International Journal of Advanced Engineering: Artificial Intelligence, Robotics and Automation
ISSN XXXX-XXXX
Vol.:1 Issue: 1
Published online on: Dec 2024
PP. 59- 68

Covid-19 Diagnosis from Chest X-Ray Images using Transfer Learning and Data Augmentation

Sandeep K Tiwari[1], Shashi Pratap Tomar[2], Shashank Swami[3], Surjeet Singh Parihar[4]
[1,2,3,4]Vikrant University, Gwalior, India

Abstract: The World Health Organisation (WHO) provided an update on the COVID-19 coronavirus illness on December 31, 2019. COVID-19 is a coronavirus disease that spread around the world in 2019. An viral sickness that affected people of all ages spread to more than 100 nations, causing a worldwide health disaster. This illness is extremely infectious because it may spread from person to person by respiratory droplets. The second wave caused numerous liver issues, pneumonia, respiratory failure, cardiovascular disorders, etc. and almost killed billions of people. This may be both symptomatic and asymptomatic in certain people, increasing their communicability as a result. A recent development that is beneficial in practically all study fields is machine learning. It is quite possible to handle this situation by using these methods to diagnose corona. There are several ways to test for the corona virus, but they are all time-consuming, costly, heavily reliant on test kits, more likely to result in false negatives, and subject to human error. The state-of-the-art of the covid diagnosis utilising chest X-ray pictures is presented in this article, which may serve as guidance for both doctors and technicians. Various machine learning (ML) and specifically Deep learning (DL) model are trained, which also classifies the images into normal, pneumonia, and covid images using a small dataset. The impact of Transfer learning and data augmentation is also studied in these schemes and using the best model, a 95% overall accuracy, 90% precision, and 90% F1 score is attained.

Keywords: COVID 19, Machine Learning, Deep Learning, Chest X-ray images, Covid Detection

I. Introduction:

As of March 2021, the World Health Organisation (WHO) reported 124 billion above instances of covid worldwide. Furthermore, this disease causes 2 billion additional deaths [1]. Cough, fever, and breathing and smelling problems are the most typical symptoms of this illness. There are situations where individuals are asymptomatic, hence it is crucial to identify infected patients as soon as possible. Reverse transcription-polymerase chain reaction (RRT-PCR) testing is the most effective technique. However, it has a lower recall and produces more false negatives. Consequently, X-ray and chest CT scans are first utilised to check for covid. This illness, which is an induced form of pneumonia, is extremely communicable between people.

Although this illness has common symptoms including a cough, cold, and fever, certain more severe ones, like breathing difficulties, can even cause significant organ failures and death [2]. In several additional procedures, radiography, often known as X-rays, is used to examine the organs and structures of the chest. The majority of COVID-19's traits, however, are comparable to those of other types of pneumonia, which is the main cause for concern. As a result, CNN-based deep learning (DL) is a promising approach for automatic feature extraction that has been adopted by the majority of authors.

This pandemic has resulted in massive job losses, economic losses due to downtime, and reduced productivity. This epidemic required major efforts to develop test kits such as reverse transcription polymerase chain reaction (RT-PCR) [3] kits, rapid test kits for antigen testers. However, these tests take time and require huge resources, and this is where artificial intelligence or machine learning (ML) plays an important role in classifying covid cases. These ML models can be used to predict infection rates and recovery rates using chest X-rays, CT scans [4] or blood samples. In the first wave, patients with severe symptoms were admitted to the hospital or intensive care unit. But actual asymptomatic people (those without symptoms) could not seek medical care and went unnoticed and unaccounted for. But they are once again infecting a huge population. During the second wave, many infectious agents and asymptomatic infections were tested and recorded. Covid variants change over time and their symptoms also vary, making detection difficult.

Corresponding Author’s E-mail ID:sandeep72128@gmail.com

There are various ML and Deep learning models such as Random forest, decision tree, bayesian network, support vector machine (SVM), convolutional neural networks and many others [5]. During this pandemic, many organizations are handing over their data sets for public support, so the use of this technology was possible.

The pre-screening of patients before to testing, non-contact approaches for diagnosis, and low-cost solutions may all be aided by the use of ML and deep learning models. Additionally, a doctor's error (human error) can affect how well infected individuals are screened, and the present test has greater negative false-positive rates of about 15-20% [6]. Additionally, RT-PCR tests are not very sensitive. Some ML technique needs less time to train than CNN, but CNNs based DL models have better performance that ML models and thus be the DL models are employed in this investigation. Data availability is always an issue, this article therefore explored Transfer learning and data augmentation technique. Performance on pretrained models is compared by training them on Covid X-ray image data.

The presented paper has the following structure: A brief literature review on the function of ML and DL in Covid-19 is included in section 2. The approach is presented in Section 3, and the backdrop of the CNN model and VGG16 and VGG19 model is presented in Section 4. Section 5 provides an illustration of the dataset's description and data preparation. Section 6 contains the work's conclusions and related discussion. The findings and the planned activities in this area are elaborated in Section 7.

II. Literature Review:

We outline the state-of-the-art artificial intelligence (machine learning) approaches employed in Covid-19 in this section. A health disaster is being brought on by this illness.

Since coughing is a primary sign of this illness, [7]authors used smart phones to record the coughs of healthy (COVID-19 negative) and COVID-19 positive individuals. They discovered that samples of coughs with covid positivity are 15%–20% shorter than samples of coughs without covid. The Resnet50 classifier produced the best results when used with a leave-p-out cross-validation strategy and six distinct ML algorithms. To determine the severity of the Corona Virus Disease, Computer Tomography (CT) images are employed, along with a deep classification method based on convolution and deconvolution local enhancement [8]. These procedures increase the contrast between the area of localised lesions and the abdomen and get middle-level characteristics from the COVID-19 cavity. However, it can successfully detect whether each feature channel's feature vector contains COVID-19 image features. To increase the effectiveness of lung CT scans, some writers [9] fed pictures from super-resolution CT scans into extremely deep, super-residual neural networks. To cut down on training time, they employed pre-trained models that were already in use. Grey Level Co-occurrence Matrix (GLCM), Local Directional Pattern (LDP), Grey Level Run Length Matrix (GLRLM), and other manually produced feature extractions from CT images can also be employed. They created the dataset by patching together full photos into grids of 16x16, 32x32, 48x48, and 64x64 pixels[10]; they then retrieved feature information from the patches and improved overall accuracy using SVM. The GLSZM technique had a 99.68% accuracy rating. SVM is widely utilised by many authors, hence our programme tries to draw attention to it.

Scat-NET, a 25 layered CNN model merged with scatter gram pictures, was suggested by the authors in [11]. They suggested setting up the Scat-NET model during the CT scan test. However, the COVID-19 system for automated illness identification from CT scans is limited by huge datasets, ambiguity in the features, and model recall or accuracy. As a result, the authors of [12] suggest a technique for accurately and recall ably diagnosing CT images. They investigate trade-offs between them since recall is frequently a better measure than accuracy. They made use of the transfer learning technique, or pre-trained machine learning models. Four CNN models are used in the proposed stacked ensemble: VGG-19, ResNet-101, DenseNet-169, and WideResNet-50-2.

Another study that using Deep CNN to automatically identify COVID-19 pneumonia patients from digital chest x-ray radiographs [13]. The Inception V3 Deep CNN model with transfer learning produced an overall accuracy of at least 98%. Federated learning, also known as distributed learning, was incorporated for covid identification in [14]. During the model training phase, they determined the variables such activation function, model optimizer, learning rate, number of rounds, and data size impacting model accuracy and loss. They discovered that SGD optimizer and softmax activation function provide superior prediction accuracy and loss.

Another approach for diagnosing lung cancer is by using X-ray imaging. The classifier ensemble approach was adopted by [15] because it is challenging to distinguish between chest X-ray pictures of common pneumonia, Covid positive, and healthy lungs. They used the Choquet fuzzy integral to boost each classifier's accuracy. InceptionV3, DenseNet121, and VGG19 were utilised to train the baseCNN classifiers (which include two dense layers and one softmax layer). Since very few authors have examined the covid diagnosis using X-ray pictures, we employed X-ray images in this study to identify the diseases.

Currently, the covid-19 standard test is the reverse transcription polymerase chain reaction (rRT-PCR). However, it takes 3–4 hours to get findings, and false-negative rates (15–20%) are greater. Additionally, these tests call for accredited labs, pricey apparatus, and skilled personnel to test the patients.In order to categorise hemato chemical results from regular blood examinations, the authors utilised two ML models [16]. Based on the clinical interpretations of the blood test samples, they distinguished between covid positive and not. Another study [17] that used blood samples

Corresponding Author's E-mail ID:sandeep72128@gmail.com

initially employed the random forest technique to identify eleven significant characteristics or indices from the blood. Since it is difficult to identify persons wearing masks during the COVID epidemic, using ML models, those wearing masks can be recognised. For face mask identification, authors in [18] utilised a hybrid ML model with two components. They started by extracting features from the Resnet50 model, and then they used ensemble algorithms, decision trees, and Support Vector Machines (SVM) to classify face masks.

Textual clinical reports may occasionally fall into one of four categories. This may be accomplished by combining machine learning (ML) algorithms with feature engineering approaches such as term frequency/inverse document frequency (TF/IDF), bag of words (BOW), and report length [19]. These characteristics are subject to ML classifier implementation, such as Logistic regression and Multinomial Naive Bayes classifiers.

The identification of the covid severity can be done in another study [20] to make the risk calculation process easier. 32 highly linked features to determine the severity of the COVID-19. To train the model, 28 features were ultimately chosen after inter-feature redundancies among the 32 features were discovered. A total accuracy of 81.48% was attained.To analyse the Arabic-language Twitter data, several academics employ software that combines unsupervised Latent Dirichlet Allocation (LDA) and other ML techniques. During the COVID-19 pandemic, they seek to identify governmental pandemic response strategies and public concerns [21]. These schemes does not focus on the issue of data set limitations and thus we propose a scheme that performs well even with less data.

In this study, we combine the transfer learning approach [22, 23] with data augmentation techniques [24, 25] for the classification of covid using patients X-ray pictures.

II. Proposed Methodology:

The proposed technique is thoroughly outlined in this section. The flowchart representation of the suggested technique is provided in figure 1. The flow chart's significant building blocks are described below:

- Data Gathering: The dataset source is used to first obtain the chest X-ray pictures [26]. The next section contains further information about the dataset.
- Hyper parameter: The hyper parameters are defined and set same each model to compare them. This allows comparison of all schemes under same settings and generation of the best results possible. In CNN hyper parameters are learning rate, optimizer, etc. Early stopping adopted with patience 5 for original data and 10 with augmented data.
- Data augmentation: Various data augmentation techniques are possible such as rotation, flipping, cropping, etc. Some of thsem are adopted in this study.
- CNN training: The complete dataset is divided into train and test folder. We evaluated the accuracy with given dataset with and without data augmentation.
- Performance metrics: Training accuracy, testing accuracy, training loss, testing loss and confusion matrix are used to assess the model's performance.

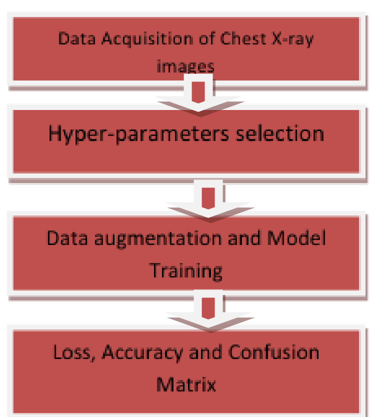


Fig. 1 Proposed methodology for the Chest X-ray image classification to detect COVID

IV. Deep Learning Models:

In the following study three models have been analyzed, namely a simple CNN, VGG16, and VGG19.

CNN model: Convolutional, pooling, and fully connected layers make up the three layers of a CNN, which is the most used model. The foundational component of the CNN is the convolution layer. It carries the majority of the computational burden on the network. This layer creates a dot product between two matrices, one of which is the kernel—a collection of learnable parameters—and the other of which is the constrained area of the receptive field. Compared to a picture, the kernel is smaller in space but deeper.

This indicates that the kernel height and width will be spatially tiny if the picture consists of three (RGB) channels, but the depth will go up to all three channels. By calculating an aggregate statistic from the surrounding outputs, the pooling layer substitutes for the network's output at certain points. This aids in shrinking the representation's spatial size, which lowers the amount of computation and weights needed. Each slice of the representation is subjected to the pooling procedure separately. All neurons in the layer that is fully linked are fully connected to all neurons in the layer below and above it. Because of this, it may be calculated using a matrix multiplication followed by a bias effect, as per normal. The representation between the input and the output is mapped using this layer.

- VGG16: A ConvNet is another name for a CNN, which is a type of neural network. One of the top computer vision models to date is the CNN variant known as VGG16. This model's developers analysed the networks and enhanced the depth using an architecture with incredibly tiny (3x3) convolution filters, which demonstrated a notable advancement over the state-of-the-art setups. The depth was increased to 16–19 weight layers, yielding around 138 trainable parameters. The "16" in VGG16 stands for 16 weighted layers. 13 convolutional layers, 5 Max Pooling layers, 3 Dense layers and a total of 21 layers make up VGG16, although only sixteen of them are weight layers, also known as learnable parameters 3. The input tensor size for VGG16 is 224, 244 and has three RGB channels.

- VGG19: VGG19 is a variation on the VGG model, and it has 19 layers total (16 convolutional, 3 fully connected, 5 maxpool, and 1 softmax). 19.6 billion FLOPs make up VGG19 and share same basic idea as the VGG16 model, with three extra convolutional layers.

V. Description of Dataset:

The dataset has been downloaded from kaggle.com . 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 111 images of Covid-19 patients, 70 normal chest X-rays and 70 pneumonia images are

present in the training folder. So in total 251 above images are present for training.

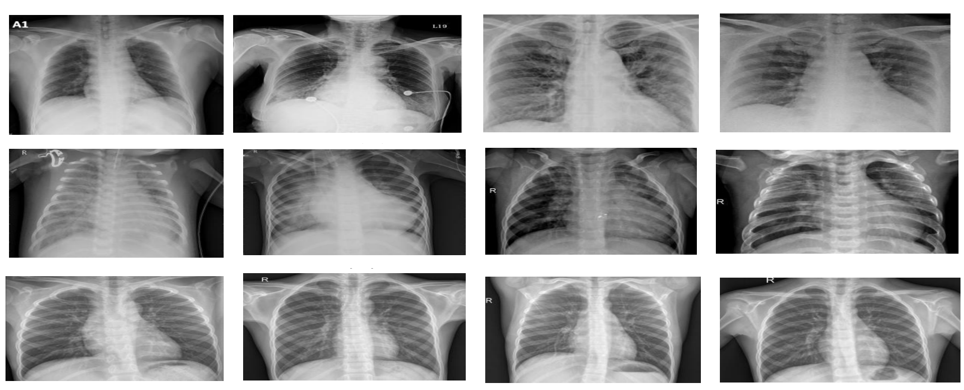


Fig. 2 shows the images of all the three classification types such as covid X-rays, pneumonia and normal X-rays images.

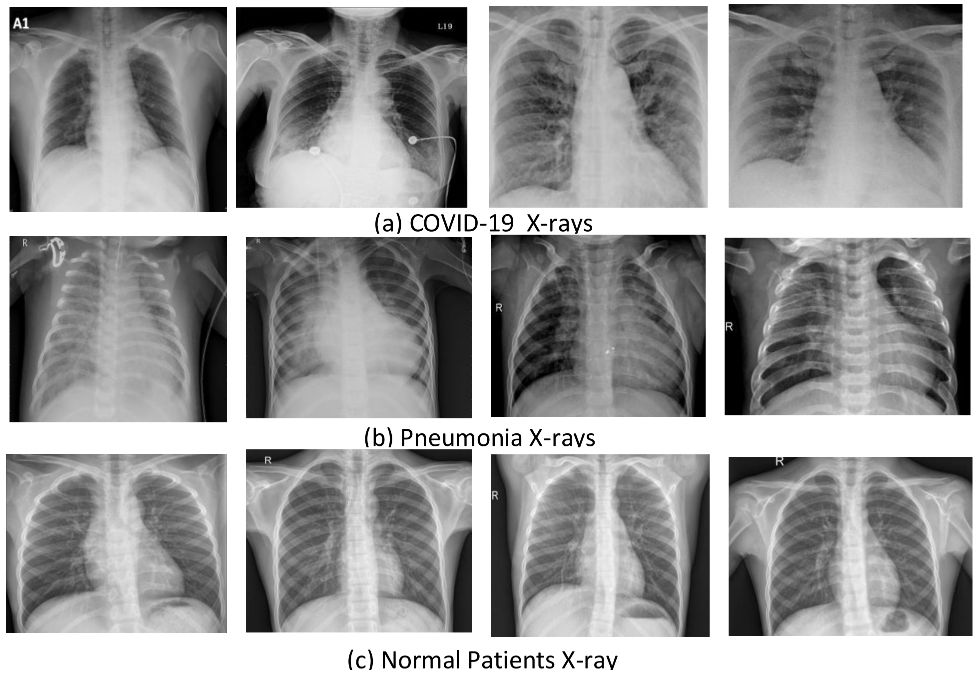
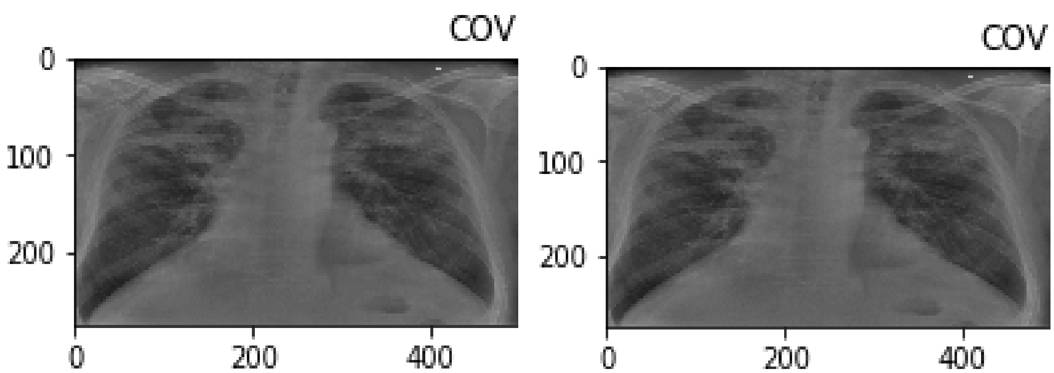


Fig. 3. X-ray Images Dataset of three classes as (a) COVID-19 , (b)Pneumonia and, (c) Normal X-rays.

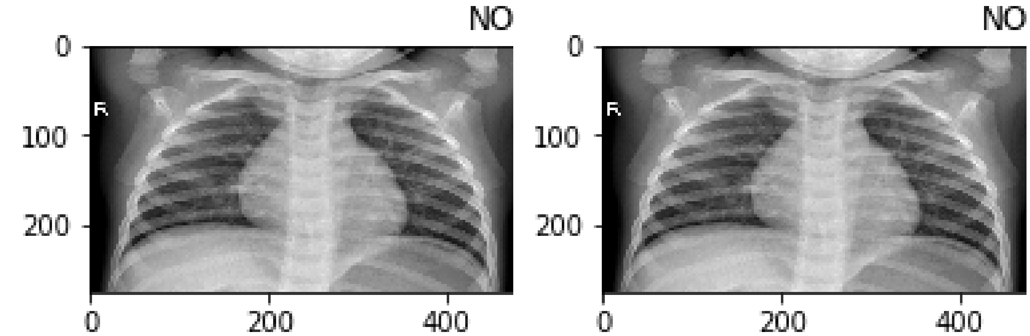
The images are initially downsized to 224x224 size as part of the preprocessing procedure. These pictures have three different colour channels since they are RGB photos.

VI. Results and Discussion:

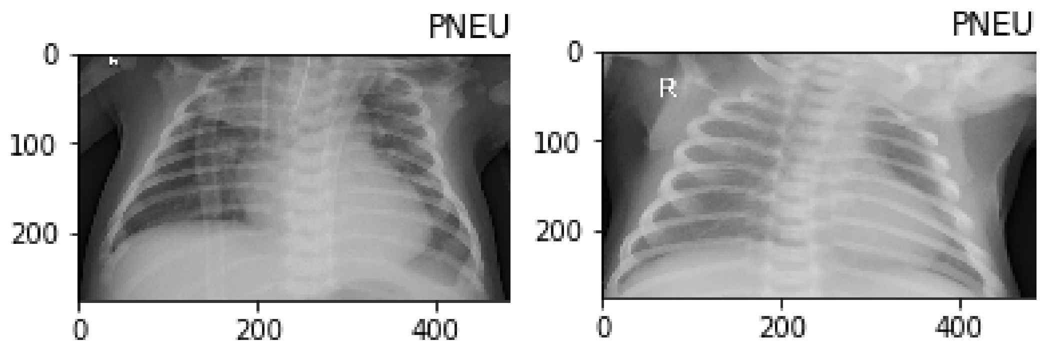
The CNN, VGG16, and VGG19 models trained on given data and the corresponding classes and predicted classes are displayed in Fig. 6 below. For demonstration purposes, the labels on the photos were both actual and predicted. Using COV the Covid prediction is displayed, Normal with NO, and Pneumonia with PNEU.



(a) COVID-19 X-rays



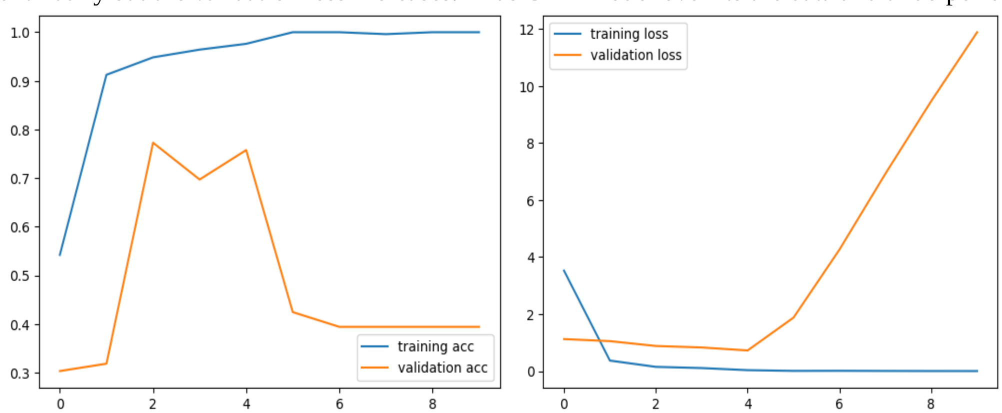
(b) Normal X-rays



(c) Pneumonia X-rays

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

The model training was performed for three models that's is CNN, VGG16, and VGG19. The plots in Fig. 7(a), Fig. 8(a) and Fig. 9(a) shows the performance of the respective models while training in terms of epochs versus accuracy on training data and validation data. Similarly, plots in Fig. 7(b), Fig. 8(b) and Fig. 9(b) shows the performance of the respective models while training in terms of epochs versus loss on training data and validation data. Clearly from Fig. 7 it can be seen that the CNN model overfits the data and thus memorize the training data leading to increase in training accuracy but a very high decrease in validation accuracy. In contrast, the training loss for training data decreases continually but the validation loss increases. Thus CNN model overfits the data and underperform.



(a) (b)

Fig. 5 (a) Epochs versus Training and validation accuracy (b) Epoch versus training and validation loss of CNN model.

Similarly from Fig. 8 and 9 it can be seen that the pre-trained models VGG16 and VGG19 model follows the original matrices curve giving slightly higher accuracy and slightly lower loss for the training data than the validation data. It implies that the model learns effectively using the pre-trained models using transfer learning without overfitting using less data. Thus transfer learning is a preferable approach for model training with less data.

Corresponding Author's E-mail ID:sandeep72128@gmail.com

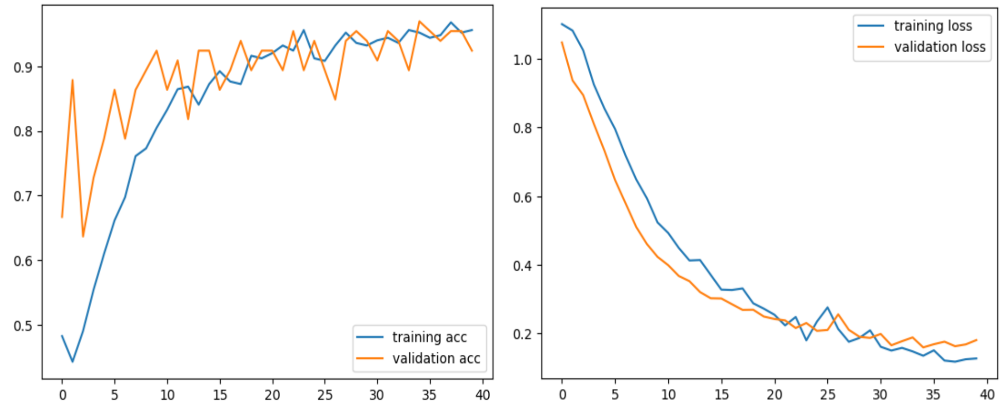


Fig. 6 (a) Epochs versus Training and validation accuracy (b) Epoch versus training and validation loss of VGG16 model.

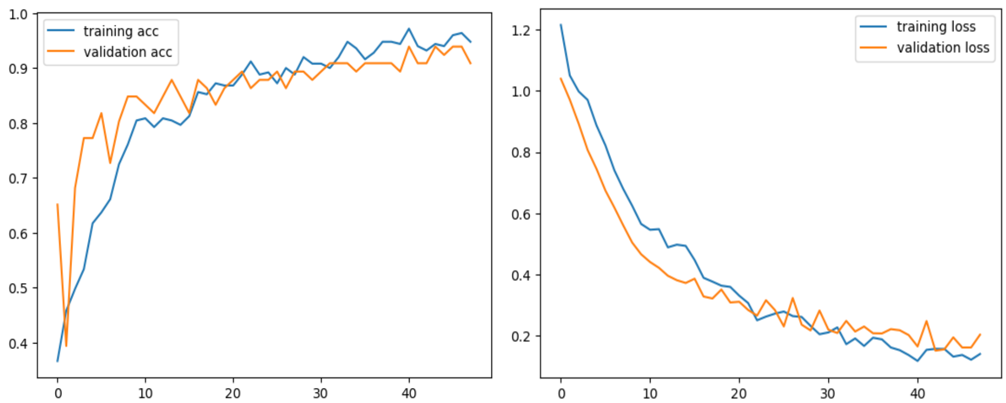
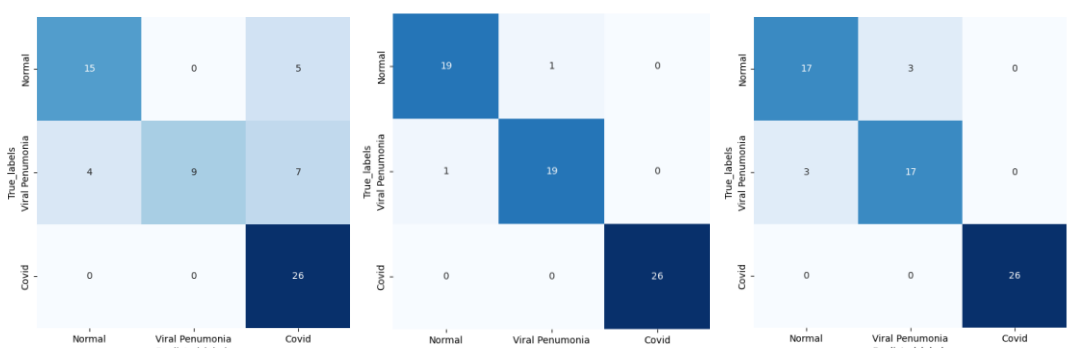


Fig. 7 (a) Epochs versus Training and validation accuracy (b) Epoch versus training and validation loss of VGG19 model.

The confusion metrics of the three schemes are shown in figure 10. Clearly, VGG16 outperforms CNN and VGG19 models due to higher parameters and more overfitting.



(a) (b) (c)

Figure 8: Confusion metric (a) CNN model (b) VGG16 (c) VGG19

Clearly from Figure 10 it can be seen that VGG16 have False positive and False negative. The precision, recall and F1-scores of the different schemes in shown in Table1 with the original data. Class 3 belongs to COVID-19 X-rays, Class 2 belongs to Pneumonia X-rays and Class 1 belongs to Normal X-rays. It also depicts that VGG16 outperforms other models. The non diagonal elements should be ideally zero but this scenario these are relatively less in number. With the VGG their is 100 percent correct classification for Covid X-rays however some miss classification occur for other classes.

Corresponding Author's E-mail ID:sandeep72128@gmail.com

Table 1: The evaluation parameters without data augmentation

Method	Precision			Recall			F1-score		
	Class			Class			Class		
	1	2	3	1	2	3	1	2	3
CNN	0.79	1	0.68	0.75	0.45	1	0.77	0.62	0.81
VGG 16	0.95	0.95	1	0.95	0.95	1	0.95	0.95	1
VGG19	0.85	0.85	1	0.85	0.85	1	0.85	0.85	1

The accuracy comparison of the three schemes on the training, validation and the test data is performed in Table 2 on the original and augmented data. As shown in below table that VGG16 achieved better train accuracy when compared to other schemes but the difference of increases even more for the validation and test data, due to overfiting. Additionally using data augmentation the accuracy of VGG 16 model is further improved.

Table 2: The accuracy and loss on the test data with and without data augmentation.

Method	Original data			Augmented data		
	Train	Validation	Test	Train	Validation	Test
CNN	0.78	0.76	0.76	0.529	0.530	0.53
VGG 16	0.968	0.969	0.97	0.980	0.984	0.98
VGG19	0.964	0.910	0.91	0.940	0.910	0.91

Not only accuracy is improved using the data augmentation of VGG16 model but also the precision, recall and F1 score as mentioned in Table 3.

Table 3: The evaluation parameters with data augmentation

Method	Precision			Recall			F1-score		
	Class			Class			Class		
	1	2	3	1	2	3	1	2	3
CNN	0.67	1	0.46	0.1	0.35	1	0.17	0.52	0.63
VGG 16	1	0.95	1	0.95	1	1	0.97	0.98	1
VGG19	0.85	0.85	1	0.85	0.85	1	0.85	0.85	1

Therefore it can be seen that pre-trained model outperforms the others models and data augmentation in them helps achieving better accuracy but with less parameters I.e. too deep models like VGG19 again have poor performance due to overfitting. Therefore from given schemes VGG16 is the best model.

VII. Conclusion and Future Scope:

The risk that COVID-19 poses to human lives has caused a dramatic deterioration in world health. However, the testing process as it is now is more expensive and lacks quick inference. ML is crucial in the early detection and prompt treatment of COVID. It is a testing approach that costs very little and doesn't require employees to have any specialized training. The development of AI and ML in technology has affected every aspect of life, including medicine, and has produced encouraging outcomes in the field of health care. By examining this data, it is simple to make a choice and diagnose covid utilizing CT scan, X-ray, and cough sounds.

Corresponding Author's E-mail ID:sandeep72128@gmail.com

These instruments can do a preliminary evaluation of suspected patients, assisting them in receiving prompt medical attention and a recommendation for isolation. This study concentrated on identifying COVID using chest X-ray pictures. X-ray pictures are used less frequently than CT scan images, which were previously well-liked. Additionally, supervised learning yields superior results to unsupervised learning techniques. When the dataset is small, transfer learning is a superior strategy, and data augmentation also aids in improving accuracy. When the 251 image dataset from CNN, VGG16, and VGG19 models were examined, VGG16 surpassed all other schemes with the maximum accuracy of 98%.
However, employing a huge dataset can increase accuracy even more. Recurrent supervised learning is a preferable option for achieving higher accuracy since distinct symptoms are appearing when the covid phase changes with the variations. Additionally, employing the transfer learning, federated learning, and incremental learning strategies, we may create convolutional neural networks using SVM. In the future, more effective acceleration techniques can be used to execute these models on hardware with limited resources.

References:

[1] Coronavirus update (live): 28,988,031 cases and 925,320 deaths from COVID-19 virus pandemic - worldometer.
[2] Sudre, Carole H.,"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, "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"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.
[10] Arora, Vinay, "Transfer learning-based approach for detecting COVID-19 ailment in lung CT scan." *Computers in biology and medicine* (2021): 104575.
[11] Barstugan, Mucahid, Umut Ozkaya, and Saban Ozturk. "Coronavirus (covid-19) classification using ct images by machine learning methods." *arXiv preprint arXiv:2003.09424* (2020).
[12] Tuncer, Seda Arslan,. "Scat-NET: COVID-19 Diagnosis with a CNN Model using Scattergram Images." *Computers in Biology and Medicine* (2021): 104579.
[13] 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.
[14] Asif, Sohaib, "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.
[15] Abdul Salam, Mustafa, Sanaa Taha, and Mohamed Ramadan. "COVID-19 detection using federated machine learning." *Plos one* 16.6 (2021): e0252573.
[16] Dey, Subhrajit, "Choquet Fuzzy Integral-based Classifier Ensemble Technique for COVID-19 Detection." *Computers in Biology and Medicine* (2021): 104585.
[17] Brinati, Davide,. "Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study." *Journal of medical systems* 44.8 (2020): 1-12.
[18] Wu, Jiangpeng,. "Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results." *MedRxiv* (2020).
[19] Loey, Mohamed,. "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.
[20] Khanday, Akib MohiUd Din,. "Machine learning based approaches for detecting COVID-19 using clinical text data." *International Journal of Information Technology* 12.3 (2020): 731-739.
[21] Yao, Haochen, "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.
[22] Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., &Ganslandt, T. (2022). Transfer learning for medical image classification: A literature review. *BMC medical imaging*, *22*(1), 69.

Corresponding Author's E-mail ID:sandeep72128@gmail.com

[23] Alzubaidi, L., Fadhel, M. A., Al-Shamma, O., Zhang, J., Santamaría, J., Duan, Y., & R. Oleiwi, S. (2020). Towards a better understanding of transfer learning for medical imaging: a case study. *Applied Sciences*, *10*(13), 4523.

[24] Chen, H., Liu, Z., Liu, Z., Zhang, P., & Yan, Y. (2019). Integrating the data augmentation scheme with various classifiers for acoustic scene modeling. *arXiv preprint arXiv:1907.06639*.

Corresponding Author's E-mail ID:sandeep72128@gmail.com