Analysis of Covid19 Disease using Machine Learning

Divya. D [#1], Renukaradhya P C [*2]

[#]*PG Student, CSE Department , Sridevi institute of technology and management, Visveswaraya Technological University*

** Assistant Professor, CSE Department, Sridevi institute of technology and management, Visveswaraya Technological University*

divyadachar@gmail.com rpcmtech@gmail.com

Abstract *- COVID-19 outbreaks only affect the lives of people, they result in a negative impact on the economy of the country. On Jan. 30, 2020, it was declared as a health emergency for the entire globe by the World Health Organization (WHO). By Apr. 28, 2020, more than 3 million people were infected by this virus and there was no vaccine to prevent. The WHO released certain guidelines for safety, but they were only precautionary measures. The use of information technology with a focus on fields such as data Science and machine learning can help in the fight against this pandemic. It is important to have early warning methods through which one can forecast how much the disease will affect society, on the basis of which the government can take necessary actions without affecting its economy. A deep CNN architecture has been proposed in this paper for the diagnosis of COVID-19 based on the chest X-ray image classification. Due to the nonavailability of sufficient-size and good-quality chest X-ray image dataset, an effective and accurate CNN classification was a challenge. To deal with these complexities such as the availability of a very-small-sized and imbalanced dataset with image-quality issues, the dataset has been preprocessed in different phases using different techniques to achieve an effective training dataset for the proposed CNN model to attain its best performance. preprocessing stages of the datasets performed in this study include dataset balancing, medical experts' image analysis, and data augmentation. experimental results have shown the overall accuracy as high as 99.5% which demonstrates the good capability of the proposed CNN model in the current application domain.*

Keywords: Covid19, CNN, Machine learnng, deep learning, Chest X-Ray

I. INTRODUCTION

In this era of automation, artificial intelligence and data science have important role in the health care industry. These technologies are so well-connected that medical professionals can easily manage their roles and patient care. All health care organizations work hard to develop an automated system that can be used to accept the challenges faced in health care. Scientists are working on machine learning (ML) to develop smart solutions to diagnose and treat disease.)e virus called the severe acute respiratory syndrome
coronavirus 2 (SARS-CoV-2) had been discovered in late 2019.)e virus which originated in China became a cause of a disease known as Corona Virus Disease 2019 or COVID- 19.)e World Health Organization (WHO) declared the disease as a pandemic in March 2020 [1, 2]. According to the reports issued and updated by global healthcare authorities and state governments, the pandemic affected millions of people globally.)e most serious illness caused by COVID- 19 is related to the lungs such as pneumonia.)e symptoms of the disease can vary and include dyspnea, high fever, runny nose, and cough.)ese cases can most commonly be diagnosed using chest X-ray imaging analysis for the abnormalities [3]. X-radiation or X-ray is an electromagnetic form of penetrating radiation.)ese radiations are passed through the desired human body parts to create images of internal details of the body part.)e X-ray image is a representation of the internal body parts in black and white shades. X-ray is one of the oldest and commonly used medical diagnosis tests. Chest X-ray is used to diagnose the chest-related diseases like pneumonia and other lung diseases [4], as it provides the image of the thoracic cavity, consisting of the chest and spine bones along with the soft organs including the lungs, blood vessels, and airways.)e X-ray imaging technique provides numerous advantages as an alternative diagnosis procedure for COVID-19 over other testing procedures.)ese benefits include its low cost, the vast availability of X-ray facilities, noninvasiveness, less time consumption, and device affordability.)us, X-ray imaging may be considered a better candidate for the mass, easy, and quick diagnosis procedure for a pandemic like COVID-19 considering the current global healthcare crisis.

II. RELATED WORK

Deep learning has shown a dramatic increase in the medical applications in general and specifically in medical imagebased diagnosis. Deep learning models performed prominently in computer vision problems related to medical image analysis. ANNs outperformed other conventional models and methods of image analysis [7, 8]. Due to the very promising results provided by CNNs in medical image analysis and classification, they are considered as de facto standard in this domain [9, 10]. CNN has been used for a variety of classification tasks related to medical diagnosis such as lung disease [10], detection of malarial parasite in images of thin blood smear [11], breast cancer detection [12], wireless endoscopy images [13], interstitial lung disease [14], CAD-based diagnosis in chest radiography [15], diagnosis of skin cancer by classification [16], and automatic diagnosis of various chest diseases using chest X-ray image classification [17]. Since the emergence of COVID-19 in December 2019, numerous researchers are engaged with the experimentation and research activities related to diagnosis, treatment, and management of COVID-19.

Researchers in [18] have reported the significance of the applicability of AI methods in image analysis for the detection and management of COVID-19 cases. COVID-19 detection can be done accurately using deep learning models' analysis of pulmonary CT [18]. Researchers in [19] have designed an open-source COVID-19 diagnosis system based on a deep CNN. In this study, tailored deep CNN design has been reported for the detection of COVID-19 patients using X-ray images. Another significant study has reported on the X-ray dataset comprising X-ray images belonging to common pneumonia patients, COVID-19 patients, and people with no disease[20].)e study uses the state-of-the-art CNN architectures for the automatic detection of patients with COVID-19.

Transfer learning has achieved a promising accuracy of 97.82% in COVID-19 detection in this study. Another recent and relevant study has been conducted on validation and adaptability of Decompose-, Transfer-, and Compose-type deep CNN for COVID-19 detection using chest X-ray image classification [21].)e authors have reported the results of the study with an accuracy of 95.12%, sensitivity of 97.91%, and specificity of 91.87%.

III. MATERIALS & METHODS

Dataset:

In the experiments of this study, a primary dataset containing 178 X-ray images has been used as a base dataset. Of 178 images, 136 X-ray images belonged to confirmed COVID-19 patients and other 42 images belonged to normal or people with other diseases like pneumonia.)e dataset used is available on GitHub [5].) basic dataset consists of two classes of COVID-19 with 136 samples and others with 42 samples.)us, the dataset was imbalanced and needed preprocessing to achieve promising results. As a first attempt, CNN was trained on the given original dataset and around 54% accuracy was achieved, which was not worthy of the current application domain. The main dataset sources used in this study are enlisted as follows:

(i) Primary chest X-ray image dataset of COVID-19 patients collected from GitHub.)e dataset has been collected by the University of Montreal's Ethics Committee no. CERSES-20-058-D from different hospitals and clinics [5].

(ii) For dataset balancing, a collection of chest X-ray images were collected from Kaggle [22].

(iii) Independent validation dataset containing a collection of 100 COVID-19 X-ray images for the realworld testing of the proposed CNN was collected from IEEE DataPort [6]. experiments have been conducted using Core i7 7thgeneration machine with 8 GB RAM, Microsoft Windows 10 platform using Python language with Anaconda 3 software and Jupyter Notebook.

A. Dataset Preprocessing

Balancing Dataset Classes.

To balance the given dataset, in order to improve the performance of the proposed CNN models in the detection of COVID-19 cases, 136 normal chest X-ray images have been used.)ese concatenated extra X-ray images were downloaded from Kaggle [22]. After balancing the dataset when the models have been trained again on the resulted dataset, the accuracy of the given CNN models was improved to 69%. Still, the performance given by the models in terms of accuracy and other measures was not justified as an effective system for COVID-19 detection.

Analysis of X-Ray Images by Medical Experts.

A deep analysis was done on the X-ray images by medical specialists. Out of 135 X-ray images of confirmed COVID-19 patients, only a set of 90 X-ray images was selected as a perfect candidate to train the models.)e resulted dataset now was reduced to 90 COVID-19-confirmed cases and 90 normal X-ray images.)e resulted dataset was again used in training the proposed CNN model; there was again an improvement in the performance of the model. Specifically, the accuracy was increased to 72% in the given scenario. Still, because the dataset was not containing a sufficient number of images for an effective training, there was not a significant increase in the accuracy and other performance metrics.

IV. ARCHITECTURE OF THE MODEL

The Proposed CNN Architecture.)e proposed CNN model consists of 38 layers in which 6 are convolutional (Conv2D), 6 max pooling layers, 6 dropout layers, 8 activation function layers, 8 batch normalization layers, 1 flatten layer, and 3 fully connected layers; CNN model input image shape is (150, 150, 3), i.e., 150-by-150 RGB image.
In all Con2D layers, a 3 × 3 size kernel has been used but the filter size after every two Con2D layers increases. At the 1st and 2nd layers of Con2D, 64 filters have been used to learn from input and the 3rd and 4th layers of Con2D use 128 filters, and at the 5th and 6th layers, 256 filters have been used. After each Con2D layer, the max pooling layer with 2 × 2 pooling size has been used, the batch normalization layer has been used with the axis argument, the activation layer has been used with the ReLU function, and the dropout layer has been used with 20% dropout rate.

The output of 256 output neurons of the final Con2D layer is followed by max pooling, batch normalization, activation, and dropout layer. Since the final pooling and convolutional layer gives a three-dimensional matrix as output, to flatten the matrix, a flattening layer has been used which converts them into a vector that will be input for 3 dense layers. 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. The model parameter values are given in Table 3, and the model architecture is given in Figure 4. During the initial experiments, the CNN has been used with different configurations in terms of the usage of number of convolution layers in the model. The decision of how many convolution layers used in the model was made by using an incremental approach. First, the CNN was tested using only one convolutional layer and the results were analysed. Then, the CNN was built with two layers and results were analysed and so on. The approach had been continued till the results provided by the model were accurate and effective. The final model which was very feasible according to its results consisted of six convolution layers. The results of each increment of the model have been reported in the Results section.

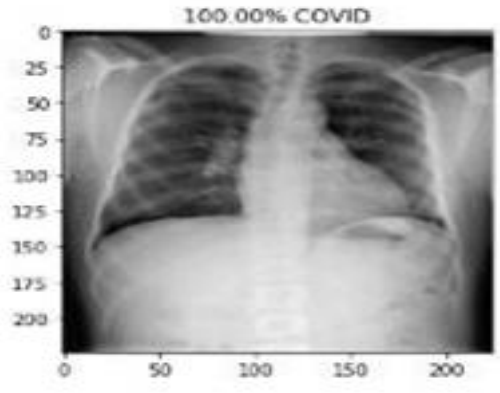
V. RESULTS AND DISCUSSION

After preprocessing of the dataset, the final dataset consisted of a total of 900 X-ray images. For training and testing the proposed CNN, the dataset was partitioned into two subsets.

The training dataset contained 400 COVID-19 X-ray images and 400 normal X-ray images, making a total of 800 X-ray images. The testing dataset similarly contained 100 X-rayimages, in which 50 X-ray images were from each class COVID-19 positive and normal. Then, the training subset containing 800 X-ray images has been passed to the model with 25% validation size. So, out of 800 X-ray images, with each epoch, 600 X-ray images train the model, and 200 X-ray Images validate the model. As mentioned in the proposed architecture of the CNN model, it consisted of 38 layers in which 6 are convolutional, 6 max pooling layers, 6 dropout layers, 8 activation function layers, 8 batch normalization layers, 1 flattening layer, and 3 fully connected layers. 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,
with the model parameter values. To evaluate the overall performance, in addition to accuracy, other important metrics have been adopted in this study including *F*1 score, precision, sensitivity, specificity, and ROC AUC. The scores of these parameters are reported in

The confusion matrix of the model is shown in Figure. According to the confusion matrix, the CNN model test uses the 100 X-ray images from the GitHub dataset, where 50 images belong to the COVID-19 class and 50 to the normal images. The CNN model shows significant performance on testing and predicts all 100 images correctly with 0% error rate as reported in the confusion matrix. Training accuracy of the CNN according to remains consistent after the 5 epochs and the CNN also shows a consistent validation accuracy after the 25 epochs. The training loss of CNN is minimum and consistent from the 1st epoch while validation loss becomes minimum after 5



epochs and remains consistent till the last epoch. The above results show the efficiency of the CNN model proposed in this study.

Figure 1a: Chect X-Ray

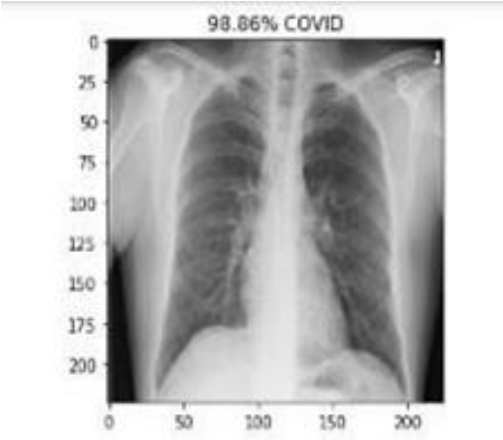


Figure 1b: Chect X-Ray

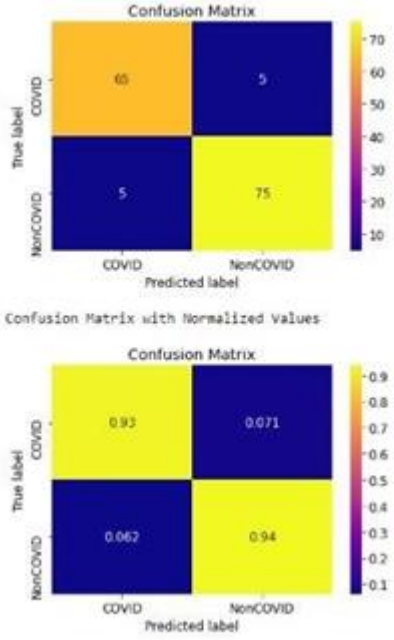


Figure 2: Confusion Matrix

REFERENCES

[1] D. Cucinotta and M. Vanelli, "WHO declares COVID-19 a pandemic," *Acta Biomedica: Atenei Parmensis*, vol. 91, pp. 157–160, 2020.

[2] F. Rustam, A. A. Reshi, A. Mehmood et al., "COVID-19 future forecasting using supervised machine learning models," *IEEE Access*, 2020.

[3] D. J. Cennimo, "Coronavirus disease 2019 (COVID-19) clinical presentation," vol. 8, pp. 101489–101499, 2020, https://emedicine.medscape.com/article/2500114-clinical#b2, 2020. Online.

[4] J. P. Cohen, "Github Covid19 X-ray dataset," 2020, https://github.com/ieee8023/covid-chestxray-dataset, 2020. Online.

[5] Z. H. Chen, "Mask-RCNN detection of COVID-19 pneumonia symptoms by employing stacked autoencoders in deep unsupervised learning on low-dose high resolution CT," IEEE Dataport, 2020.

[6] M. Ahmad, "Ground truth labeling and samples selection for hyperspectral image classification," *Optik*, vol. 230, Article ID 166267, 2021.

[7] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *Proceedings of the 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, pp. 844–848, Singapore, December 2014.

[8] M. Umer, S. Sadiq, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood, "A novel stacked CNN for malarial parasite detection in)in blood smear images," *IEEE Access*, vol. 8, pp. 93782–93792, 2020.

[9] M. Sharif, M. Attique Khan, M. Rashid, M. Yasmin, F. Afza, and U. J. Tanik, "Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images," *Journal of Experimental& .eoretical Artificial Intelligence*, pp. 1–23, 2019.

[10] N. Asada, K. Doi, H. MacMahon et al., "Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung diseases: pilot study," *Radiology*, vol. 177, no. 3, pp. 857–860, 1990.

[11] S. Katsuragawa and K. Doi, "Computer-aided diagnosis in chest radiography," *Computerized Medical Imaging andGraphics*, vol. 31, no. 4-5, pp. 212–223, 2007.

[12] Y. Dong, Y. Pan, J. Zhang, and W. Xu, "Learning to read chest X-ray images from 16000+ examples using CNN," in *Proceedings of the 2017 IEEE/ACM International Conference on Connected Health: Applications*, Systems and Engineering Technologies (CHASE), pp. 51–57, Philadelphia, PA, USA, July 2017.

[13] D. Dong, Z. Tang, S. Wang et al., ")e role of imaging in the detection and management of COVID-19: a review," IEEE Reviews in Biomedical Engineering, vol. 14, pp. 16–19, 2020.

[14] P. Mooney, "Kaggle X rays dataset," 2020, https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia Online.

[15] Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, p. 60, 2019.