0.0462	0.0231
	70%	0.9530	0.9364	0.9792	0.9858	0.6100	0.047	0.0208
	80%	0.9577	0.9417	0.9828	0.9880	0.6060	0.0423	0.0172
	90%	0.9588	0.9443	0.9806	0.9871	0.6011	0.0412	0.0194

Table 2
The evaluation results of the SVM-based classification models

Method	Training rate	Accuracy	Sensitivity	Specificity	Precision	Prevalence	Error Rate	False Pos.
LBP	50%	0.9350	0.9993	0.8677	0.8879	0.5129	0.0650	0.1323
	60%	0.9380	0.9996	0.8724	0.8932	0.5163	0.0620	0.1276
	70%	0.9447	0.9992	0.8850	0.9053	0.5239	0.0553	0.1150
	80%	0.9454	0.9995	0.8844	0.9066	0.5293	0.0547	0.1156
	90%	0.9498	0.9998	0.8956	0.9122	0.5200	0.0502	0.1044
HOG	50%	0.8513	0.9536	0.8086	0.8189	0.5176	0.1487	0.1914
	60%	0.8857	0.9803	0.8204	0.8330	0.4979	0.1143	0.1796
	70%	0.9011	0.9968	0.8163	0.8342	0.4863	0.0989	0.1837
	80%	0.9074	0.9999	0.8200	0.8402	0.4881	0.0926	0.1800
	90%	0.9147	0.9996	0.8324	0.8529	0.4935	0.0853	0.1676
Haralick	50%	0.9364	0.9995	0.8684	0.8914	0.5205	0.0636	0.1316
	60%	0.9436	0.9995	0.8819	0.9033	0.5250	0.0565	0.1181
	70%	0.9515	0.9994	0.8978	0.9167	0.5309	0.0485	0.1022
	80%	0.9538	0.9998	0.9024	0.9199	0.5291	0.0463	0.0976
	90%	0.9586	0.9997	0.9112	0.9288	0.5382	0.0414	0.0888



(a) The KNN-based classification models (b) The SVM-based classification models

Fig. 7. The average results of the classification models.

dimensional space.

Subsequently, in the classification results of the SVM for the LBP, HOG, and Haralick features, the Haralick-SVM model outperforms the other two models in which the average accuracy score of 5-folds is 94.88%, as shown in Fig. 7(b). Moreover, it has the highest sensitivity of 99.96%, highest specificity of 89.23%, highest precision of 91.2%, the lowest error rate of 5.13%, and lowest false positive of 10.77%. The HOG-SVM comes second and slightly lower than the Haralick-SVM with an average accuracy score of 5-folds of 94.26% and HOG-KNN performance relatively lower than both of them in which the average accuracy score of 5-folds is 89.2%. The average prevalence of Haralick-SVM is slightly better than the other five models.

The results validate the ability of the proposed method for early detection and classification of COVID-19 through image processing using X-ray images [2,3,8]. The combinations of the feature extraction operators and classifiers outcome six models, namely LBP-KNN, HOG-KNN, Haralick-KNN, LBP-SVM, HOG-SVM, and Haralick -SVM on 5,000 X-ray images. The results further show that the chest X-ray images are considered one of the best means for the detection of COVID-19 [7,11,12,13]. However, the limitation of this work includes the availability of limited samples of tested X-ray image cases, so the models have not been tested in big data for further verification of the research findings [29,30,31].

Conclusion

By applying the proposed method, which is the classification of X-ray images of corona patients, the test results have shown that it is possible through X-ray images to detect the disease by training the machine learning algorithms on an image dataset. A set of images are taken from the Kaggle website, which includes X-ray images of normal and abnormal cases of tested people (about 5,000 images) that are tested with results through a different group of random samples taken from total images for a number of iterations with different training size as explained before.

The development methodology of this work includes preprocessing, segmentation, feature extraction, and classification. The preprocessing includes image noise removal, image thresholding, and morphological operation. The segmentation is performed by Region of Interest (ROI) detection. The feature extraction includes multiple operators of Local binary pattern (LBP), Histogram of Gradient (HOG), and Haralick features. Finally, classification is performed by K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). Subsequently, a combination of six models LBP-KNN, HOG-KNN, Haralick-KNN, LBP-SVM, HOG-SVM, and Haralick-SVM are proposed. The six models are tested, and the accuracy, error rate, sensitivity, false-positive rate, specificity, precision, and prevalence of the models are calculated. The obtained results are relatively high in which the diagnosis accuracies of all tested cases are between 89.2% and 98.66% on average. The LBP-KNN model outperforms the other models in which it achieves an average accuracy of 98.66%, the sensitivity of 97.76%, the specificity of 100%, precision of 100%, the error rate of 1.34%, and zero false positive. Using more than one method of feature extraction and classification, the results are confirmed and validated. The future work includes using other combinations of feature extraction and classification operators such as the Gabor filter and random forest, and designing and testing the proposed system on real devices such as the radiographic thorax.

Declarations

Funding: This article is funded by the projects SP2021/45 and SP2021/32, assigned to VSB-Technical University of Ostrava, the Ministry of Education, Youth and Sports in the Czech Republic.

CRediT authorship contribution statement

Jamal N. Hasoon: Conceptualization, Funding acquisition, Investigation, Resources, Supervision, Writing – original draft, Writing – review & editing. **Ali Hussein Fadel:** Conceptualization, Investigation, Software, Visualization. **Rasha Subhi Hameed:** Conceptualization, Data curation, Software, Validation, Visualization, Writing – review & editing. **Salama A. Mostafa:** Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Software, Supervision, Writing – original draft, Writing – review & editing. **Bashar Ahmed Khalaf:** Formal analysis, Funding acquisition, Resources, Validation. **Mazin Abed Mohammed:** . **Jan Nedoma:** .

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the Department of Computer Science, Mustansiriyah University, and Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia. Also, it is supported by the Department of Computer Science, University of Diyala, Diyala, Iraq.

Data availability

The used dataset in this work is public and available online in a data repository.

Human and animal rights

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Atangana A, Araz Sİ. Mathematical model of COVID-19 spread in Turkey and South Africa: theory, methods, and applications. Advances in Difference Equations 2020; 2020(1):1–89.
[2] A. S. Al-Waisy M. Abed Mohammed S. Al-Fahdawi M. S. Maashi B. Garcia-Zapirain K. Hameed Abdulkareem et al. COVID-DeepNet: hybrid multimodal deep learning system for improving COVID-19 pneumonia detection in chest X-ray images 67 2 2021 2409 2429.
[3] Hemdan, E. E. D., Shouman, M. A., & Karar, M. E. (2020). Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. *arXiv preprint arXiv:2003.11055*.
[4] WHO Coronavirus Disease (COVID-19) Dashboard, available online, https://covid19.who.int/?gclid=EAIaIQobChMIt- aK0_v36QIVkn4rCh2v Ew0nEAAYASAAEgKqnfD_BwE, Accessed on 10/6/2020.
[5] Yan L, Zhang HT, Xiao Y, Wang M, Sun C, Liang J, et al. Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. MedRxiv. 2020.
[6] Zhou T, Lu H, Yang Z, Qiu S, Huo B, Dong Y. The ensemble deep learning model for novel COVID-19 on CT images. Appl Soft Comput 2021;98:106885. https://doi.org/10.1016/j.asoc.2020.106885.
[7] Al-Waisy AS, Al-Fahdawi S, Mohammed MA, Abdulkareem KH, Mostafa SA, Maashi MS, et al. COVID-CheXNet: hybrid deep learning framework for identifying COVID-19 virus in chest X-rays images. Soft Comput 2020:1–16.
[8] Din RU, Shah K, Ahmad I, Abdeljawad T. Study of transmission dynamics of novel COVID-19 by using mathematical model. Advances in Difference Equations 2020; 2020(1):1–13.
[9] Sethy PK, Behera SK. Detection of coronavirus disease (covid-19) based on deep features. Preprints 2020;2020030300:2020.
[10] Xu X, Jiang X, Ma C, Du P, Li X, Lv S, et al. A deep learning system to screen novel coronavirus disease 2019 pneumonia. Engineering 2020;6(10):1122–9.
[11] Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med 2020;121:103792. https://doi.org/10.1016/j.compbiomed.2020.103792.
[12] Narin, A., Kaya, C., & Pamuk, Z. (2020). Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. arXiv preprint arXiv:2003.10849.
[13] Cohen, J. P., Dao, L., Morrison, P., Roth, K., Bengio, Y., Shen, B., & Duong, T. Q. (2020). Predicting covid-19 pneumonia severity on chest x-ray with deep learning. arXiv preprint arXiv:2005.11856.
[14] Mohammed, Naman Goyal. Circle and Ellipse drawing Algorithms. (2017) Roll No.: UE143059UIET, PU, Chandigarh. https://www. Scribd.com/document/356462640/Circle-and-Ellipse-Drawing-Algorithm.
[15] Liu Y, Ge Y, Wang F, Liu Q, Lei Y, Zhang D, et al. A Rotation Invariant HOG Descriptor for Tire Pattern Image Classification. ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019.
[16] Khaleefah, S. H., Mostafa, S. A., Mustapha, A., & Nasrudin, M. F. (2019). The ideal effect of Gabor filters and Uniform Local Binary Pattern combinations on deformed scanned paper images. *Journal of King Saud University-Computer and Information Sciences*.
[17] N. Zayed H.A. Elnemr 2015 2015 2015 1 7.
[18] Setiawan AS, Elysia, Wesley J, Purnama Y. Mammogram classification using law's texture energy measure and neural networks. Procedia Comput Sci 2015;59:92–7.
[19] Thanh Noi P, Kappas M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. Sensors 2018;18(1):18.
[20] Mostafa SA, Mustapha A, Mohammed MA, Hamed RI, Arunkumar N, Abd Ghani MK, et al. Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease. Cognit Syst Res 2019;54:90–9.
[21] Taneja S, Gupta C, Aggarwal S, Jindal V. In: March). MFZ-KNN—A modified fuzzy based K nearest neighbor algorithm. IEEE; 2015. p. 1–5.
[22] Han Y, Li J, Li JZ, Xing HW, Yang AM, Pan YH. In: September). Demonstration of SVM Classification Based on Improved Gauss Kernel Function. Cham: Springer; 2016. p. 189–95.
[23] Obaid OI, Mohammed MA, Ghani MKA, Mostafa A, Taha F. Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. *International Journal of*. Eng Technol 2018;7(4.36):160–6.
[24] Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., and Shen, D. (2020). Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. *IEEE reviews in biomedical engineering*.

[25] Ng MY, Lee EY, Yang J, Yang F, Li X, Wang H, et al. Imaging profile of the COVID-19 infection: radiologic findings and literature review. *Radiology: Cardiothoracic. Imaging* 2020;2(1):e200034.
[26] Atangana A, araz SİĞret. A novel Covid-19 model with fractional differential operators with singular and non-singular kernels: Analysis and numerical scheme based on Newton polynomial. Alexandria Engineering Journal 2021;60(4): 3781–806.
[27] Zhang Z. A novel covid-19 mathematical model with fractional derivatives: Singular and nonsingular kernels. Chaos, Solitons Fractals 2020;139:110060. https://doi.org/10.1016/j.chaos.2020.110060.
[28] Umar M, Sabir Z, Raja MAZ, Amin F, Saeed T, Guerrero-Sanchez Y. Integrated neuro-swarm heuristic with interior-point for nonlinear SITR model for dynamics of novel COVID-19. Alexandria Engineering Journal 2021;60(3):2811–24.
[29] Mohammed MA, Abdulkareem KH, Al-Waisy AS, Mostafa SA, de la Torre I, Díez. Benchmarking Methodology for Selection of Optimal COVID-19 Diagnostic Model Based on Entropy and TOPSIS Methods. IEEE Access 2020;8:99115–31.
[30] Shoaib M, Raja MAZ, Sabir MT, Bukhari AH, Alrabaiah H, Shah Z, et al. A stochastic numerical analysis based on hybrid NAR-RBFs networks nonlinear SITR model for novel COVID-19 dynamics. Comput Methods Programs Biomed 2021;202:105973. https://doi.org/10.1016/j.cmpb.2021.105973.
[31] Ioannis D. Apostolopoulos Tzani A. Mpesiana Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks 43 2 2020 635 640.

Jamal N. Hasoon[a], Ali Hussein Fadel[b], Rasha Subhi Hameed[b], Salama A. Mostafa[c,*], Bashar Ahmed Khalaf[d], Mazin Abed Mohammed[e], Jan Nedoma[f]

[a] *Department of Computer Science, Mustansiriyah University, 10001 Baghdad, Iraq*
[b] *Department of Computer Science, University of Diyala, 32001 Diyala, Iraq*
[c] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Johor, Malaysia*
[d] *Department of Medical Instruments Engineering Techniques, Bilad Alrafidain University College, 32001 Diyala, Iraq*
[e] *College of Computer Science and Information Technology, University of Anbar, Anbar 31001, Iraq*
[f] *Department of Telecommunications, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, 70800 Ostrava, Czech Republic*

[*] Corresponding author.
E-mail addresses: jamal.hasoon@uomustansiriyah.edu.iq (J.N. Hasoon), salama@uthm.edu.my (S.A. Mostafa), bashar@bauc14.edu.iq (B.A. Khalaf), mazinalshujeary@uoanbar.edu.iq (M.A. Mohammed), jan.nedoma@vsb.cz (J. Nedoma).