COVID-19 Detection from Chest X-ray Scans using Machine Learning

Duaa F. Eljamassi
Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine
duaa.jamassi@gmail.com

Ashraf Yunis Maghari
Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine
amaghari@iugaza.edu.ps

***Abstract*— Recently, the virus (COVID-19) has spread widely throughout the world and has led to the examination of large numbers of suspected cases using standard COVID-19 tests and has become pandemic. Everyday life, public health and the global economy have been destroyed. The pathogenic laboratory tests such as Polymerase chain reaction (PCR) take a long time with false negative results and are considered the gold standard for diagnosis. Therefore, there was an urgent need for rapid and accurate diagnostic methods to detect COVID-19 cases as soon as possible to prevent the spread of this epidemic and combat it. Applying advanced artificial intelligence techniques along with radiography may be helpful in detecting this disease. In this study, we propose a classification model that detect the infected condition through the chest X-ray images. A dataset containing chest x-ray images of normal people, people with pneumonia such as SARS, streptococcus and pneumococcus and other patients with COVID-19 were collected. Histogram of oriented gradients (HOG) is used for image features extraction. The images are then classified using Support Vector Machines (SVM), random forests and K- nearest neighbors (KNN), with classification rate 98.14%, 96.29% and 88.89% respectively. These results may contribute efficiently in detecting COVID-19 disease.**

Keywords*— *COVID-19, HOG, SVM, KNN, Classification, X-ray images.

I. INTRODUCTION

In late December 2019, the previously reported new Coruna virus outbreak in Wuhan, China (2019-nCoV), which later spread to 26 countries around the world and It is considered an acute disease and may also be fatal with 2% of deaths [1]. The diagnosis of COVID-19 is related to Fever, cough, fatigue and muscle pain in the human body during the early stages, as patients suffer from abnormal conditions in the CT and X-ray chest images. X-rays are the first technique to play a significant role in the diagnosis of COVID-19 disease [2]. Due to the sensitivity and expediency of chest X-rays and computed tomography (CT) imaging, they are considered as the potential screening tools [2]. X-ray and CT images have been utilized by some countries such as Italy, China and Spain in addition to reverse transcription polymerase chain reaction (RT-PCR) to examine and manage the patient's condition [3].

Machine learning algorithms, called also classifiers, can receive input data, process it using statistical analysis, and predict the output based upon the data type available. All these algorithms have various applications such as image recognition, voice recognition, predictions, sentiment analysis, etc. The development of machine learning has a great advantages in clinical decision making and development computer-aided systems [4]. The application of automated learning methods in diagnostics in the medical field was welcomed by doctors, to become an aid to them [5].

Recently, many researches were conducted on detecting COVID-19 using advanced artificial intelligence techniques. There are some researches, related to deep learning, have used chest X-ray image data to detect the disease. In [6], Abiyev and Ma'aitah used CNN models to classify the data using chest x-ray images of pneumonia and for the normal where 70% of the data was for training, and the accuracy was 89.57%. Chouhan et al. [7] used deep learning models to detect the pneumonia images using normal, virus pneumonia, and bacterial pneumonia images as datasets. They proposed a deep learning framework for pneumonia detection using transfer learning concept. The image features were extracted using NN models based on ImageNet to train five different models. Then, all pretrained models were combined using an ensemble model. They achieved a classification accuracy of 96.39%. Also, In [4], Sethy et al. used classification model to detect (COVID-19) Disease based on Deep Features and SVM, i.e. ResNet50 plus SVM, the average accuracy was 95.33%, which was the average of 20 independent executions.

This paper focuses on using machine learning approaches with chest X-ray images to detect if the image is infected with COVID-19 or not. Some researchers used machine learning classifiers for detecting COVID-19. For example, Sarhan [8] used SVM and wavelets to detect COVID-19 cases in chest x-ray Images. He used a threshold value of 903, decomposition level of two, and a dB 1 wavelet and scored an accuracy of 94.5%. Ohata et al. [9] applied many classifiers on 194 X-ray images for healthy patients and the same number of images for patients infected with COVID-19. They applied the transfer learning and CNNs trained with ImageNet to extract the images features. Then, CNNs are combined with KNN, Naive Bayes, Random Forest, MLP, and SVM. The best accuracy of 98.5% was obtained when they used the MobileNet with SVM classifier (linear kernel). In [10], Yadav and Jadhav used SVM as a classification method to classify pneumonia data into three categories: viral pneumonia, normal, and bacterial pneumonia. They used a deep learning approach such as VGG-16,

© IEEE 2021. This article is free to access and download, along with rights for full text and data mining, re-use and analysis.

DOI 10.1109/ICPET51420.2020.00009

InceptionV3 models. Their results showed a good accuracy of 96.6%. Moreover, et al. [11] used the sigmoid function to classify pneumonia data to determine classification probabilities, they achived an accuracy of 93.73%. Recently, Chakraborty [12] used a dataset of X-rays images of normal people, pneumonia patients and COVID-19 patients, as 100 images for each class. Transfer learning was used with VGG-16 model with significantly high accuracy in detecting COVID-19 in test data.

In our work, however, a classification model is proposed to classify the chest X-ray images of COVID-19, pneumonia patient and normal people using datasets with relatively more images. The aim of this research is to diagnose and classify Covid-19 disease, healthy people and pneumonia patients through chest X-ray images. Symptom of pneumonia in most cases and Covid-19 virus have the same ranges, both are diseases associated with the lungs. Hence, in our study, we prepared three classes of datasets contains chest X-ray images of COVID-19 patient, pneumonia patient and normal people. The dataset consists of 520 chest x-ray images. The images are preprocessed and prepared to be used by three machine learning algorithms; random forests, K-nearest neighbors (KNN) and SVM algorithm.

The paper is organized as follows: Section two describes the proposed prediction model. Section three demonstrates the experiments and results, finally the paper is concluded in Section four.

II. THE PROPOSED PREDICTION MODEL

Our proposed prediction model consists of five main steps. The first step being data collection, followed by data preprocessing which prepare the chest X-ray images for feature extraction where Histogram of oriented gradients (HOG) feature is extracted (third step). In the fourth step, HOG feature is trained by the classifiers to detect the infected condition of COVID-19. The final step is evaluating the performance of the three used classifiers. Fig. 1 gives the basic overview of proposed model.

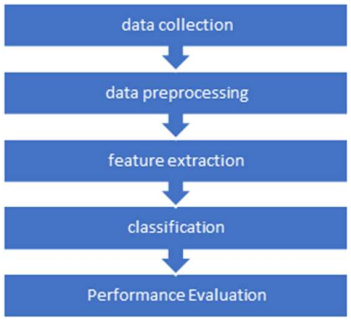

Fig. 1. Proposed classification model

A. Chest X-ray image Datasets

In our study, we prepared three classes of chest X-ray images dataset (COVID-19 patients, normal people, and pneumonia patients). It contains 520 images (120 COVID-19 images, 200 pneumonia images and 200 healthy images). COVID-19 is a new disease, so, the number of COVID-19 chest x-ray images is limited. The datasets were collected from the GitHub website and the Kaggle website [13]. Then, the datasets are divided into 80% for training and 20% to test the classifiers. Fig. 2 shows an example of chest X-ray image Datasets.

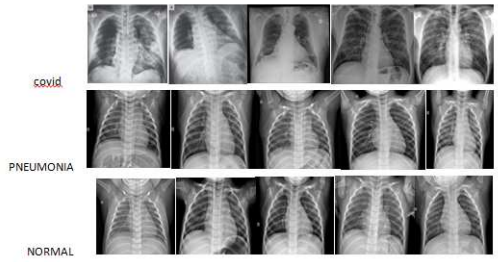

Fig. 2. Example of chest X-ray image Datasets

B. Data Pre-Processing

The dataset is already preprocessed by [13] and resizing it to *64×128* pixels so that it is ready for training and testing our model.

C. Feature Extraction

After pre-processing, features are extracted from images by using Histogram of oriented gradients (HOG). The basic advantages are describing the shape and contour properties of an image. Fig. 3 shows an example of HOG feature extracted from an x-ray image [14].

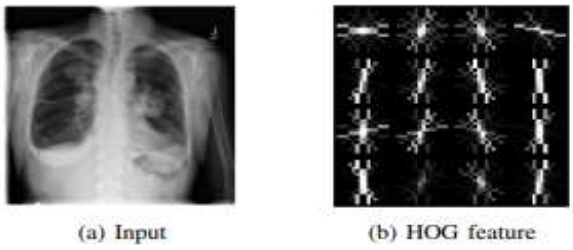

Fig. 3. Feature extracted by histogram of oriented gradients (HOG).

D. Classification

Support vector machine (SVM), K-nearest neighbors (KNN) and random forests were used for Classification.

a) Support Vector Machine (SVM)

SVM is a supervised algorithm used for both data classification and regression challenges. It is often used for data classification problems. SVM uses the kernel trick as a technique to have a linear hyper-plane between classes. The advantages of SVM compared to some classification algorithms are its speed, efficiency, and accuracy [15]. It is also considered as the optimal method for small dataset classification [15].

b) K-Nearest Neighbors (KNN)

KNN is one of the nonparametric learning algorithms that used for classification and regression [16]. It is also the simplest machine learning algorithms that has many applications in different fields such as pattern recognition and analysis of images [17]. KNN is a typical type of learning scheme based on

instances which make it an efficient algorithm. KNN requires an integer k to determine the total number of neighbors used for classification, a training data set and a metric to measure closeness between neighbors. When k is equal to 1, then the sample class is determined by the nearest neighbors [16]. The Euclidean distance is usually used to determine the closeness between two points in the training dataset [17].

c) Random Forest

It is a machine learning technique which comprises a large number of decision trees used for class prediction. The random forest is an ensemble model uses the results from many different decision trees to predict a class. Several individual decision trees are created and the class with the most votes is used to calculate the response [18]. Mostly, ensemble model outperforms better than any one of the individual models. Random forest can rank the variables importance in a natural way for the classification problem. Many algorithms such as Build Tree, Split, Can Split, etc. can be used by random forest [19].

d) Performance Evaluation

Equations (1) to (4) give the accuracy, precision, F1-score and recall respectively. They are used to evaluate the three classifies.

$$Acc = \frac{\text{Number of correctly classified images}}{\text{Total Number of images}} \times 100\% \quad (1)$$

$$P = \frac{\text{Number of correctly classified images per class}}{\text{Total Number of images per class}} \times 100\% \quad (2)$$

$$R = \frac{\text{Number of correctly classified images}}{\text{N. of expected images in the corresponding class}} \times 100\% \quad (3)$$

$$F - score = 2 \times \frac{\text{P}\times\text{R}}{\text{P}+\text{R}} \times 100\% \quad (4)$$

III. EXPERIMENTS AND RESULTS

In this paper we used Toshiba laptop with 4 GB RAM and Python as programming language. The extracted features from the images are given as an input to the three classifiers. Table I provide the classification accuracies of the three classifiers. The results show that SVM scored the highest accuracy with 98.14%. On the other hand, our proposed model (SVM with HOG) was compared with some related works as shown in Table II. It shows that our model outperformed the woks presented by Sethy et al. [4] and Sarhan [8] and comparable with the results obtained by Ohata et al. [9].

Our results are consistent with the results obtained by Ohata et al. [9] where the MobileNet and SVM classifier with a linear kernel are used with accuracy of 98.5%. On the other hand, our results outperformed the results obtained by Sethy et al. [4] where SVM plus ResNet50 classification model of COVID-19 images was used with average accuracy of 95.33% among 20 independent executions. It also outperformed the results obtained by Sarhan [8] where wavelets and SVM were applied to detect COVID-19 in chest x-ray images with accuracy of 94.5%. These results demonstrate that using HOG feature with SVM can produce very good results.

TABLE I. ACCURACY RESULTS FOR CLASSIFICATION

	SVM	**Random forest**	**KNN**
Accuracy	98.14%	96.29%	88.89%

TABLE II. THE ACCURACY OF THE PROPOSED MODEL COMPARED WITH SOME RECENT WORKS

	Method	**Bst Accuracy**
OUR MODEL	SVM with HOG	98.14%
Sethy et al. [4]	ResNet50 plus SVM	95.33%
Sarhan [8]	SVM with wavelets	94.5%
Ohata et al. [9]	MobileNet with the SVM	98.5%

Further, Table III, Table IV and Table V provide the performance metrics (Precision, Recall, F1-Score) for SVM, Random Forest, and KNN respectively. The results show that the performance of SVM and random forest are better than the KNN.

TABLE III. CLASSIFICATION REPORT USING SVM.

	Normal	**Pneumonia**	**COVID-19**
Precision	1	0.98	0.95
Recall	1	0.98	0.95
F1-score	1	0.95	0.95

TABLE IV. CLASSIFICATION REPORT USING RANDOM FOREST

	Normal	**Pneumonia**	**COVID-19**
precision	0.93	0.98	1
recall	1	0.93	0.95
F1-score	0.97	0.95	0.98

TABLE V. CLASSIFICATION REPORT USING KNN

	Normal	**Pneumonia**	**COVID-19**
precision	0.84	0.91	1
recall	0.98	0.95	0.59
F1-score	0.90	0.93	0.74

IV. CONCLUSIONS

According to Worldometer website [20], the coronavirus COVID-19 is affecting 218 countries and territories around the world with more than 60 million infected cases. Everyday life, public health and the global economy have been destroyed. The pathogenic laboratory tests such as Polymerase chain reaction (PCR) take a long time with false negative results. In this study, we suggested a machine learning model to classify and detect cases of COVID-19 from chest X-rays. Our model used three classification algorithms; SVM, KNN and random forests. The datasets contained chest x-ray images of normal people, people with pneumonia such as SARS, streptococcus and pneumococcus and other patients with COVID-19. The images features are extracted using Histogram of oriented gradients (HOG) technique. The results showed that SVM scored the highest accuracy with 98.14%. KNN and Random Forest scored 96.29 and 88.89 accuracies.

The classification accuracy could be more effective if we use more images, whereas the restrictions on the study was using a limited number of chest X-ray images of COVID19 patient. The classification can be also further improved by utilizing more image features.

REFERENCES

[1] Z. Xu *et al.*, "Pathological findings of COVID-19 associated with acute respiratory distress syndrome," *Lancet Respir. Med.*, vol. 8, no. 4, pp. 420–422, 2020, doi: 10.1016/S2213-2600(20)30076-X.

[2] S. Hussain and A. Khan, "Coronavirus Disease Analysis using Chest X-ray Images and a Novel Deep Convolutional Neural Network," no. April, 2020, doi: 10.13140/RG.2.2.35868.64646.

[3] D. Caruso *et al.*, "Chest CT features of COVID-19 in Rome, Italy," *Radiology*, p. 201237, 2020.

[4] P. K. Sethy, S. K. Behera, P. K. Ratha, and P. Biswas, "Detection of coronavirus Disease (COVID-19) based on Deep Features and Support Vector Machine," *Int. J. Math. Eng. Manag. Sci.*, vol. 5, no. 4, pp. 643–651, 2020, doi: 10.33889/ijmems.2020.5.4.052.

[5] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Comput. Biol. Med.*, vol. 121, no. April, p. 103792, 2020, doi: 10.1016/j.compbiomed.2020.103792.

[6] R. H. Abiyev and M. K. S. Ma'aitah, "Deep convolutional neural networks for chest diseases detection," *J. Healthc. Eng.*, vol. 2018, 2018.

[7] V. Chouhan *et al.*, "A novel transfer learning based approach for pneumonia detection in chest X-ray images," *Appl. Sci.*, vol. 10, no. 2, p. 559, 2020.

[8] A. Sarhan, "Detection of COVID-19 Cases In Chest X-ray Images Using Wavelets And Support Vector Machines," pp. 1–13, 2020, doi: 10.21203/rs.3.rs-37558/v1.

[9] E. F. Ohata *et al.*, "Automatic detection of COVID-19 infection using chest X-ray images through transfer learning," *IEEE/CAA J. Autom. Sin.*, 2020, doi: 10.1109/JAS.2020.1003393.

[10] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *J. Big Data*, vol. 6, no. 1, p. 113, 2019.

[11] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," *J. Healthc. Eng.*, vol. 2019, 2019.

[12] S. Chakraborty, "An attempt- Detection of COVID-19 presence from Chest X-ray scans using CNN & Class Activation Maps," 2020. [Online]. Available: https://towardsdatascience.com/detection-of-covid-19-presence-from-chest-x-ray-scans-using-cnn-class-activation-maps-c1ab0d7c294b. [Accessed: 03-Jun-2020].

[13] P. Moony, "Chest X-Ray Images (Pneumonia)." [Online]. Available: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia/data. [Accessed: 04-Feb-2020].

[14] S. Reza, O. B. Amin, and M. M. A. Hashem, "A Novel Feature Extraction and Selection Technique for Chest X-ray Image View Classification," in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*, 2019, pp. 189–194.

[15] M. Meng and C. Zhao, "Application of support vector machines to a small-sample prediction," *Adv. Pet. Explor. Dev*, vol. 10, no. 2, pp. 72–75, 2015, doi: http://dx.doi.org/10.3968/7830.

[16] I. A. A. Amra and A. Y. A. Maghari, "Students performance prediction using KNN and Naïve Bayesian," in *2017 8th International Conference on Information Technology (ICIT)*, 2017, pp. 909–913, doi: 10.1109/ICITECH.2017.8079967.

[17] A. Maghari, I. Venkat, I. Y. Liao, and B. Belaton, "Adaptive face modelling for reconstructing 3D face shapes from single 2D images," *IET Comput. Vis.*, vol. 8, no. 5, pp. 441–454, 2014, doi: 10.1049/iet-cvi.2013.0220.

[18] N. Horning, "Random Forests: An algorithm for image classification and generation of continuous fields data sets," in *Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan*, 2010, vol. 911.

[19] M. A. M. Alhaj and A. Y. A. Maghari, "Cancer survivability prediction using random forest and rule induction algorithms," in *2017 8th International Conference on Information Technology (ICIT)*, 2017, pp. 388–391, doi: 10.1109/ICITECH.2017.8080031.

[20] Worldometer, "COVID-19 Coronavirus Pandemic," 2020. [Online]. Available: https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1? [Accessed: 28-Nov-2020].