Sanjukta Nad

Develop a prediction model from Covid-19 Xray dataset

Problem Statement:

The 2019 novel coronavirus (COVID-19) diagnosis is confirmed using polymerase chain reaction (PCR), infected patients with pneumonia may present on chest X-ray and computed tomography (CT) images with a pattern that is only moderately characteristic for the human eye. COVID-19's rate of transmission depends on our capacity to reliably identify infected patients with a low rate of false negatives. In addition, a low rate of false positives is required to avoid further increasing the burden on the healthcare system by unnecessarily exposing patients to quarantine if that is not required. Along with proper infection control, it is evident that timely detection of the disease would enable the implementation of all the supportive care required by patients affected by COVID-19.

Artificial intelligence (AI) techniques in general and convolutional neural networks (CNNs) in particular have attained successful results in medical image analysis and classification. A deep CNN architecture has been proposed in this project for the diagnosis of COVID-19 based on the chest X-ray image classification.

The goal of this project to use available chest radiograph images with clinical findings associated with COVID-19 as a training data set, mutually exclusive from the images with confirmed COVID-19 cases, which will be used as the testing data set.

Deep learning has the potential to revolutionize the automation of chest radiography interpretation. The key component in deep learning research is the availability of training and testing data set, whether or not it is accessible to allow reproducibility and comparability of the research.

Exploratory Data Analysis:

The dataset I have used (https://www.kaggle.com/datasets/khoongweihao/covid19-xray-dataset-train-test-sets) contained testing and training directories. Originally, the training directory has files of 74 normal x-ray images and 74 pneumonia x-ray images. The testing directory has files of 20 normal x-ray images and 20 pneumonia x-ray images.

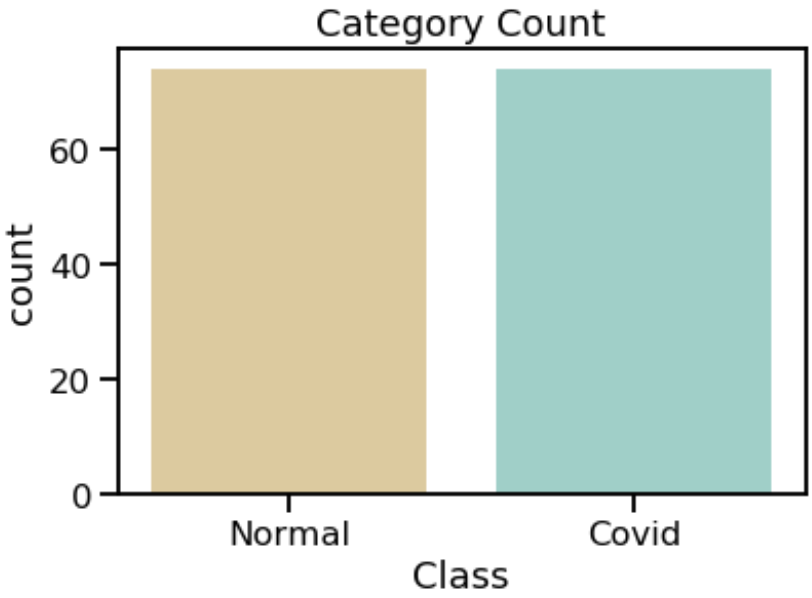

Fig 1: Category count Plot for Normal and Covid class

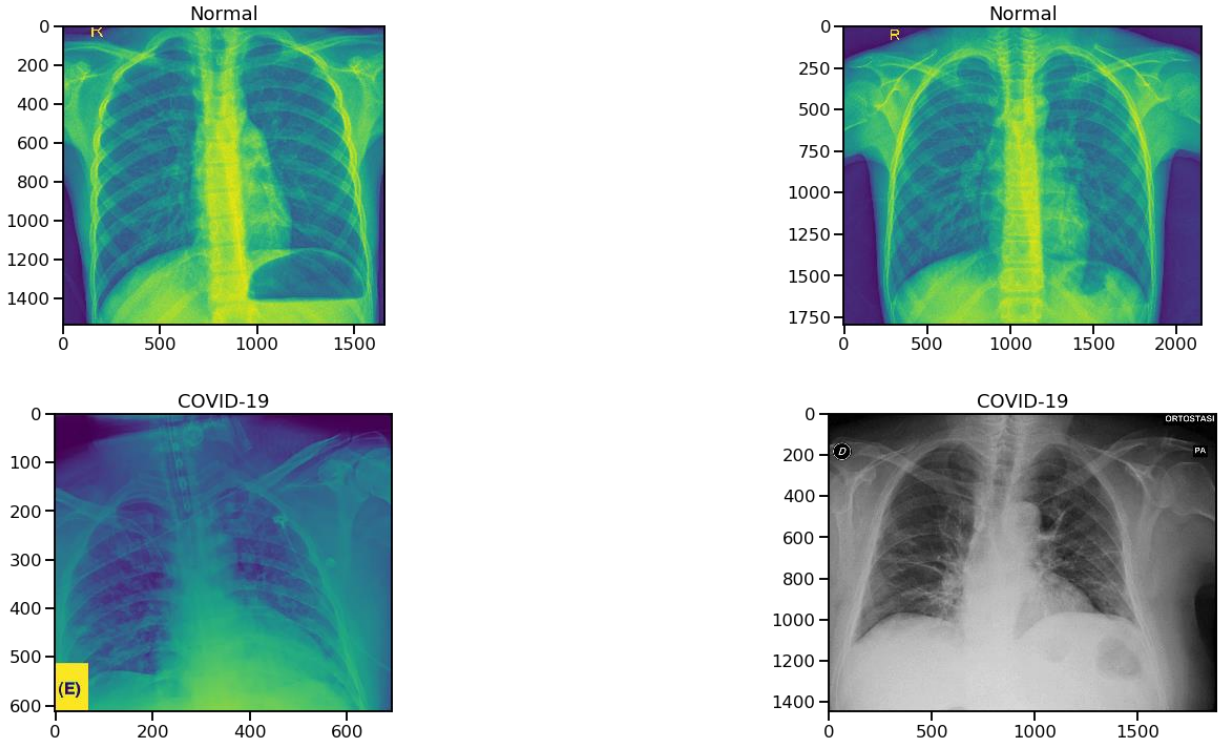

Fig 2: Plotting Normal VS Covid in Grid for random files

Training and Modelling:

Independent sets were used for each training, validation, and testing phase. The depth of deep learning architecture is important for many visual detection applications. With recent opacity-related finding as an important characteristic in COVID-19 patients, this project is aimed at developing a deep learning model for the prediction of COVID-19 cases based on available images.

The Convolutional Neural Networks (CNNs) are inspired by visual system of human brain. The idea behind the CNNs thus is to make the computers capable of viewing the world as humans view it. This way CNNs can be used in the fields of image recognition and analysis, image classification, and natural language processing.

Generally, CNNs contain the convolutional, max pooling, and nonlinear activation layers. The convolutional layer, considered as a main layer of a CNN, performs the operation called “convolution” that gives CNN its name. Kernels in the convolutional layer are applied to the layer inputs. All the outputs of the convolutional layers are convolved as a feature map. In this study, the Rectified Linear Unit (ReLU) has been used in the activation function with a convolutional layer which is helpful to increase the nonlinearity in input image, as the images are fundamentally nonlinear in nature. Thus, CNN with ReLU in the current scenario is easier and faster.

The pooling layer or subsampling layer is also an important building block of CNN. On each feature map extracted through the convolution layer, the pooling layer operates independently. To minimize overfitting and the number of extracted features, it decreases the spatial size of the feature map and returns the important features. Pooling can be the max, average, and sum in the CNN model. In this study, max pooling has been used because others may not identify the sharp features easily as compared to max pooling. In addition, the batch normalization layer has been used in this study as it involved the training of a very deep neural network. So the technique adjusts the scaling and activation to normalize the input layer and speed up the learning procedure between hidden units.

The dropout layer with a 25% dropout rate has also been used for input layer and 10% for the dense layer, which drops the neurons during the training chosen at random to reduce the overfitting problem. Towards the last stage of the CNN used in the study, there is a flatting layer to convert the output of convolutional layers into a single-dimensional feature vector. After flattening, the vector data is given as an input to the next layers of the CNN called fully connected layers or dense layer.

CNN model input image shape is (150, 150, 3), i.e., 150-by-150 RGB image. In Con2D layer, 32 filters with 3×3 size kernel have been used. The max pooling layer with 2×2 pooling size has been used.

This study uses CNN for binary classification; that is the reason for using the binary crossentropy (BCE) loss function. In binary classification since only one output node is needed to classify the data to one of the two given classes, so in the case of BCE loss function, the output value is being given to a sigmoid activation function. The output given by the sigmoid activation function lies between 0 and 1. It finds the error between the predicted class and the actual class. The “Adam” optimizer has been used which changes the attribute weight and learning rate to reduce the loss of the learning model.

Parameter	Value
Input dimension	(150, 150, 3)
Filter to learn	64, 128
Max pooling	2 × 2
Activation functions	ReLU, sigmoid
Dropout rate	25%
Kernel size	3 × 3
Epochs	20
Optimizer	Adam
Loss function	binary_crossentropy

Table 1: The model parameter values

For training and testing the proposed CNN, the dataset was partitioned into two subsets. The training dataset contained 74 COVID-19 X-ray images and 74 normal X-ray images, making a total of 148 X-ray images. The testing dataset similarly contained 40 X-ray images, in which 20 X-ray images were from each class COVID-19 positive and normal. Then, the training subset containing 148 X-ray images has been passed to the model with 10% validation size. So, out of 148 X-ray images, with each 20 epoch, 133 X-ray images train the model, and 15 X-ray Images validate the model.

Fig 3 represents the learning curve for training and validation data over number of iterations.

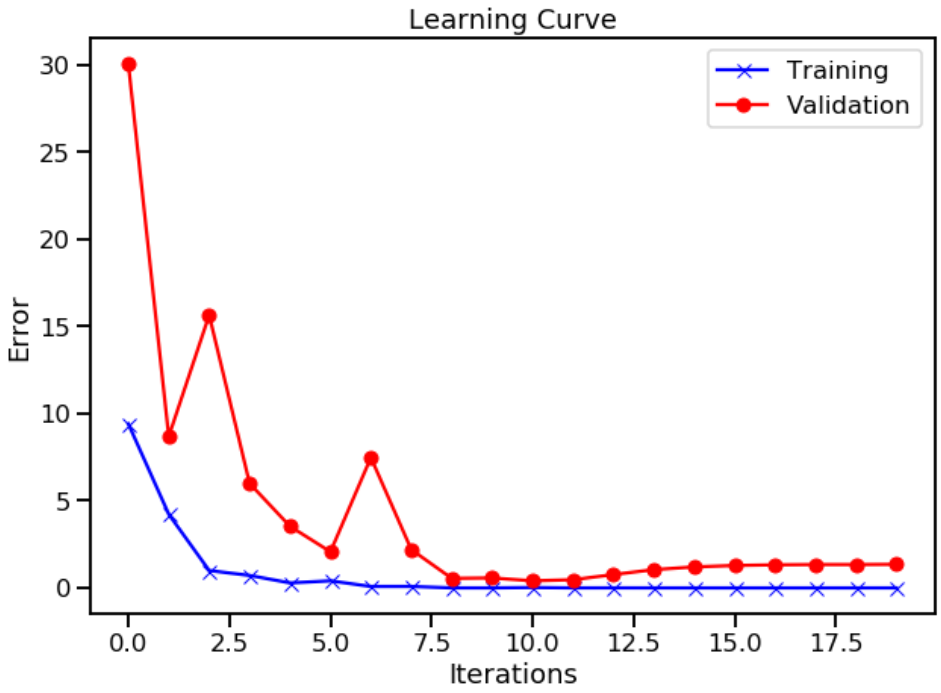

Fig 3: Iteration curve

Performance Evaluation - Accuracy, Classification Report & Confusion Matrix:

The CNN model thus achieved an extraordinary performance with an accuracy of 100% with the test data subset used from the processed dataset of this study with a precision of 1.0. The confusion matrix of the model is shown in Figure 4.

According to the confusion matrix, the CNN model test uses the 40 X-ray images from the GitHub dataset, where 20 images belong to the COVID-19 class and 20 to the normal images. The CNN model shows significant performance on testing and predicts all 40 images correctly with 0% error rate as reported in the confusion matrix.

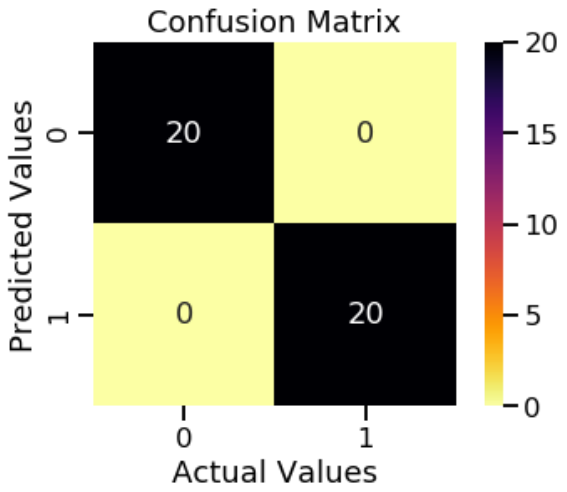

Fig 3: Confusion Matrix

Conclusion and future work:

This goal of this project to demonstrate the effective and accurate diagnosis of COVID-19 using CNN which was trained on chest X-ray image datasets. The model training was performed to attain the maximum accuracy and performance.

In the course of this project effective exploratory image data analysis on Covid vs Normal Images using various techniques was done. Then CNN Model is applied and the experimental results have shown the overall accuracy as high as 100% which demonstrates the good capability of the proposed CNN model in the current application domain.

In future I will be considering using a larger dataset to improve generalization. Due to the nonavailability of sufficient-size and good-quality chest X-ray image dataset, an effective and accurate CNN classification is a challenge. I would like to preprocess the data set like data augmentation. It will be also interesting to do a comparative analysis of proposed CNN model by performance comparisons with some of the prominent machine learning models such as RF, GBM, SVC, LR, and KNN. If it can be proved that the proposed CNN has outperformed all the models particularly when each model was tested on the independent validation dataset that will be also an interesting observation.

Finally, CNN-based diagnosis using X-ray imaging can be very effective for medical sector to handle the mass testing situations in pandemics like COVID-19.