2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) | 978-1-6654-1460-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICT4SD50815.2021.9396890

A Comparative Study of Classification Approaches for COVID-19 Prediction

S. M. Mahedy Hasan[1], Md. Fazle Rabbi[2], Arifa Islam Champa[3], Md. Asif Zaman[4]

[1]Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi-6204, Bangladesh
[2,3]Department of Computer Science & Engineering, Islamic University of Technology, Gazipur, Bangladesh
[4]North South University

***Abstract*— The novel coronavirus (COVID-19), a highly infectious disease that first found at Wuhan Province of China in Dec 2019, spread worldwide in some months and already become a pandemic. Covid-19 has already changed the world economic structure, people's religious, political, social life, public health structure, people's daily life structure and also made millions of people jobless. The only way to fight this epidemic is to identify the infected person as soon as possible and separate them from a healthy person, so that they can't infect anyone again. At present, RT-PCR is currently used to detect coronavirus patients around the world. But the World Health Organization (WHO) said that RT-PCR suffers from low sensitivity and low specificity for early-stage cases. Recent research has shown that chest CT scan images play a beneficial role in identifying coronavirus cases. In this study, we compared the performances of four classification algorithms, such as Random Forest (RF), Support Vector Machine (SVM), Extra Trees (ET), and Convolutional Neural Network (CNN) for classifying COVID-19 cases and proposed a prediction model based on classification results. The result shows that our proposed CNN model outperformed the other classification algorithms and obtained an accuracy of 98.0%.**

Keywords*— *Coronavirus (COVID-19), Machine Learning, Deep learning, Convolutional Neural Network, CT scan images

I. INTRODUCTION

Coronavirus Disease 2019 (COVID 19) was emerged from Wuhan, China, and spread all over China. Eventually, it escalates around the world within a short time and turns into a world pandemic. As of 11th August 2020, the number of infected cases is 20,254,662, and death cases are 738,930 worldwide [1]. By detecting COVID-19 cases early, the patients can be secluded so that non-infected persons remain safe. Currently, the global method to diagnose COVID patients is the reverse transcription-polymerase chain reaction (RT-PCR). However, the primary issue is that it suffers from low sensitivity and specificity [2]. Besides, because of the scarcity of RT-PCR test kits in the remote areas, the doctors recommend utilizing medical images for screening COVID-19 [3][4]. Computed Tomography (CT) scan image carries valuable details for detecting positive COVID-19 patients [5][6]. Despite the benefits of the CT scan image, it may contain similar features between COVID-19 and other pulmonary diseases; thereby, screening is hard. Machine learning and deep learning techniques come in handy in extracting and detecting features from radiological images in recent times.

This study has employed several machine learning and deep learning techniques to detect COVID-19 positive patients from CT scan images. We have applied a relatively big dataset comprising of 2481 images. After comparing those methods, we have perceived the best model and urged that the model could verify patients with COVID-19. The remaining fragment of the paper is organized as follows: Section II describes the related works performed in recent times. Section III describes about the used dataset in this paper. Section IV represents the pipeline of the research methodology. Section V analyses the experimental outcomes. Section VI presents the proposed prediction model and section VII concludes the paper.

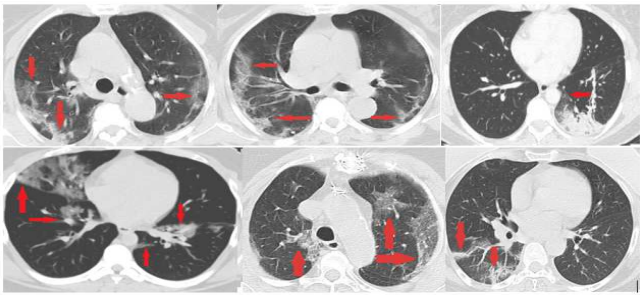
II. LITERATURE REVIEW

From the emergence of the pandemic, an automatic screening system of COVID-19 becomes a top priority for the research community. Kang et al. [7] proposed a multi-view representation learning technique that can automatically diagnose from COVID-19. To validate their model, they applied 2522 CT scan images. Their method reached accuracy, sensitivity and specificity of 95.5%, 96.6% and 93.2%, respectively. Li et al. [8] suggested COVNet, a deep learning automatic framework, to identify COVID-19 accurately using chest CT. Chest CT consisting of 4356 images were employed in building their model. This model achieved a sensitivity of 87% and an AUC of 0.95 while detecting corona patients from other pneumonia patients. Xu et al. [9] designed a deep learning model, ResNet, for early screening of COVID-19. A sum of 618 pulmonary CT samples was used while constructing the model. A final accuracy of 86.7% was attained distinguishing COVID-19 from Influenza-A pneumonia and healthy cases. Ardakani et al. [10] compared 10 (ten) convolutional neural networks i.e. VGG-16, VGG-19, AlexNet, GoogleNet, SqueezeNet, ResNet-18, ResNet-50, ResNet-101, MobileNet-V2 and Xception for distinguishing corona (COVID) and other pneumonia (Non-COVID). They used 1020 slices of CT images. Moreover, they observed that ResNet 101 and Xception acquired the highest AUC of 0.994 and recommended Resnet 101 to characterize and detection of COVID patients. X. Bai et al. [11] presented a deep neural network architecture, EfficientNet, and applied CT slices from 1,186 patients into that architecture. While differentiating between COVID and Non-COVID, their introduced system achieved 96% accuracy, 95% sensitivity and 96% specificity. Shi et al. [12] applied Random Forest (RF) as a machine learning algorithm for screening COVID-19. They utilized CT images of 2685 patients to evaluate their presented model. After assessing the 5-fold cross-validation technique, the model reached accuracy, sensitivity, and specificity of 87.9%, 90.7%, and 83.3%, respectively. Ozkaya et al. [13] generated 3000 patch images from 150 CT images and further applied ranking and fusion techniques on those images. They employed Support Vector Machine (SVM) for classification, and before that,

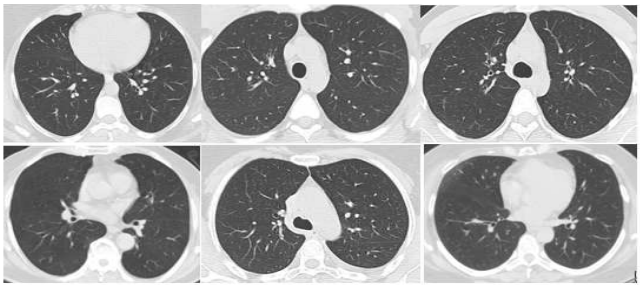
they used pre-trained CNN models as a part of the transfer learning method. Their presented procedure attained 98.27% accuracy, a precision of 97.63%, and a sensitivity of 97.6%. Alom et al. [14] presented an efficient deep learning approach, Residual RCNN, with transfer learning, for detecting COVID-19. Earlier, they employed the NABLA-N network for infected area segmentation, which enhances the outcome of classification. They applied both X-ray and CT images for the evaluation of the proposed method. They acquired 84.67% and 98.78% accuracy from X-ray and CT images, respectively. Sahlol et. al [15] utilized marine predator algorithm to extract deep features for COVID-19 classification. Abbas et. al [16] proposed a DeTrac model based on chest x-ray images to classify COVID-19 cases. Sadik et. al [17] compared the performances of different machine learning algorithms for COVID-19 prediction.

III. Dataset Description

In this research, CT scan images were collected from a publicly available data repository named Kaggle for classifying COVID-19 patients [18]. In this dataset, there were 2482 chest CT images collected from Sao Paolo, Brazil. In this dataset, 1252 chest CT scan images contained positive COVID-19 cases and other 1230 chest CT scan images had other pulmonary diseases that meant negative COVID-19 cases. Some samples of CT scan images of the chest from this dataset are shown in Fig. 1.



(a)



(b)

Fig.1. Samples of CT scan images of (a) COVID-19 positive cases and (b) COVID-19 negative cases. Red arrows in (a) indicates the contaminated region.

IV. Methodology

A. Image Preprocessing

All the images obtained from the dataset were of distinct sizes. We adopted the resize function from Python Open CV to bring all the images back to the same size. Color space conversion was performed after bringing all the images to a uniform size. All of the images were converted to the gray color space from the BGR color space. The preprocessing phase was then finished by converting all the images into arrays for further processing. The flowgraph of our research work is shown in Fig.2.

B. Classification

1) Convolutional Neural Network

Over the past few years, deep learning has achieved an emerging interest in medical research such as object recognition, brain tumor segmentation, and classification, breast cancer detection, cervical cancer recognition, etc. CNN is a part of deep learning, is mostly applied in solving computer vision type problems. The complete architecture of CNN comprises three layers, such as a convolutional layer, a pooling layer followed by a fully connected layer. The first two layers extract deep features from the input image, and a fully connected layer maps those extracted features to the output layer.

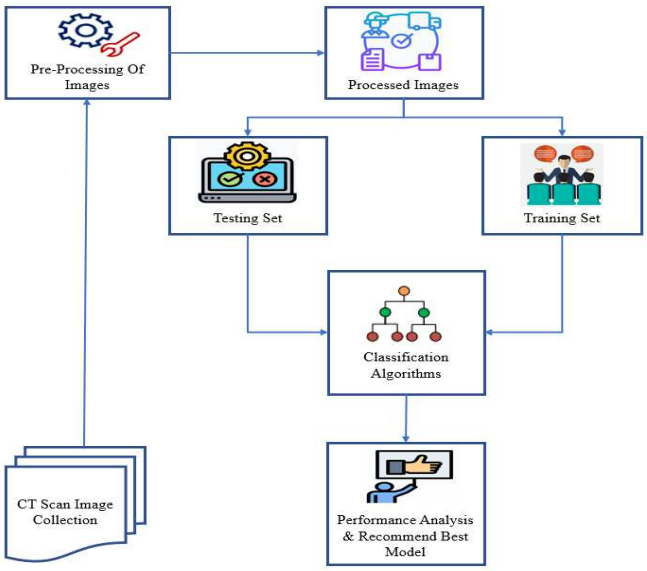


Fig. 2. Workflow of Our Research Methodology

Convolutional Layer: Convolutional layer performs the feature extraction tasks. In the linear convolutional process, a feature detector or kernel is used to extract features from the input image. Element-wise product operation is conducted between input tensor and kernel to produce a convolved image, also known as a feature map or activation map. The primary purpose of this convolution is to reduce the size of the input image. Mathematically, the following equation represents the convolutional operation:

$$(f * g)(t) = \int_{-\infty}^{+\infty} f(T)\, g(t-T)\, dT \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where, f(t) is a function of an input image, g(t) function of feature detector, t at any moment and T represents the amount of shift.

After convolution, the Rectified Linear Unit (RELU) breaks up the linearity and enhancing the non-linearity in the network. The output of the function is outlined as follows:

$$f(x) = \max(0, x) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where, x represents the input to the neuron.

Pooling Layer: Pooling extracts features from the convolved image to minimize the size of the feature map to create a pooled feature map. Later, the pooled feature map is flattened into a one-dimensional column and fed into an artificial neural network for additional processing.

Fully Connected Layer: This full connection is composed of a consortium of an input layer with an output layer through a fully connected layer, depicted in Fig.3. Here, the fully connected layers are also known as a special type of hidden layer as all nodes are fully connected. In these layers, ReLU is employed as an activation function.

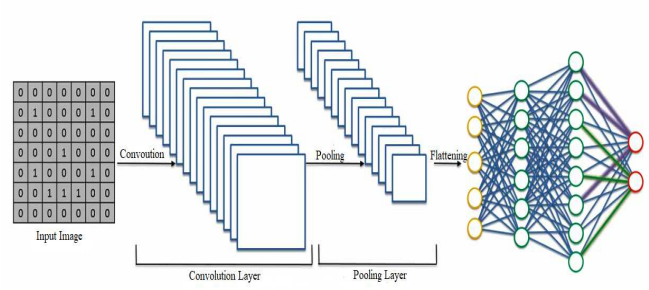


Fig. 3. Complete Architecture of CNN [22]

The final prediction is assessed in the output layer, where softmax or sigmoid are utilized as an activation function. The information goes through ANN in this way. The function of the prediction error or loss is also measured, which defines how well the network is functioning. This function needs to be minimized to optimize the network. To get a perfect prediction, the prediction error is backpropagated through the network.

2) Random Forest(RF)

Random forest, an ensemble Machine Learning (ML) method, which is generally used for solving classification and regression problems [19]. This algorithm creates forest by combining the multiple numbers of trees. In classification, the higher number of trees in the forest provides better results in prediction. Random forest classifier works by creating decision trees on data samples, taking a prediction from each tree, and finding the best solution using majority voting.

3) Support Vector Machine (SVM)

SVM is a supervised ML algorithm used to solve both classification and regression related tasks. The main intention of SVM is to obtain a maximum marginal hyperplane by dividing the datasets into distinct classes. It can be done in two steps:

1. At first, SVM creates hyperplanes in an iterative manner that separates the levels in the best way.
2. Secondly, it will select the hyperplane that segregates the classes accurately.

When the data points are not linearly separable, then SVM employs a kernel method. There are many different types of kernels, but for this research, Radial Basis Function (RBF) has been used. The following equation represents the RBF kernel:

$$K(x,x') = exp\left(-\frac{(||x-x'||)^2}{2\sigma^2}\right) \dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

Where, $(||x - x'||)^2$ defines the squared Euclidean distance between the two feature vectors and σ is a free parameter.

4) Extra Trees (ET)

ET is an ensemble ML technique generally used for solving classification and regression type problems. Extra tree classifier generates an immense number of unpruned DT's from the training dataset. The majority voting technique is used to make predictions where each decision tree gives a vote, and the highest voted prediction is considered the final classification result. Unlike the RF algorithm, it chooses the random features to split the node instead of the best features.

C. Evaluation Measures

To evaluate the efficiency of this proposed model, we employed K-fold cross-validation techniques to randomly split the dataset into different k subsets to make the training set and test set. In this study, 10-fold cross-validation was adopted to lessen bias and variance. It is known to all that accuracy is a standard evaluation metric for measuring the performance of any prediction model. In this study, six (6) different evaluation measures were considered, such as accuracy, AUC, sensitivity, specificity, precision, and recall. The confusion matrix manifests the overall performance of any prediction model. The confusion matrix is introduced in Table I. By utilizing this confusion matrix, those six measures can be calculated.

TABLE I. CONFUSION MATRIX

Actual	**Predicted**	
	Positive	***Negative***
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (4)$$

$$Precision = \frac{TP}{FP+TP} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (5)$$

$$Sensitivity = \frac{TP}{FN+TP} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (6)$$

$$F1-Score = \frac{2TP}{FN+FP+2TP} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (7)$$

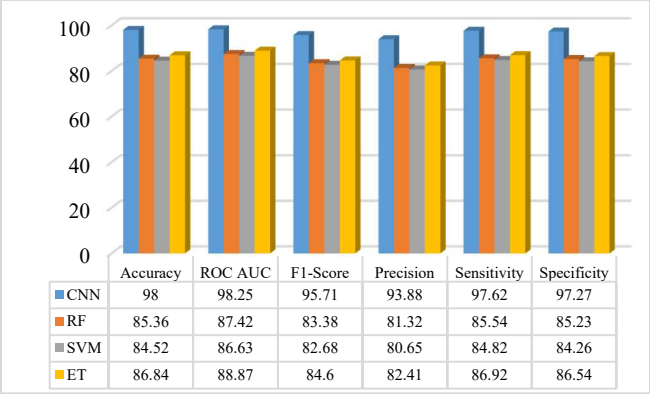
$$Specificity = \frac{TN}{FP+TN} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (8)$$

Receiver Operating Characteristic Curve (ROC AUC): ROC AUC is a pictorial representation that demonstrates the performance of a two-class classification system.

V. EXPERIMENTAL RESULTS

In this research, Google Colaboratory [20] was used, which is a cloud service based on python programming language for developing applications provided by Google. For faster processing, virtual Tensor Processing Unit (TPU) was used.

We divided our experimental dataset into a 75:25 ratio, where 75% of the data was used as training, and 25% of the data as testing [21]. In this research, we compared the performances of four classification techniques for classifying COVID-19 cases are presented in Fig. 4. From the experimental results, it is shown that RF, SVM, ET, and CNN provide 85.36%, 84.52%, 86.84%, and 98% accuracy respectively in classifying coronavirus patients. Furthermore, while reviewing experimental outcomes in terms of ROC AUC, CNN has found with the highest ROC AUC value of 98.25%. Scrutinizing the F1-score values of the four classification algorithms, it is seen that CNN, RF, SVM and ET have obtained 95.71%, 83.38%, 82.68% and 84.6% respectively. Same as before, in terms of F1-score value, CNN has delivered the best result than the other classification methods. Apart from that, while we analyzing the experimental outcomes in terms of precision, sensitivity and specificity values, it is seen that CNN has delivered the best results such 93.88%, 97.62% and 97.27% respectively and outperformed the other classifiers same as before. Therefore, From Fig. 4, we can say that the CNN classifier has outperformed the remaining classification algorithms in terms of prediction performance for the Covid-19 classification.



	Accuracy	ROC AUC	F1-Score	Precision	Sensitivity	Specificity
CNN	98	98.25	95.71	93.88	97.62	97.27
RF	85.36	87.42	83.38	81.32	85.54	85.23
SVM	84.52	86.63	82.68	80.65	84.82	84.26
ET	86.84	88.87	84.6	82.41	86.92	86.54

Fig. 4. Performance Analysis of Classification Algorithms for COVID-19 Cases Detection

VI. Discussion

In this section, we will discuss the proposed prediction model in detail. From the results section, we have already seen that CNN classifier has outperformed other algorithms in terms of performance. From this research, we can say that our proposed CNN model can be used for detecting COVID-19 cases through chest CT scan images. The complete structure of our proposed CNN model is depicted in Fig. 5.

The images in the experimental dataset were of different sizes, so they were brought to a uniform size of 64 * 64 pixels. Then 64 filters of 3 * 3 sizes were used in the first convolution layer to extract features from input images.

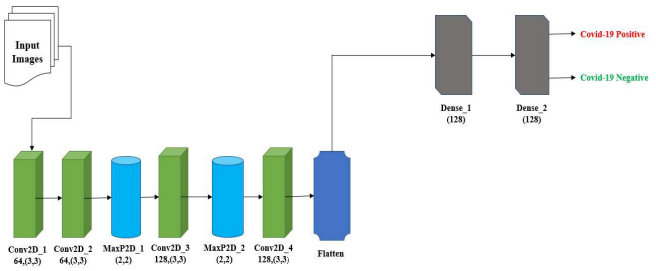


Fig. 5. Structure of Proposed CNN Model

A second convolution layer was added as same as the configuration of the first convolution layer, and ReLU was used as the activation function. Max Pooling of 2 * 2 size was later used to reduce the size of the feature map. Then another convolution layer was added where there were 128 kernels of 3*3 sizes. Max Pooling was then added as before to make the size of the feature map smaller. Then the final convolution and the max pulling layer were added again and then flattened into a one-dimensional array. A total of two dense/hidden layers of 128 units and 256 units were then added, where ReLU was used as an activation function. Only one node was used to get the final prediction about COVID-19 classification in the final output layer. As it was a binary classification problem, that's why the sigmoid function was adopted as an activation function. The proposed CNN model used Adam as an optimizer and binary cross-entropy to calculate the loss function.

To make the proposed CNN model more robust and avoid overfitting and underfitting, the entire operation was performed up to 100 number of epochs with a batch size of 32. Fig. 6 shows that at the beginning of the epochs, training loss was extremely high, and accuracy was quite low.

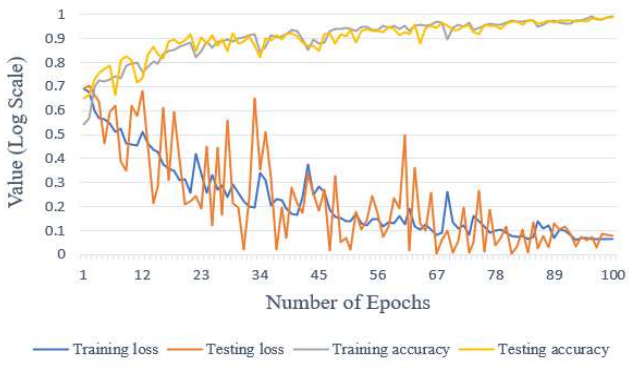
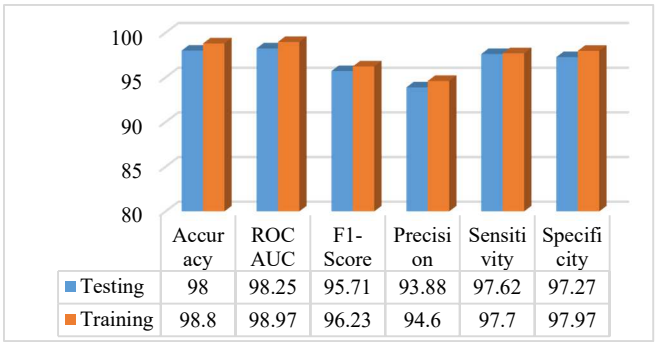


Fig. 6. Training and Testing loss, Training and Testing Accuracy Curves of Proposed CNN Model



	Accuracy	ROC AUC	F1-Score	Precision	Sensitivity	Specificity
Testing	98	98.25	95.71	93.88	97.62	97.27
Training	98.8	98.97	96.23	94.6	97.7	97.97

Fig. 7. Comparison of Training and Testing Performances

In each step, weights were updated continuously following their batch size and forwarded to the network. After completing each epoch, loss value gradually decreased, and accuracy increased. After concluding all of the 100 epochs, the proposed CNN model obtained an accuracy of 98.0%. The overall training and testing performances of CNN is represented and compared in Fig. 7.

VII. Conclusion

RT-PCR is currently being used for coronavirus detection in almost all countries around the world. RT-PCR is giving more false-negative results, and at the same time, there is an extreme shortage of this RT-PCR kits all over the world. So, chest CT scan images along with artificial intelligence techniques can play an active role in solving this problem and helping humankind to overcome this crisis. In our study, we compared four machine learning and deep learning algorithms performance to identify and classify coronavirus patients accurately. Analyzing the performance of four algorithms, it can be said that that the CNN algorithm can extract hidden information from CT scan images to identify COVID-19 patients. Our proposed CNN model has achieved excellent accuracy of 98.0%. This proposed model can be used as an alternative tool or assistive tool along with Rt-PCR in rural areas where there is a lack of adequate identification kits and expert physicians. This cost-effective prediction model will be able to identify the coronavirus patients in a minute so that the affected people cannot infect anyone later, and the community spread is stopped. If the size of the dataset had been larger, we could have developed a more robust prediction model to identify Covid-19 patients. This prediction model can be stored in the cloud for taking chest CT scan images as input and give results within a minute. We will collect CT scan images of Covid-19 patients from local hospitals, clinics, and diagnostic centers in Bangladesh and evaluate them with our developed models.

References

[1] "CORONAVIRUS PANDEMIC," [Online]. Available: https://www.worldometers.info/coronavirus/. [Accessed: 11-August-2020].

[2] G. Bleve, L. Rizzotti, F. Dellaglio and S. Torriani "Development of reverse transcription (RT)-PCR and real-time RT-PCR assays for rapid detection and quantification of viable yeasts and molds contaminating yogurts and pasteurized food products." Applied and environmental microbiology vol. 69,7 (2003): 4116-22. doi:10.1128/aem.69.7.4116-4122.2003.

[3] A. Lal, A. K. Mishra and K. K. Sahu, "CT chest findings in coronavirus disease-19 (COVID-19)," J Formos Med Assoc. 2020;119(5):1000-1001. doi:10.1016/j.jfma.2020.03.010.

[4] C. Long et al., "Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT?," Eur J Radiol. 2020;126:108961. doi:10.1016/j.ejrad.2020.108961.

[5] D. Dong et al., "The Role of Imaging in the Detection and Management of COVID-19: A Review." IEEE reviews in biomedical engineering vol. 14 (2021): 16-29. doi:10.1109/RBME.2020.2990959.

[6] F. Shi et al., "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19," in IEEE Reviews in Biomedical Engineering, vol. 14, pp. 4-15, 2021, doi: 10.1109/RBME.2020.2987975.

[7] H. Kang et al., "Diagnosis of Coronavirus Disease 2019 (COVID-19) With Structured Latent Multi-View Representation Learning," IEEE transactions on medical imaging vol. 39,8 (2020): 2606-2614. doi:10.1109/TMI.2020.2992546.

[8] L. Li et al., "Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy," Radiology vol. 296,2 (2020): E65-E71. doi:10.1148/radiol.2020200905.

[9] X. Xu et al., "Deep learning system to screen coronavirus disease 2019 pneumonia," arXiv preprint arXiv:2002.09334, 2020.

[10] A. A. Ardakani, A. R. Kanafi, U. R. Acharya, N. Khadem and A. Mohammadi, "Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks," Computers in Biology and Medicine, vol. 121, pp. 103795, 2020.
https://doi.org/10.1016/j.compbiomed.2020.103795.

[11] H. X. Bai et al., "Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT," Radiology vol. 296,3 (2020): E156-E165. doi:10.1148/radiol.2020201491.

[12] F. Shi et al., "Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification," Physics in medicine and biology, 10.1088/1361-6560/abe838. 19 Feb. 2021, doi:10.1088/1361-6560/abe838.

[13] U. Özkaya, S. Öztürk and M. Barstugan, "Coronavirus (COVID-19) Classification Using Deep Features Fusion and Ranking Technique," Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach. Studies in Big Data, vol 78. Springer (2020), Cham. https://doi.org/10.1007/978-3-030-55258-9_17.

[14] M. Z. Alom et al., "COVID_MTNet: COVID-19 Detection with MultiTask Deep Learning Approaches," 2020 arXiv:2004.03747.

[15] A. T. Sahlol et al., "COVID-19 image classification using deep features and fractional-order marine predators algorithm," Sci Rep 10, 15364 (2020). https://doi.org/10.1038/s41598-020-71294-2.

[16] A. Abbas, M. M. Abdelsamea and M. M. Gaber, "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network," Appl Intell (2020). https://doi.org/10.1007/s10489-020-01829-7.

[17] R. Sadik, M. L. Reza, A. A. Noman, S. A. Mamun, M. S. Kaiser and M. A. Rahman, "COVID-19 Pandemic: A Comparative Prediction Using Machine Learning," Int J Auto AI Mach Learn. 2020;1(1): 01 - 16.

[18] E. Soares, P. Angelov, S. Biaso, M. H. Froes, and D. K. Abe, "SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification," medRxiv (2020). doi: https://doi.org/10.1101/2020.04.24.20078584.

[19] S. M. M. Hasan, M. A. Mamun, M. P. Uddin and M. A. Hossain, "Comparative Analysis of Classification Approaches for Heart Disease Prediction," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, 2018, pp. 1-4.

[20] "Google Colaboratory," [Online]. Available: colab.research.google.com. [Accessed: 12-May-2020].

[21] G. E. P. Box and R. D. Meyer, "An Analysis for Unreplicated Fractional Factorials," Technometrics, vol. 28, no. 1, pp. 11-18, 1986.

[22] "Deep Learning A-Z™: Hands-On Artificial Neural Networks," [Online]. Available: https://www.udemy.com/course/deeplearning/. [Accessed: 28-Aug-2020].