Evaluation of Machine Learning Algorithms in the Classification of Normal, Viral Pneumonia and COVID-19 patients

Rohith N Reddy

Biomedical Engineer, Panacea Medical Technologies
Benguluru, Karnataka, INDIA - 560066

---***---

Abstract -This study presents the evaluation of machine learning algorithms and implementation of the best classification technique in the early phase detection of the global outbreak known as Coronavirus (COVID-19), which is named by the World Health Organization (WHO). The experiment was conducted on chest radiographic (x-ray) images. The radiologists make use of these images for detecting the abnormalities in the lungs. The clinical experts specify that the COVID-19 virus needs to be identified in the early phase, as the COVID-19 shows fewer behavioral differences from viral pneumonia. Initially, we evaluated machine learning algorithms: SVM and KNN. A dataset of 300 chest x-ray images (normal, viral pneumonia, and COVID-19) was used. The first-order parameters, Gray Level Co-occurrence Matrix (GLCM) and Region properties were used as feature extraction methods. Both Support Vector Machines (SVM) and k-Nearest Neighbor (KNN) models were trained with the extracted features for different kernel and distance functions. Five-fold cross-validation was implemented during the classification process. The best classification accuracy of 74.7% was obtained using the Linear SVM classifier. Later, Linear SVM trained model was used in the classification of Normal, Viral Pneumonia, and COVID-19 patients.

***Key Words*:**Classification, COVID-19, Feature Extraction, Gray Level Co-occurrence Matrix, k-Nearest Neighbor, Radiographic Images, Support Vector Machine, Viral Pneumonia.

1.INTRODUCTION

According to most of the world's leading transmittable disease experts, the coronavirus (COVID-19) which started spreading from Wuhan, China is now developed into a pandemic that circles the globe. This coronavirus acts more like the highly transmissible influenza than scientists have found in its slow-moving viral relatives, SARS and MERS. The number of laboratory-confirmed cases has risen over the last 6 weeks from about 50 in China to over 27,56,000 in at least 23 countries; there have been more than 1,92,000 deaths. It's a huge leap beyond what virologists saw when SARS and MERS came up. Researchers are yet to say who is at the highest risk of developing a serious or life-threatening disease, and what factors may protect against the disease [1].

Coronaviruses are a large virus family usually targets the respiratory organ. SARS is believed to have developed in China from bats to civet cats to humans, MERS has spread from bats to camels into Middle East humans. As of now, still, there is no proper evidence from where the COVID-19 virus came into existence. Pneumonia is one of the major effects of coronavirus. Pneumonia is an infectious disease that affects the lungs causing symptoms like dry cough, fever, shortness of breath, and rapid breathing. This disease is usually caused by bacteria and viruses such as Influenza A and B viruses, Respiratory Syncytial virus, SARS-CoV-2 (COVID-19) virus, Adenoviruses, Rhinoviruses, and Parainfluenza viruses [2]. Radiographs and Computed Tomography images of lungs are used for the diagnosis of pneumonia. Therefore, clinical experts make use of these imaging modalities for the detection of COVID-19 in the early phase. Machine learning algorithms are vastly used in the biomedical image processing applications for increasing the image quality, segmentation of organs and anatomical structures, texture classification and detection of tumor nodules, and so on.

Xu et al. [3] classified CT images of COVID-19 into three classes as COVID-19, Influenza-A viral pneumonia, and healthy cases. They obtained images from the hospitals in the Zhejiang region of China. The dataset consisted of a total of 618 images, which includes 219 images from 110 patients with COVID-19, 224 images of 224 patients with Influenza-A viral pneumonia, and 175 images of 175 healthy people. They classified the images with a 3D-dimensional deep learning model and achieved an 87.6% overall classification accuracy. Shan et al. [4] developed a deep learning-based system for segmenting and quantification of the infected regions as well as the entire lung on chest CT images. They used 249 COVID-19 patients and 300 new COVID-19 patients for validation in their study. They obtained the Dice similarity coefficient as 91.6%. The normal delineation system often takes 1 to 5 hours; however, their proposed system reduced the delineation time to four minutes. Mucahid et al. [5] used 150 CT images for COVID-19 classification. Before the classification process, the four different datasets were created from the 150 CT images and the samples of datasets were labeled as coronavirus / non-coronavirus (infected/non-infected). Features extraction methods and SVM are used during the classification of the coronavirus images.

In this study, a total of 300 chest x-ray images of three classes namely, Normal (100), Viral Pneumonia (100), and COVID-19 (100) are used for classification. Fourteen different features were extracted from the images and trained SVM and KNN models were obtained. Further, models were evaluated for prediction accuracy and the best model was used in the diagnosis of the COVID-19 disease.

2.MATERIALS AND METHODS

2.1 Dataset and Feature Extraction

In the proposed method, a dataset of 100 normal, viral pneumonia and COVID-19 each x-ray chest images are taken for experimentation. The x-ray images initially undergo pre-processing operations such as image filtering by median filter and image enhancement by adaptive histogram equalization. Later, the pre-processed images are ROI selected with a window of 25 x 25 pixels, and fourteen different features are extracted. First Order Parameters: Mean, Standard Deviation, Skewness, Kurtosis, Entropy, and Variance. Gray Level Co-occurrence Matrix: Contrast, Correlation, Energy, and Homogeneity. Region Properties: 4-connectivity, 8-connectivity, Eccentricity, and Solidity.

2.2 Train Classifier Models

The extracted features are used to train the classifier models: SVM and KNN. Support Vector Machines (SVM) models were trained with different kernel functions: Linear, Cubic, and Quadratic. Whereas, k-Nearest Neighbor (KNN) models were trained with different distance metric: Euclidean (Fine KNN and Medium KNN), and Minkowski (Cubic KNN). The cross-validation method was used of 5 k-Folds to partition the data into where each fold is held out in turn for testing. This ensures the performance of the model using the data inside the fold, then calculates the average test error overall fold. This method gives a good estimate of the predictive accuracy of the final trained model. The prediction accuracy of the six trained models, three SVM, and three KNN models are found out (table1). Based on the results obtained, the Linear SVM model showed more accuracy of 74.7% over other trained models.

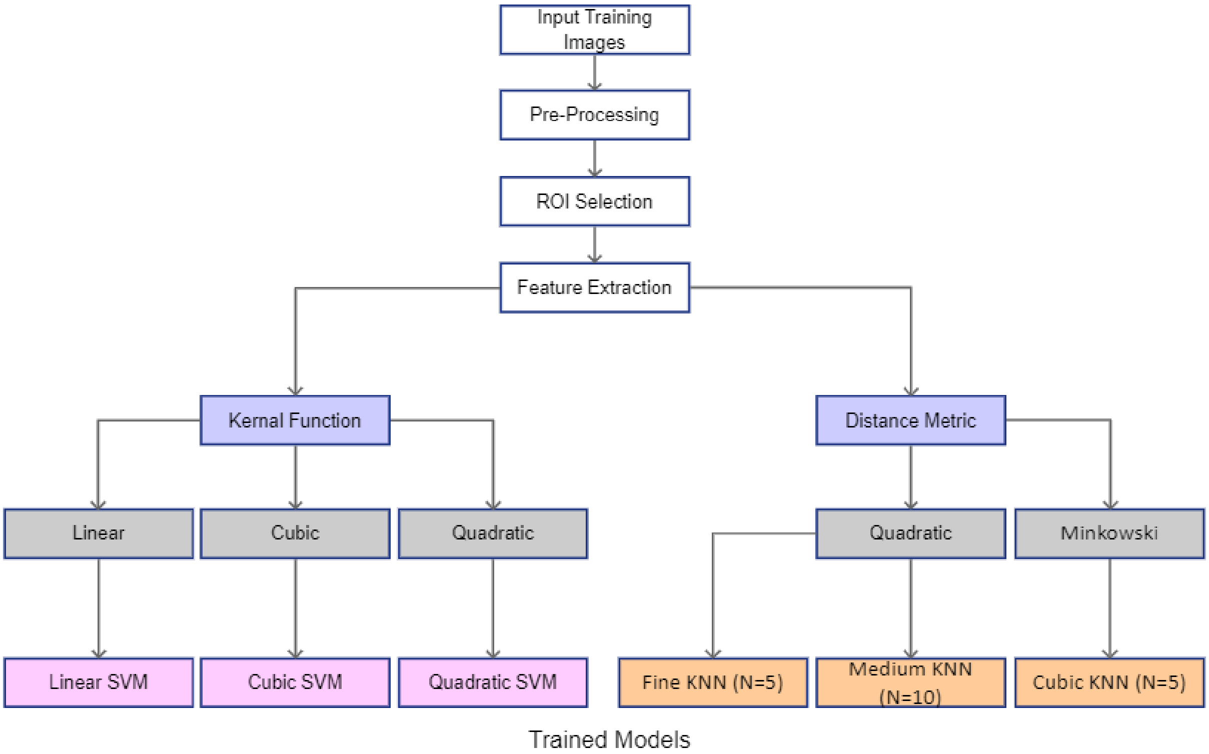


Figure1 Training Model Flow Diagram

2.3 Classification of Normal Lung, Viral Pneumonia and COVID-19 Patients

A computer vision system was implemented with a graphical user interface: Image Loading, Pre-processing (Filtering and Enhancement), Region Selection (ROI), Feature Extraction, and Classification. The fourteen extracted features will be passed through the trained classifier model. Here, we have used the Linear SVM model for the classification of the Normal, Viral Pneumonia, and COVID-19 patients.

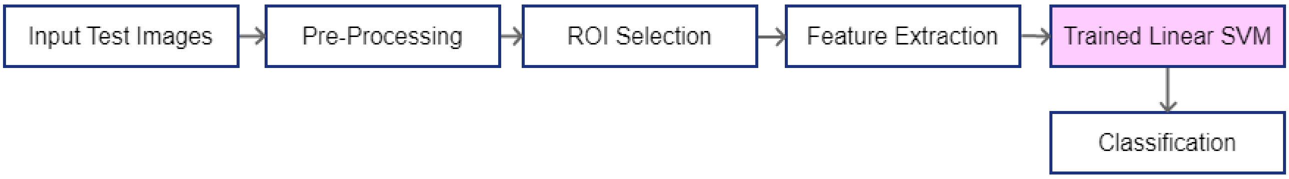


Figure2 Classification Flow Diagram

3.MATERIALS AND METHODS

This section describes the results of experiments carried out in this research. Two classification algorithms are trained and the prediction accuracy of the models is evaluated. The algorithm which gives better accuracy is considered the most efficient when applied for classification of normal, viral pneumonia and COVID-19 patients. The training process for both SVM and KNN models remain the same as described in the methodology. The SVM trained models with different kernel functions scatter plot of trained data and confusion matrices are shown below:

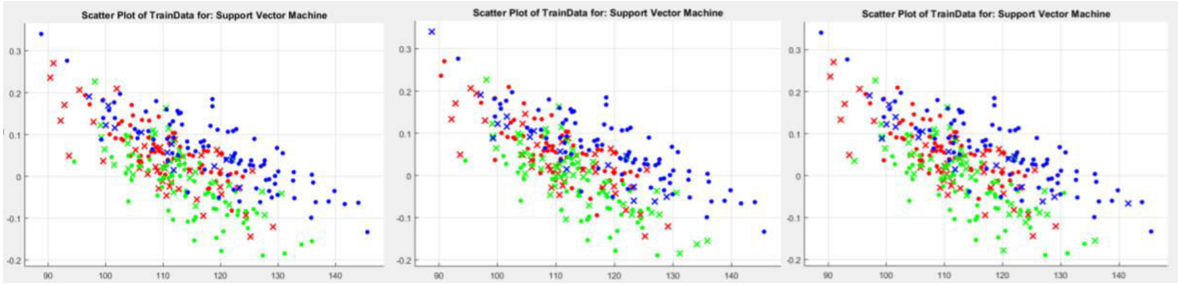
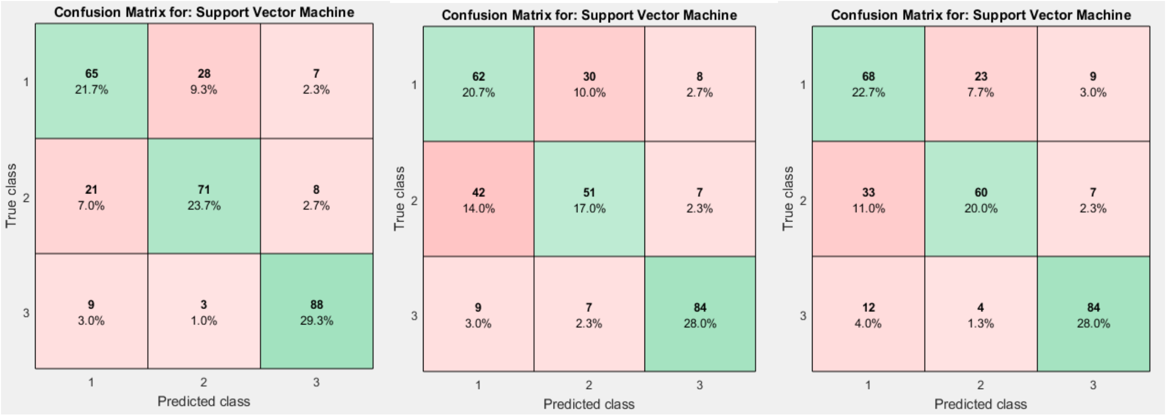


Figure3 SVM Scatter Plot of Trained Data with Kernel Functions (a) Linear (b) Cubic (c) Quadratic



Confusion Matrix for: Support Vector Machine

True class \ Predicted class	1	2	3
1	65 21.7%	28 9.3%	7 2.3%
2	21 7.0%	71 23.7%	8 2.7%
3	9 3.0%	3 1.0%	88 29.3%

Confusion Matrix for: Support Vector Machine

True class \ Predicted class	1	2	3
1	62 20.7%	30 10.0%	8 2.7%
2	42 14.0%	51 17.0%	7 2.3%
3	9 3.0%	7 2.3%	84 28.0%

Confusion Matrix for: Support Vector Machine

True class \ Predicted class	1	2	3
1	68 22.7%	23 7.7%	9 3.0%
2	33 11.0%	60 20.0%	7 2.3%
3	12 4.0%	4 1.3%	84 28.0%

Figure4 SVM Confusion Matrix with Kernel Functions (a) Linear (b) Cubic (c) Quadratic

The KNN trained models with different distance metric (Euclidean and Minkowski) scatter plot of trained data and confusion matrices are shown below:

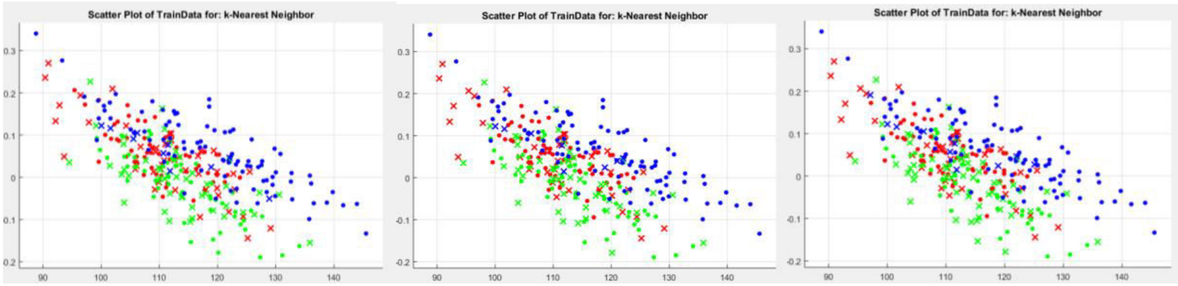
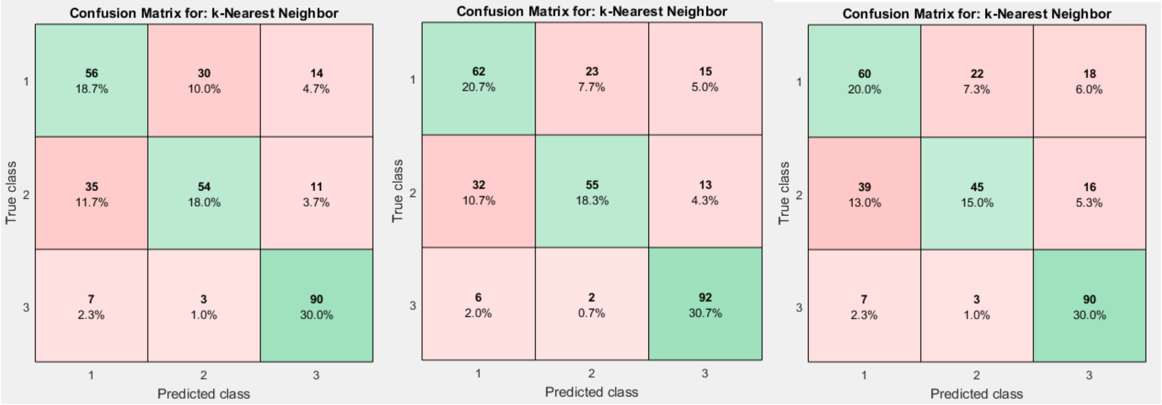


Figure5 KNN Scatter Plot of Trained Data (a) Fine KNN (b) Medium KNN (c) Cubic KNN



Confusion Matrix for: k-Nearest Neighbor

True class \ Predicted class	1	2	3
1	56 18.7%	30 10.0%	14 4.7%
2	35 11.7%	54 18.0%	11 3.7%
3	7 2.3%	3 1.0%	90 30.0%

Confusion Matrix for: k-Nearest Neighbor

True class \ Predicted class	1	2	3
1	62 20.7%	23 7.7%	15 5.0%
2	32 10.7%	55 18.3%	13 4.3%
3	6 2.0%	2 0.7%	92 30.7%

Confusion Matrix for: k-Nearest Neighbor

True class \ Predicted class	1	2	3
1	60 20.0%	22 7.3%	18 6.0%
2	39 13.0%	45 15.0%	16 5.3%
3	7 2.3%	3 1.0%	90 30.0%

Figure6 KNN Confusion Matrix (a) Fine KNN (b) Medium KNN (c) Cubic KNN

The overall prediction accuracy and error percentages of the trained SVM and KNN models are tabulated below:

Table 1 Overall Accuracy and Error results of SVM and KNN trained models on the training set.

Classifier Type	**SVM**			**KNN**		
	Linear	**Cubic**	**Quadratic**	**Fine**	**Medium**	**Cubic**
Total No. of Instances	300	300	300	300	300	300
Overall Accuracy (%)	74.7	65.7	70.7	66.7	69.7	65.0
Overall Error (%)	25.3	34.3	29.3	33.3	30.3	35.0

3.1 Linear SVM Classifier Application

A GUI is developed to analyze the input image, extract features, and to classify the results using Linear SVM trained classifier model as Normal, Viral Pneumonia or COVID-19 patients. To analyze and classify the test images, follow the below-mentioned steps.

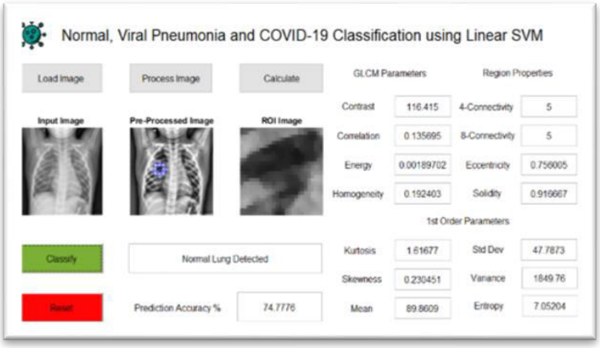
Step1: Load the input test image.

Step2: Apply Pre-processing operations (Image Filtering and Enhancement).

Step3: Cropping region of 25 x 25 pixels.

Step4: Feature Extraction (First Order, GLCM, and Region Properties).

Step5: Classification using Linear SVM trained model.


Normal, Viral Pneumonia and COVID-19 Classification using Linear SVM

GLCM Parameters
Contrast 116.415
Correlation 0.135695
Energy 0.00189702
Homogeneity 0.192403

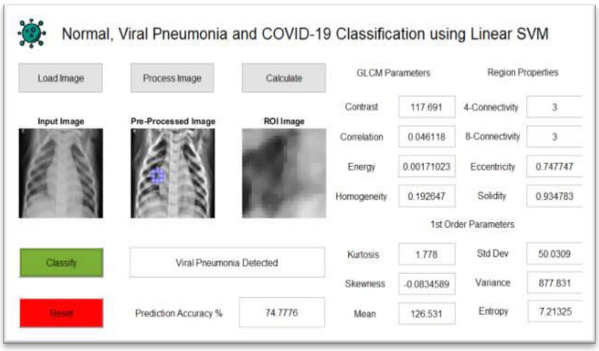
Region Properties
4-Connectivity 5
8-Connectivity 5
Eccentricity 0.756005
Solidity 0.916667

1st Order Parameters
Kurtosis 1.61677
Skewness 0.230451
Mean 89.8609
Std Dev 47.7873
Variance 1849.76
Entropy 7.05204

Normal Lung Detected

Prediction Accuracy % 74.7776

Figure7 Input Image Classified as Normal Lung


Normal, Viral Pneumonia and COVID-19 Classification using Linear SVM

GLCM Parameters
Contrast 117.691
Correlation 0.046118
Energy 0.00171023
Homogeneity 0.192647

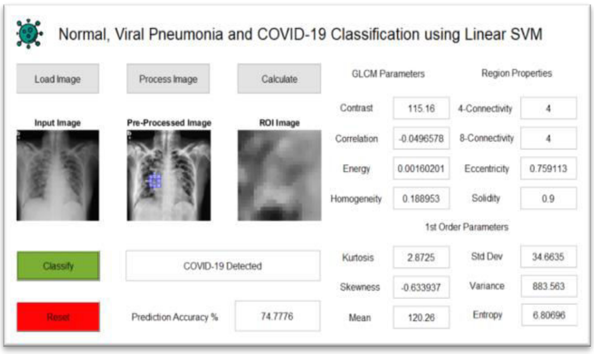
Region Properties
4-Connectivity 3
8-Connectivity 3
Eccentricity 0.747747
Solidity 0.934783

1st Order Parameters
Kurtosis 1.778
Skewness -0.0834589
Mean 126.531
Std Dev 50.0309
Variance 877.831
Entropy 7.21325

Viral Pneumonia Detected

Prediction Accuracy % 74.7776

Figure8 Input Image Classified as Viral Pneumonia


Normal, Viral Pneumonia and COVID-19 Classification using Linear SVM

GLCM Parameters
Contrast 115.16
Correlation -0.0496578
Energy 0.00160201
Homogeneity 0.188953

Region Properties
4-Connectivity 4
8-Connectivity 4
Eccentricity 0.759113
Solidity 0.9

1st Order Parameters
Kurtosis 2.8725
Skewness -0.633937
Mean 120.26
Std Dev 34.6635
Variance 883.563
Entropy 6.80696

COVID-19 Detected

Prediction Accuracy % 74.7776

Figure9 Input Image Classified as COVID-19

4.DISCUSSION AND CONCLUSION

The coronavirus showed partial similar characteristics over viral pneumonia. Therefore, the rate of the virus spreading made the situation difficult to be under control. A training dataset comprising of 100 normal, 100 viral pneumonia and 100 COVID-19 cases were used to train the classifier models i.e. SVM and KNN with different kernel functions and distance metrics respectively. First-order statistical features: Mean, Standard Deviation, Skewness, Kurtosis, Entropy, and Variance, second-order statistical features: GLCM (Contrast, Correlation, Energy, and Homogeneity) and Region Properties: 4-Connectivity, 8-Connectivity, Eccentricity, and Solidity features were extracted and used as training parameters. These parameters were used to train six models; Linear SVM, Cubic SVM, Quadratic SVM, Fine KNN Medium KNN and Cubic KNN.

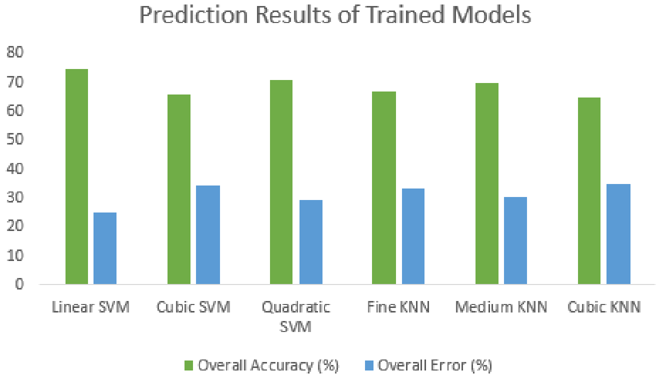


Figure10 shows the bar graph results of Predication Accuracy vs Error of Trained Models

From the above experiments, it is revealed that Support Vector Machines outperforms the k-Nearest Neighbor algorithm. Its performance is best in terms of prediction accuracy. KNN is less computationally intensive and the simplest algorithm than SVM, whereas the SVM takes plenty of time to measure all the distances and store all the training samples. Chest X-ray image results of Normal, Viral Pneumonia, and COVID-19 show fewer variations in the extracted features. Therefore, the best predication accuracy that could be achieved was 74.7% using Linear SVM trained model. This trained model was used in the implementation of the classifier application for the detection of normal, viral pneumonia, and COVID-19 patients. However, the proposed method should be trained and tested for more datasets and different features to increase prediction accuracy.

ACKNOWLEDGMENT

I would like to thank the MATLAB for the development environment and GitHub for providing the dataset of high resolution 1024 x 1024 pixels chest X-ray images (normal, viral pneumonia, and COVID-19). Also, the International Journal of Scientific Research in Engineering and Management for giving an opportunity to publish the practical research work.

REFERENCES

1. https://en.wikipedia.org/wiki/2019%E2%80%9320_coronavirus_pandemic.
2. https://arxiv.org/ftp/arxiv/papers/2003/2003.09424.pdf.
3. Oussema Zayane1, Besma Jouini1 and Mohamed Ali Mahjoub21, “Automatic lung segmentation method in CT images” published in Canadian Journal on Image Processing & Computer Vision Vol. 2, No. 8, December 2011.
4. Daniel Y. Chong, “Robustness-driven feature selection in classification of fibrotic interstitial lung disease patterns in computed tomography using 3D texture features”, 0278-0062 (c) 2015 IEEE.
5. Gehad Ismail Sayed; Mona Abdelbaset Ali; Tarek Gaber; Aboul Ella Hassanien;VaclavSnasel 2015 11th International Computer Engineering Conference (ICENCO) Year: 2015 Pages: 144 -149, DOI: 10.1109/ICENCO.2015.7416339.
6. Mark R Dension: Coronavirus Research: Keys to diagnosis, Treatment and Prevention of SARS.