ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 9, Issue 2, pp: (24-32), Month: October 2022 – March 2023, Available at: www.paperpublications.org

Covid-19 Diagnosis from X-Ray Images using Support Vector Machine

Satyendra Kumar Sagar

DOI: https://doi.org/10.5281/zenodo.7446436

Published Date: 16-December-2022

Abstract: **Coronavirus disease strike the world in 2019 and commonly called COVID-19 with its update given by the World Health Organization (WHO) on December 31, 2019. It infected more than 100 countries, an infectious disease strike the whole world and people of all age groups became a global health emergency. This disease can transmit from person to person through respiratory droplets and thus is highly contagious. The second wave almost killed billions of persons and lead to several liver problems, pneumonia, respiratory failure, cardiovascular diseases, etc. This can be symptomatic as well as asymptomatic in some patients and thus lead to increased communicability. Machine Learning is a latest trend currently useful in almost all research areas. Using these techniques to diagnose corona makes it highly feasible to cope up with this emergency. Different methods for testing corona virus are present but they require huge delay, are expensive, highly dependent test kits, higher negative false rate and prone to human errors. In this article we provide the state of the art of the covid diagnosis using Chest X ray images and this can guide both clinicians and technologists. A support vector machine is used to train the model and classify images into normal, pneumonia, and covid images. An overall accuracy of 95% is achieved using this method.**

Keywords: **COVID-19, Corona Virus, Machine Learning, Convolutional Neural Network, Support Vector Machine, X-ray images, Pneumonia.**

I. INTRODUCTION

The World Health Organization (WHO), confirmed a total of 124 billion above cases globally of covid as in March 2021. There are also 2 billion above deaths due to this disease [1]. The most common manifestations of this disease are cough, fever and difficulties in smelling and breathing. There are cases in which patients are asymptomatic and thus an early detection of the infected patients is very much important. The best testing method is reverse transcription-polymerase chain reaction (RRT-PCR) method. But it suffers from lower recall and result in more false negatives. Hence, chest CT scans and X-ray images are a preliminary used to test for covid. This disease is highly inter-person contagiousness and it is a form of induced pneumonia. This disease have symptoms like cough, cold, fever, but some severe symptoms like shortening in breath can even lead to death with major organ failures [2]. Some other tests leads to checking the organs and structures of the chest using X-rays also called as radiography. But the main concern is that most of the characteristics of COVID-19 are similar to other kinds of pneumonia. Therefore, most of the authors have used deep learning (DL) using CNN is a promising option for the automatic feature extraction.

This pandemic caused huge loss of manpower, economy due to lockdown and loss of productivity. This outbreak required huge efforts to develop the test kits such as reverse transcription polymerase chain reaction (RT-PCR) kits [3], antigen test or rapid test kits. However these tests take time, and requires huge resources and this is where Artificial intelligence or machine learning (ML) plays an essential role in covid case classification. These ML models can be used to predict infectious cases and recovery rates using chest x-ray, CT scans [4] or blood samples. In the first wave patients with severe symptoms were taken to hospital or intensive care unit. However the actual asymptomatic people (those who did not had symptoms) could not seek any medical assistance and remained undetected and uncounted. But they infected a huge population again. During the second wave many infectives along with the asymptomatic infectives were tested and got registered. The variants of covid are changing with time and their symptoms are also varying making it hard to detect.

Different ML models exists such as regression models, decision tree, support vector machine (SVM), clustering methods, and Convolutional neural network and many more [5]. In this pandemic the dataset is released by many organizations for public support, hence using this technique was possible.

Using ML models can help in pre screening of patients before test, non contact methods for diagnosis and low cost solution. Moreover, successful screening of contaminated patients, is subject to doctors error (human error) and also the current test have higher negative false reports around 15-20% [6]. Moreover, RT-PCR test have low sensitivity. Thus ML model will be used in this study like SVM, since SVM requires less time in training than CNN. Therefore we have used an SVM model.

The paper is structured as follows: structure of the given paper is as follows: section 2 provides a brief literature review on role of ML in Covid-19. Section 3 provides the methodology followed by section 4 providing the background of the SVM ML model. The description of the dataset and data preprocessing is illustrated in section 5. The results and discussion associated with work can be obtained from section 6. Section 7 elaborates the conclusion and the future work in this domain.

II. LITERATURE REVIEW

In this section we present the current state of the art of techniques of artificial intelligence (machine learning) used in Covid-19. This disease is causing a global health crisis.

The key symptom in this disease is cough and thus in [7] authors recorded coughs through smartphone of the healthy (COVID-19 negative) and COVID-19 positive persons. They found that covid positive coughs samples are 15%–20% shorter than non-covid coughs. Using six different ML techniques with leave-*p*-out cross-validation scheme to train and evaluate, best results were obtained using the Resnet50 classifier. Computer Tomography (CT) image are used to detect of level of Corona Virus Disease, using a deep classification based on convolution and deconvolution local enhancement [8]. These operation, enhances the contrast between local lesion region and abdominal cavity of COVID-19 and obtain middle level features. But it can effectively determine whether the feature vector in each feature channel contains the image features of COVID-19. Some authors [9] used super resolution CT scans images into very deep, super-residual neural networks to enhance lung CT scan efficiency. They used existing pre-trained models to decrease the training time. The hand crafted feature extraction from the CT scans can also be used such as Grey Level Co-occurrence Matrix (GLCM), Local Directional Pattern (LDP), Grey Level Run Length Matrix (GLRLM), etc. They formed the dataset by making patches (16x16, 32x32, 48x48 and 64x64) of complete images [10] and extracted feature from patches, increasing the overall accuracy using SVM. GLSZM method gave the best accuracy of 99.68%. SVM is popularly used by many authors thus it seeks attention in our application.

In [11] authors proposed Scat-NET which is 25 layered CNN model integrated with scattergram images. They proposed that the Scat-NET model can be arranged at the CT scan test. However automatic detection of disease (COVID-19) from CT scans are subject to large datasets, ambiguity in the characteristics and the model accuracy or recall. Hence authors in [12] propose a method to diagnose CT scans with high recall and accuracy. Many a times recall is better measure than precision thus they explore a trade-off among them. They used the concept of pre-trained ML models i.e. transfer learning approach. The proposed stacked ensemble use four CNN models: VGG-19, ResNet-101, DenseNet-169 and WideResNet-50-2.

Another work that used digital chest x-ray radiographs to automatically detect COVID-19 pneumonia patients using Deep CNN [13]. The Deep CNN model is Inception V3 with transfer learning gave an overall accuracy of 98% and above.

Federated learning is distributed learning, in [14], authors integrated it for covid detection. They identified the factors affecting model accuracy and loss like activation function, model optimizer, learning rate, number of rounds, and data size during the model training stage. They found that softmax activation function and SGD optimizer give better prediction accuracy and loss.

X-ray images are another diagnosis method for determining the covid in the lungs. It is difficult to classify between chest X-ray images of common Pneumonia, Covid positive, and healthy lungs and thus [15] used a classifier ensemble technique. They utilized Choquet fuzzy integral to increase accuracy of individual classifier. They used transfer learning approach to train the base CNN classifiers (with two dense layers and one softmax layer) using InceptionV3, DenseNet121, and VGG19. Only few authors have evaluated the covid diagnosis from X-ray images and thus in this work we used X-ray images for disease identification.

Currently, Reverse transcription polymerase chain reaction (rRT-PCR) is standard test for covid-19. But it requires 3-4 hours to generate results and higher false-negative rates (15-20%). Moreover these test require certified laboratories, costly equipment and trained persons to test the patients. Thus the authors used two ML models to classify hematochemical values from routine blood exams [16]. They discriminated between covid positive or not based on the clinical interpretations of blood tests samples. Another work [17] that used blood samples first extracted eleven important features/indices from blood using random forest algorithm. People wearing masks in the covid pandemic cannot be identified easily thus using ML models peoples with face mask can be identified. Authors in [18] used a hybrid ML model for face mask detection with two components. They first extracted features from Resnet50 model and then classified face masks using decision trees, Support Vector Machine (SVM), and ensemble algorithm.

In some cases the textual clinical reports can be classified into four classes. This can be achieved using ML algorithms with feature engineering techniques like Term frequency/inverse document frequency (TF/IDF), Bag of words (BOW) and report length [19]. ML classifiers are implemented on these features like Logistic regression and Multinomial Naive Bayes classifiers.

In another work [20], the identification of the severity of covid can be done to facilitate the risk estimation. 32 highly associated features to detect covid-19 severeness. However further, inter-feature redundancies among the 32 features was identified and finally selected 28 features were used to train the model. They achieved an overall accuracy of 81.48%.

Some authors use a software tool comprised of unsupervised Latent Dirichlet Allocation (LDA) and other ML methods to analyze the Twitter data in Arabic. They aim to detect government pandemic measures and public concerns during the COVID-19 pandemic [21].

In this work we use X-ray images with the support vector machine.

III. PROPOSED METHODOLOGY

In this section the proposed methodology is explained in the detailed way. The proposed methodology is shown in the flow chart representation in the figure 1. The important blocks in the flow chart are explained below:

- Data Acquisition: First the chest X-ray images are acquired from the dataset source [22]. The details of the dataset are provided in the later section.

- Grid Search for hyperparameters : The optimum parameters are obtained for the grid search such that best parameters could be obtained to generate best results. The hyperparameters for radial basis function kernel ‘c’ and ‘gamma’ in SVM were searched in the range, c: [0.1, 10, 1000] and gamma : [1, 0.01, 0.0001].

- SVM training: The complete dataset is divided into training and testing dataset. We evaluated the accuracy with different train:test ratios. For example, 75:25 means, 75 percent samples (images here) were used for training and 25 percent for testing.

- Performance parameters: Accuracy and confusion matrix are used to evaluate the performance of SVM model.

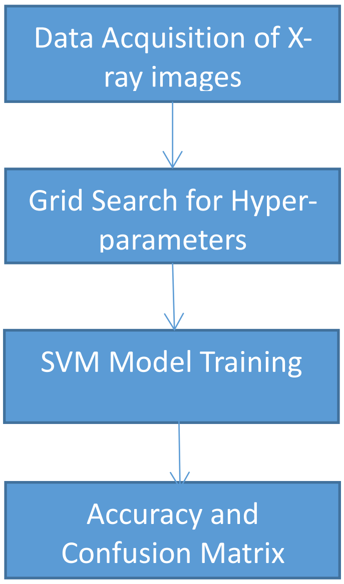

Data Acquisition of X-ray images
Grid Search for Hyper-parameters
SVM Model Training
Accuracy and Confusion Matrix

Fig. 1 Proposed methodology for the image classification.

IV. SUPPORT VECTOR MACHINE MODEL

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms that can perform two or more class classification [23]. In comparison to latest algorithms like DNN, CNN etc. they have two main advantages: higher speed and better performance with a limited number of samples (in the thousands). Thus in this paper SVM algorithm [24] is used for image classification problem. The support vector machine does not considers the complete dataset instead they consider only some of the support data points and these data points are used to design a hyperplane (assume a line in 2D) that best separates the tags. The best hyperplane is chosen such that maximization is achieved in the margins from both tags or a line whose distance to the support vector of each category is the largest. Figure 2 demonstrates the best hyperplane in 2 D.

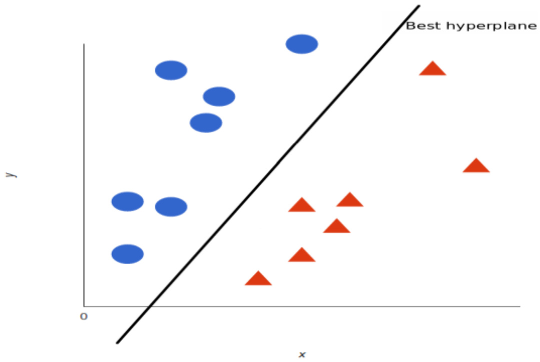


Fig. 2. A SVM hyperplane in 2-D (line) for two class problem.

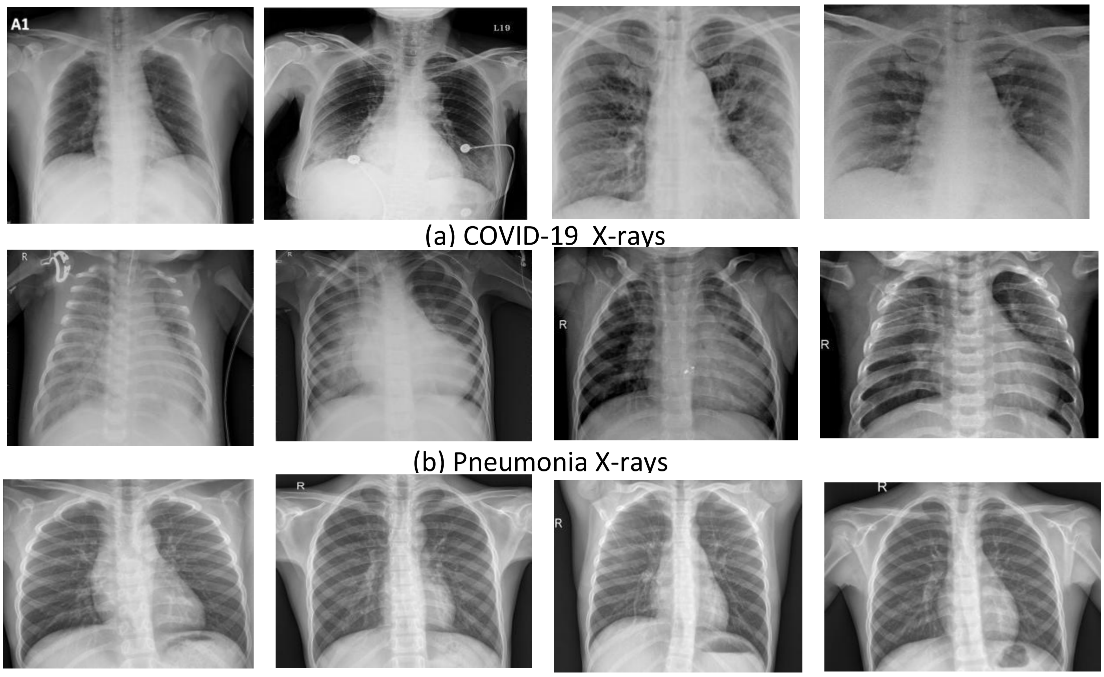
It can also be employed for regression purposes. This algorithm is based on the idea of finding a hyperplane. This plane best divides a dataset into two classes, as shown in the figure 2. Support vectors are the data points nearest to the hyperplane, that decides the decision of best hyperplane. If the Support vectors data set removed, would alter the position of the dividing hyperplane. Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. Thus our data points needs to be as far away from the hyperplane as possible such that remains the correct side of it each class.

Advantages of SVM are: Accuracy, working with smaller cleaner datasets and more efficient because it uses a subset of training points. Disadvantages of SVM are: not suitable for larger datasets as the training time with SVMs can be high and less effective on noisier datasets with overlapping classes.

V. DESCRIPTION OF DATASET

The dataset has been downloaded from kaggle.com [25]. This dataset comprise of total of two folders one for training and other for testing. These X-ray images are further classified in both training and testing folder with the name of the folders as covid-19, pneumonia and normal chest X-ray images. A total of 460 images of Covid-19 patients, 1266 normal chest X-rays and 3418 pneumonia images are present in the training folder. So in total 5000 above images are present for training.

Figure 3 shows the images of all the three classification types such as covid X-rays, pneumonia and normal X-rays images.



(a) COVID-19 X-rays

(b) Pneumonia X-rays

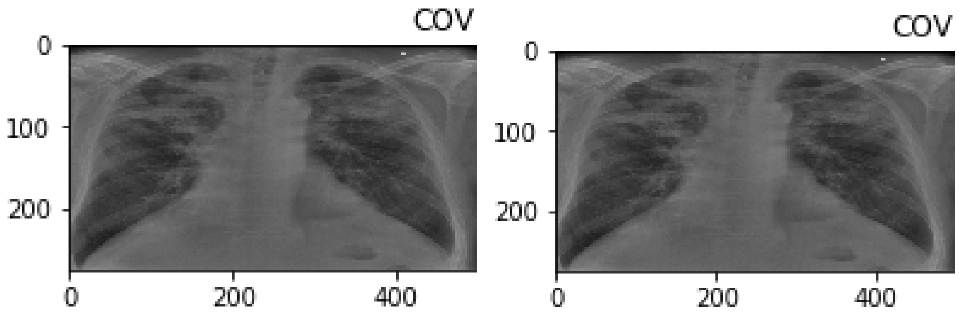
(c) Normal Patients X-ray

Fig. 3. X-ray Images Dataset of three classes as (a) COVID-19 , (b) Pneumonia and © Normal X-rays.

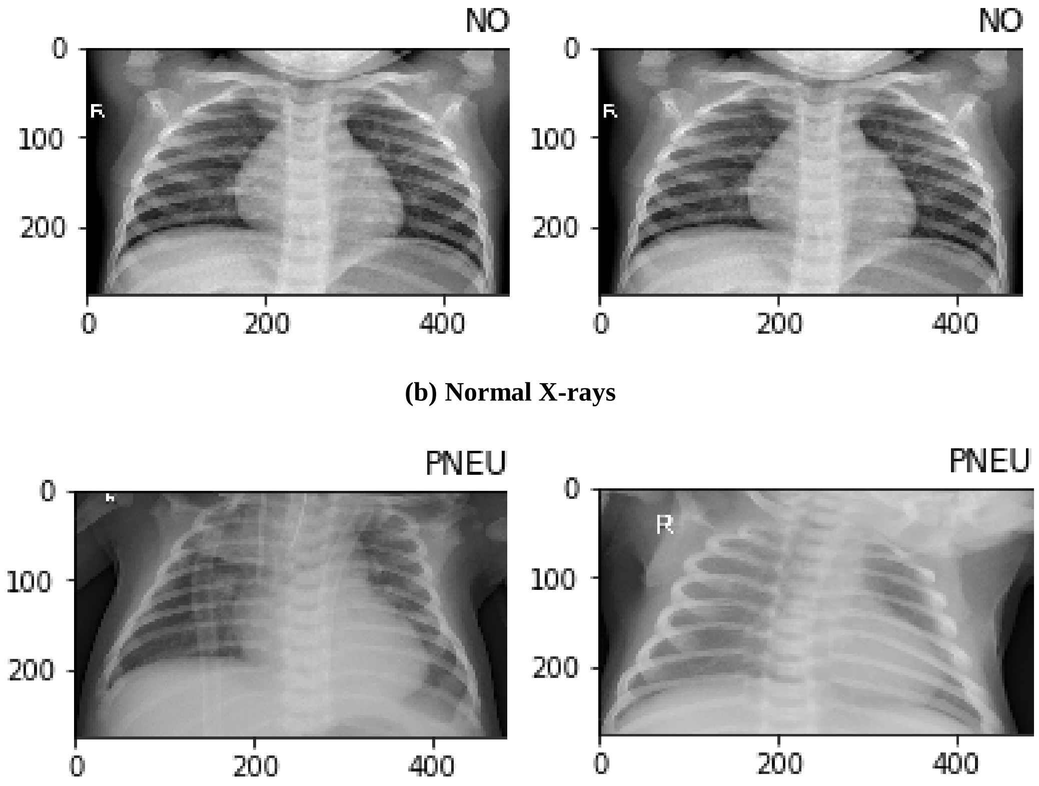
As the first step in preprocessing the images are resized to 300x300 size. These images are RGB images and thus three colour channels in total. These are converted to a 2D data and thus images are converted to grey scale. The grey images are flattened into array with each pixel as a feature and thus obtained matrix is used to feed the SVM model for training.

VI. RESULTS AND DISCUSSION

The model was trained using the SVM classifier and the reconstructed images in grey scale are shown below in figure 4 with their predicted classes. The images were resized for demonstration with their labels both actual and predicted. The radial basis kernal is used for training the classifier. The Covid prediction is shown using COV, Normal with NO and Pneumonia with PNEU.



(a) COVID-19 X-rays



(b) Normal X-rays

(c) Pneumonia X-rays

Fig. 4: Results for X-ray image classification using SVM.

The model training was performed for different training and testing samples. The precision recall and F1-scores of the different test:train splits in shown in table1. Class 1 belongs to COVID-19 X-rays, Class 2 belongs to Pneumonia X-rays and Class 3 belongs to Normal X-rays.

Table 1: The evaluation parameters for different training and testing size.

Train: Test	**Precision**			**Recall**			**F1-score**		
	Class			**Class**			**Class**		
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
70:30	0.94	0.89	**0.95**	0.88	0.90	**0.95**	0.91	0.89	**0.95**
75:25	0.89	0.87	**0.95**	0.91	0.88	**0.94**	0.90	0.88	**0.95**
98:2	**1**	0.83	0.99	**1**	0.95	0.95	**1**	0.89	0.97

The precision, Recall and F1-Score is the best (that is 100%) for class 1 using 98% samples for training. It means all the covid images are classified correctly. However when using 70:30 and 75:25 split ratios, the precision, recall and F1-score is maximum for Class 3.

Accuracy is another important performance parameter and hence we have tabulated macro average and weighted average for eacg classes at different train:test splits in table 2.

Table 2: The Accuracy parameters for different training and testing size.

Train: Test	**Macro Average Accuracy**			**Weighted Average Accuracy**		
	Class			**Class**		
	1	**2**	**3**	**1**	**2**	**3**
70:30	**0.93**	0.91	0.92	0.93	0.93	0.93
75:25	0.90	**0.91**	**0.91**	0.93	0.93	0.93
98:2	0.94	**0.97**	0.95	0.96	0.95	0.95

The average accuracy of class 1 is highest for a 70:30 split and minimum for class 2. However the best results are obtained using 98:2 split ratio i.e., 94% for class 1, 97% for class 2, and 95% for class 3.

Table 3: Confusion matrix for test train split of 70:30.

Predicted/True	Class 1	Class 2	Class 3
Class 1	135	5	13
Class 2	1	344	16
Class 3	8	39	963

Table 4: Confusion matrix for test train split of 75:25.

Predicted/True	Class 1	Class 2	Class 3
Class 1	99	1	9
Class 2	3	276	33
Class 3	9	40	816

Table 5: Confusion matrix for test train split of 98:2.

Predicted/True	Class 1	Class 2	Class 3
Class 1	9	0	0
Class 2	0	20	1
Class 3	0	4	69

Table 3, 4 and 5 shows the confusion matrix of train:test split of 70:30, 75:25 and 98:2 respectively. In can be seen from each table that the diagonal elements are highest in number which shows the correct classified samples.

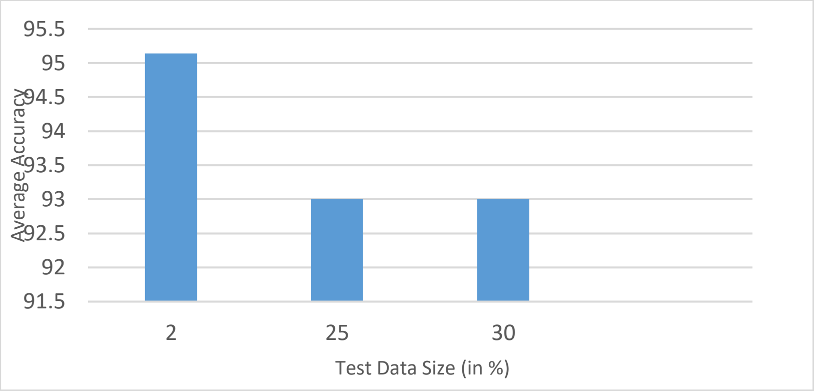


Fig. 5: Test Data Size versus the Average Accuracy for radial kernel.

The non diagonal elements should be ideolly zero but this scenario these are relatively less in number. With the train: test split of 98:2, their is 100 percent correct classification of class 1 i.e. covid X-rays however some miss classification occur for other classes.

The impact on accuracy of change in the training and testing dataset is shown the figure 5. The x-axis is the test data size in percent and the y-axis is the average accuracy using the radial basis kernel. Clearly there is a decrease in the accuracy with an increase in the test data samples. This is so because of reduction in training samples, thus it is optimum to chose 75:25 train test split .

VII. CONCLUSION AND FUTURE SCOPE:

COVID-19 has rapid global health concern because of its risk to human lifes. However the current state of testing lack in fast inference or test reports and it is higher in cost. ML plays a vital role in early diagnosis and timely treatment of covid. It is almost no cost testing method and also require no special training of the personals. This technology advancement of AI and ML has influenced every field of life even the medical field and shown promising results in health care. The decision making and diagnosis of covid can be easily achieved using CT scan images, X-ray images and cough sounds by analysing these data. These tools can carry out preliminary assessment of suspected patients and help them to get timely treatment and quarantine suggestion. This work focused on using chest X-ray images for COVID identification. CT scan images were already popular but less work is done using X-ray images. Moreover, Supervised learning show better outcomes than than Unsupervised learning methods. Therefore, SVM is a popular ML technique used almost in all the domains, such as fault detection, disease classification, data mining, credit risk etc. It is therefore preferred to use this since it is based only on the support vectors for training.

We achieved highest accuracy of 95% using the given dataset of 5000 above images.

However accuracy can be further improved by using large dataset. Also since the covid phase is changing with the varients, different symptoms are arriving and thus recurrent supervised learning can be better used to achieve superior accuracy. We can also Convolutional neural network with SVM using the transfer learning, federated learning and incremental learning approaches. In the future, better acceleration methods can be deployed to run these models on low resource devices.

REFERENCES

[1] Coronavirus update (live): 28,988,031 cases and 925,320 deaths from COVID-19 virus pandemic - worldometer.

[2] Sudre, Carole H., et al. "Attributes and predictors of Long-COVID: analysis of COVID cases and their symptoms collected by the Covid Symptoms Study App." *Medrxiv* (2020).

[3] Banaganapalli, Babajan, et al. "Multilevel Systems Biology Analysis of Lung Transcriptomics Data Identifies key miRNAs and Potential miRNA Target Genes for SARS-CoV-2 Infection." *Computers in Biology and Medicine* (2021): 104570.

[4] Roberts, Michael, et al. "Machine learning for COVID-19 detection and prognostication using chest radiographs and CT scans: a systematic methodological review." *arXiv preprint arXiv:2008.06388* (2020)

[5] Nabavi, Shahabedin, et al. "Medical Imaging and Computational Image Analysis in COVID-19 Diagnosis: A Review." *Computers in Biology and Medicine* (2021): 104605.

[6] Kwekha-Rashid, Ameer Sardar, Heamn N. Abduljabbar, and Bilal Alhayani. "Coronavirus disease (COVID-19) cases analysis using machine-learning applications." *Applied Nanoscience* (2021): 1-13.

[7] Pahar, Madhurananda & Klopper, Marisa & Warren, Robin & Niesler, Thomas. (2021). COVID-19 Cough Classification using Machine Learning and Global Smartphone Recordings. Computers in Biology and Medicine. 135. 104572. 10.1016/j.compbiomed.2021.104572.

[8] Fang, Lingling, and Xin Wang. "COVID-19 deep classification network based on convolution and deconvolution local enhancement." *Computers in Biology and Medicine* (2021): 104588.

[9] Arora, Vinay, et al. "Transfer learning-based approach for detecting COVID-19 ailment in lung CT scan." *Computers in biology and medicine* (2021): 104575.

[10] Barstugan, Mucahid, Umut Ozkaya, and Saban Ozturk. "Coronavirus (covid-19) classification using ct images by machine learning methods." *arXiv preprint arXiv:2003.09424* (2020).

[11] Tuncer, Seda Arslan, et al. "Scat-NET: COVID-19 Diagnosis with a CNN Model using Scattergram Images." *Computers in Biology and Medicine* (2021): 104579.

[12] Jangam, Ebenezer, and Chandra Sekhara Rao Annavarapu. "A stacked ensemble for the detection of COVID-19 with high recall and accuracy." *Computers in Biology and Medicine* 135 (2021): 104608.

[13] Asif, Sohaib, et al. "Classification of COVID-19 from Chest X-ray images using Deep Convolutional Neural Network." *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*. IEEE, 2020.

[14] Abdul Salam, Mustafa, Sanaa Taha, and Mohamed Ramadan. "COVID-19 detection using federated machine learning." *Plos one* 16.6 (2021): e0252573.

[15] Dey, Subhrajit, et al. "Choquet Fuzzy Integral-based Classifier Ensemble Technique for COVID-19 Detection." *Computers in Biology and Medicine* (2021): 104585.

[16] Brinati, Davide, et al. "Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study." *Journal of medical systems* 44.8 (2020): 1-12.

[17] Wu, Jiangpeng, et al. "Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results." *MedRxiv* (2020).

[18] Loey, Mohamed, et al. "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic." *Measurement* 167 (2021): 108288.

[19] Khanday, Akib Mohi Ud Din, et al. "Machine learning based approaches for detecting COVID-19 using clinical text data." *International Journal of Information Technology* 12.3 (2020): 731-739.

[20] Yao, Haochen, et al. "Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests." *Frontiers in cell and developmental biology* 8 (2020): 683. https://www.kaggle.com

[21] Durgesh, K. SRIVASTAVA, and B. Lekha. "Data classification using support vector machine." *Journal of theoretical and applied information technology* 12.1 (2010): 1-7.

[22] Widodo, Achmad, and Bo-Suk Yang. "Support vector machine in machine condition monitoring and fault diagnosis." *Mechanical systems and signal processing* 21.6 (2007): 2560-2574.

[23] https://www.kaggle.com/prashant268/chest-xray-covid19-pneumonia