Performance Evaluation of Deep Learning-Based COVID-19 Diagnosis Software: A Comprehensive Approach Using Convolutional Neural Networks and Ensemble Machine Learning

YETUNDE ESTHER OGUNWALE[1], OLUYEMISI ADENIKE OYEDEMI[2], MICHEAL OLALEKAN AJINAJA[3]

[1, 2] *University of Ilesa, Ilesa*

[3] *Federal Polytechnic Ile Oluji, Ile Oluji*

Abstract- ***Quick diagnosis of COVID-19 through chest X-ray images has gained significant attention due to its potential to aid in rapid screening. In this study, we presented a comprehensive approach utilizing convolutional neural networks (CNNs) for feature extraction from chest X-ray images, followed by an ensemble of classifiers including Decision Tree, Support Vector Machine, Random Forest, and AdaBoost for accurate classification. Our CNN architecture, trained on Google Colab with GPU runtime, comprises 20 layers incorporating Conv2D, MaxPooling2D, Dropout, and fully connected layers with ReLU activation function and a dropout threshold of 0.25, achieving an accuracy of 97.10%. By using a dataset that consists of 33,920 chest X-ray (CXR) images including 11,956 COVID-19, 11,263 Non-COVID infections (Viral or Bacterial Pneumonia), 10,701 Normal with Ground-truth lung segmentation masks provided for the entire dataset from the Kaggle COVID-19 Radiography Database. Our final ensemble classifier, employing Soft voting, attained a heightened accuracy of 97.51%. Moreover, to gain insights into the CNN's internal processes, we visualized intermediate layer activations. Subsequently, we deployed the final model using a Flask API for seamless integration into healthcare systems. Our approach promised efficient and accurate diagnosis of COVID-19 from chest X-ray images, facilitating timely patient management.***

Indexed Terms- ***Deep learning. Convolutional Neural Networks. Ensemble Learning. Chest X-ray***

I. INTRODUCTION

In the world of medical diagnostics, the rapid and accurate identification of COVID-19 has become one of the most critical challenge worldwide. Chest X-ray imaging stands as a pivotal tool in this endeavor, offering a non-invasive means to detect characteristic pulmonary abnormalities associated with the disease [1]. Using advancements in artificial intelligence, particularly convolutional neural networks (CNNs), has shown promise in automating this diagnostic process [2]. CNNs, inspired by the hierarchical structure of the human visual cortex, excel at extracting intricate features from complex data such as medical images [3].

Through successive layers of convolutions, pooling, and non-linear activations, these networks learn to discern subtle patterns indicative of COVID-19 infection, facilitating accurate classification [4]. Furthermore, ensemble learning techniques have been widely adopted to bolster the robustness and generalization capabilities of such models [5]. The journey towards an effective CNN-based diagnostic system encompasses several crucial stages. First and foremost, the architecture of the CNN must be meticulously crafted, accounting for factors such as depth, kernel size, and activation functions [6]. Additionally, the availability of high-quality, annotated datasets plays a pivotal role in training and validating these models, ensuring their reliability and generalizability [7].

Moreover, the interpretability of CNNs remains a paramount concern, particularly in medical

applications where transparency and trustworthiness are imperative [8]. Visualizing intermediate layer activations provides valuable insights into the decision-making process of these networks, elucidating the features they deem salient in distinguishing between COVID-19-positive and negative cases [9].

In this paper, we presented a comprehensive approach for automated COVID-19 diagnosis from CXR images using a CNN-based ensemble classifier. We also explored techniques for visualizing intermediate layer activations within the CNN architecture to gain insights into the model's decision-making process. Throughout this study, we aimed to contribute to the ongoing efforts in medical image analysis with the ultimate goal of improving diagnostic accuracy and facilitating knowledge discovery in healthcare domains.

The dataset consists of 33,920 chest X-ray (CXR) images [10] including 11,956 COVID-19, 11,263 Non-COVID infections (Viral or Bacterial Pneumonia), 10,701 Normal with Ground-truth lung segmentation masks provided for the entire dataset. The experiments were conducted on two CXR sets, where each set is divided into train, validation and test sets - Lung Segmentation Data and Entire COVID-QU-Ex dataset (33,920 CXR images with corresponding ground-truth lung masks).

Section 2 presents the review of relevant literature. The methodology, data source, pre-processing, and model architecture are presented in Sect. 3 while Sect. 4 focuses on results and discussions. The conclusion drawn from the research is presented in Sect. 5

II. RELATED WORKS

Several studies have investigated the application of deep learning techniques in COVID-19 diagnosis, contributing to the advancement of diagnostic tools and methods. This section reviews relevant literature that explores similar approaches to deep learning-based COVID-19 diagnosis, focusing on convolutional neural networks (CNNs) and ensemble machine learning techniques.

[11] proposed a deep learning framework utilizing a CNN architecture for COVID-19 detection from chest X-ray images. Their model demonstrated promising results in distinguishing COVID-19 cases from other pneumonia types, achieving high accuracy rates. Similarly, [12] introduced a CNN-based approach for COVID-19 diagnosis using computed tomography (CT) images. Their study emphasized the effectiveness of CNNs in automatic feature extraction and classification tasks, highlighting the potential of deep learning in aiding medical professionals in timely diagnosis. In addition to individual deep learning models, ensemble learning techniques have been explored to enhance COVID-19 diagnosis accuracy. [13] employed an ensemble of deep neural networks for the classification of COVID-19 cases based on CT images. By combining multiple models, their ensemble approach demonstrated improved generalization performance and robustness against variations in data distribution.

Furthermore, efforts have been made to integrate various modalities of medical imaging data into a unified diagnostic framework. For instance, [14] proposed a comprehensive approach that combines features extracted from chest X-ray and CT images using ensemble machine learning methods. Their study emphasized the synergistic effects of integrating multiple imaging modalities for more accurate and reliable COVID-19 diagnosis. [15] proposed a CNN-based model for COVID-19 detection from chest X-ray images, incorporating transfer learning to leverage pre-trained networks. Their approach showcased promising results in terms of accuracy and computational efficiency, highlighting the potential of transfer learning in medical imaging analysis.

[16] addressed the urgent need for accurate COVID-19 diagnosis by using deep convolutional neural networks (DCNN) and transfer learning. Through a comprehensive evaluation of eight pre-trained models on chest X-ray images, the study demonstrated the efficacy of fine-tuning DenseNet121, achieving a remarkable test accuracy of 98.69% and a macro f1-score of 0.99 for four-class classification. Notably, the findings revealed that only 62% of total parameters needed retraining, underscoring the computational efficiency and accuracy of the fine-tuned models.

This study in [17] utilized X-ray image datasets containing cases of bacterial pneumonia, confirmed Covid-19, and normal incidents to automatically detect Coronavirus disease using state-of-the-art convolutional neural network architectures and transfer learning. The experiment involved two datasets, sourced from public medical repositories, comprising a total of 1427 and 1442 X-ray images respectively. Results indicated that deep learning with X-ray imaging showed promise in extracting significant biomarkers related to Covid-19, achieving high accuracy (96.78%), sensitivity (98.66%), and specificity (96.46%), suggesting the potential for incorporating X-rays into the disease diagnosis process pending further evaluation by the medical community.

To address the challenges posed by the COVID-19 pandemic, particularly in the context of diagnosing the virus accurately amidst a surge in cases, [18] utilized Convolutional Neural Networks (CNNs) on X-ray images, the study aimed to automate COVID-19 detection, providing a scalable solution for hospitals overwhelmed with patient volumes. Experimental findings demonstrated the effectiveness of CNNs in achieving precise and accurate COVID-19 detection, with an impressive accuracy of 96.8%. The COVID-19 pandemic, declared by the WHO in 2019, has resulted in over 6.18 million confirmed cases and 104,000 deaths globally, underscoring the urgency for effective diagnostic methods. [19] developed a deep learning model based on CNN algorithms, utilizing chest X-ray images for early COVID-19 diagnosis, achieving notable accuracy rates above 95% through modified architectures such as EfficientNet, Inception MobileNetV2, ResNet, and Xception. The COVID-19 pandemic has profoundly impacted global health and economies, necessitating early and accurate screening methods to curb further transmission. [20] introduced a Mask R-CNN approach for detecting ground-glass opacities (GGOs) in chest CT images of COVID-19 patients, achieving a high accuracy of 98.25% during instance segmentation, thus offering a valuable tool to expedite screening and validation processes for healthcare professionals.

[21] employed convolutional neural networks (CNN) for feature extraction from CT exams and XGBoost for classification, achieving high accuracy (95.07%) and demonstrating its potential as a diagnostic aid system for specialists. The methodology consisted of using a CNN to extract features from 708 CTs, 312 with COVID-19, and 396 Non-COVID-19. After the extracted data, the team used XGBoost for classification. The results show an accuracy of 95.07, recall of 95.09, precision of 94.99, F-score of 95, AUC of 95, and a kappa index of 90.

The limitations of this work presented in this section include the study's reliance on a single modality. Also, most of the work had restriction of practical application and scalability of the developed deep learning models. Without a dedicated platform or software solution, healthcare providers may face challenges in deploying and utilizing these models effectively in clinical settings. Others are CNNs can be prone to overfitting, especially when trained on limited or imbalanced datasets. More specifically, the objective of the research was to create a system allowing users, particularly medical professionals, to upload patient images, enabling the system to provide a diagnosis indicating whether the X-ray image shows signs of COVID-19 or appears normal. Additionally, the study explored the integration of three distinct machine learning models to enhance performance, evaluating their effectiveness using various metrics.

III. METHODOLOGY

In the section below, we presented four important techniques in our research.

3.1 Data Source

The dataset consists of 33,920 chest X-ray (CXR) images [10] including 11,956 COVID-19, 11,263 Non-COVID infections (Viral or Bacterial Pneumonia), 10,701 Normal with Ground-truth lung segmentation masks provided for the entire dataset. The experiments were conducted on two CXR sets, where each set is divided into train, validation and test sets - Lung Segmentation Data and Entire COVID-QU-Ex dataset (33,920 CXR images with corresponding ground-truth lung masks).

3.2 Data Preprocessing

This marks the initial stage of our research. Following data collection, we proceeded to randomly select 2531 COVID X-ray images for training, 723 for validation,

and 362 for testing purposes. A similar process was carried out for normal X-ray images within the dataset. Subsequently, all images were scaled to dimensions of (299,299,3) and normalized. After preprocessing, the data was divided into train and validation data. The test data was used for performance analysis of the system. CNN was used for feature extraction of images as stated earlier. An Ensemble model was developed based on the four machine learning models used. The architecture of the system is shown in Fig. 1.

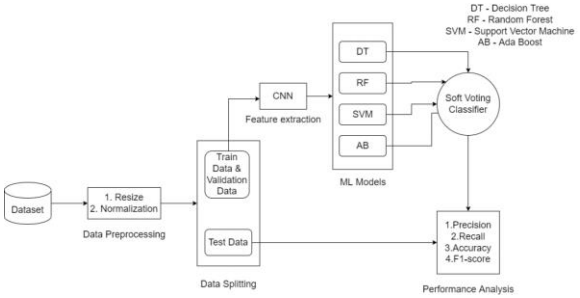

Fig. 1 Architecture of the Proposed Covid-19 Diagnosis System

3.3 Feature Extraction using CNN

Feature extraction using Convolutional Neural Networks (CNNs) is a critical process in image analysis and computer vision tasks. It involves extracting relevant patterns or features from raw image data to represent them in a more abstract and meaningful way. To reduce dimensionality of the image data and capture essential information from images, such as edges, textures, shapes, and other discriminative patterns, feature extraction process is paramount.

CNNs consist of multiple convolutional layers. Each convolutional layer applies a set of learnable filters (kernels) to the input image, performing convolutions to produce feature maps. These feature maps represent the presence of specific patterns or features at various spatial locations within the input image. Mathematically, the operation of applying a filter W_i to a portion of the input image X can be represented as a convolution operation followed by a bias term and an activation function as shown in Eq. 1.

$$Z_i = f\left(\sum_{l=1}^{d}\sum_{m=1}^{d}\left(X_{(l,m)} * W_{i(l,m)}\right) + b_i\right) \quad (1)$$

where Z_i is the output feature map corresponding to the i-th filter, $X_{(l,m)}$ is the input image patch centered around pixel *(l,m)*, $W_{i(l,m)}$ is the corresponding filter weights, b_i is the bias term for the *i-th* filter and *f* is the activation function such as ReLU.

For pooling layers after each convolutional layer, pooling layers (max pooling) was applied to reduce the spatial dimensions of the feature maps. Mathematically, the max pooling operation can be represented as $Y_{(i,j)}$ in Eq. 2.

$$Y_{(i,j)} = \max_{(p,q)\ \epsilon\ pooling\ region}\left(Z_{(i+p,j+q)}\right) \quad (2)$$

where $Y_{(i,j)}$ is the output of the max pooling operation at position *(i, j)*, $\left(Z_{(i+p,j+q)}\right)$ is the feature map value at position *(i + p,j + q)*. The max pooling operation is applied over a predefined pooling region.

Non-linear activation functions ReLU (Rectified Linear Unit) was applied after each convolutional and pooling operation to introduce non-linearity and enable the network to learn complex representations. The Activation functions introduces non-linearity into the network which is represented in Eq. 3.

$$f(x) = \max(0, x) \quad (3)$$

Eventually, the feature maps are flattened into a vector representation, which serves as the input to fully connected layers (also known as dense layers) in the network. For the research work, the flatten operation result was 25088 which corresponds to the total number of neurons in the output feature maps of the preceding convolutional and pooling layers. The fully connected layers further process the extracted features to perform tasks like classification or regression as represented in Eq. 4.

$$A = f(W X + b) \quad (4)$$

where A is the output vector, W is the weight matrix of the layer, X is the input vector, b is the bias vector and *f* is the activation function. These layers combine the learned features from previous layers and map them to the desired output classes. For the research, there are 64 individual neurons in the layer meaning there are 64 sets of weights and biases that will be learned during the training process. To curtail overfitting, a dropout of 0.25 was used during training.

Fig. 2 shows the diagrammatic representation of each process.

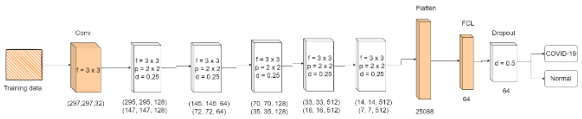

Fig. 2 Step-by-Step Representation of CNN for feature extraction

The CNN has 20 layers of various types including Conv2D, MaxPooling2D, Dropout and FCL. ReLu activation function was used for the inner layers and dropout threshold of 0.25. The CNN learns to extract meaningful features from a labeled training dataset consisting of images and corresponding ground-truth labels. The trained CNN is evaluated on a separate test dataset that it hasn't seen during training. This ensures unbiased assessment of the model's performance on unseen data. The feature extraction process remains the same during testing, but the extracted features are used for inference or prediction without further parameter updates.

In CNN, the number of epochs refers to the number of times the entire training dataset is passed forward and backward through the neural network. Each epoch consists of one forward pass (computing predictions and losses) and one backward pass (updating weights using backpropagation). For the purpose of this research, 50 epochs were used due to complexity of the model, size of the training dataset and learning rate which determines the size of the steps taken during gradient descent optimization. To measure the performance of the epoch function, Fig. 3 shows the accuracy function and Fig. 4 shows the loss function graph. It can be seen on the graph (Fig. 3) that both training accuracy and validation accuracy increase rapidly and approach 1.0 as the number of epochs increases. Also, both training loss and validation loss decrease rapidly and approach a minimum value as the number of epochs increases. The graph suggests that the CNN model is effectively learning from the training data and generalizing well to unseen data which is due to the early epoch and overfitting avoidance in the model.

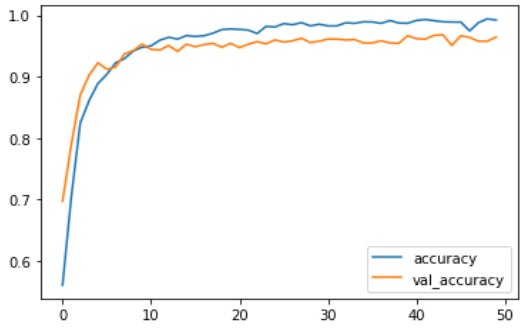

Fig. 3 Accuracy/Validation Accuracy Graph

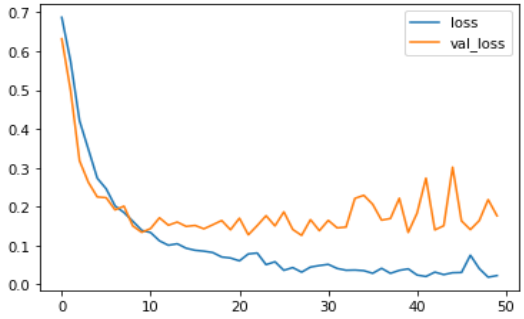

Fig. 4 Loss/Validation Loss Graph

Th next stage is to performance analysis of the CNN model. Table 1 shows the classification report for the binary classification between COVID-19 and Normal cases. Precision measures the proportion of true positive predictions (correctly identified COVID-19 cases) out of all positive predictions (all cases predicted as COVID-19). A precision of 0.9749 for COVID-19 and 0.9671 for Normal indicated that the model had a high percentage of correct positive predictions for both classes. Recall measures the proportion of true positive predictions (correctly identified COVID-19 cases) out of all actual positive cases (all COVID-19 cases in the dataset). A recall of 0.9669 for COVID-19 and 0.9751 for Normal indicated that the model captured a high percentage of actual positive cases for both classes. F1-score is the harmonic mean of precision and recall, providing a single metric that balances both precision and recall. A high F1-score (0.9709 for COVID-19 and 0.9711 for Normal) indicated a good balance between precision and recall for both classes. Support represents the number of samples in each class in the dataset. There are 362 samples for both COVID-19 and Normal classes. Accuracy measures the overall correctness of the model's predictions across all classes. An accuracy of 0.9710 indicated that the

model correctly predicted the class for approximately 97.10% of the samples. The macro average calculates the average of precision, recall, and F1-score across all classes. In this case, the macro average for precision, recall, and F1-score is 0.9710, indicating balanced performance across classes. The weighted average calculates the average of precision, recall, and F1-score, weighted by the number of samples in each class. In this case, the weighted average for precision, recall, and F1-score was 0.9710, indicating balanced performance considering the class distribution.

Table 1 Classification report for CNN Model

	precision	recall	F1-score	Support
Covid	0.9749	0.9669	0.9709	362
Normal	0.9671	0.9751	0.9711	362
Accuracy			0.9710	724
Macro avg	0.9710	0.9710	0.9710	724
Weighted avg	0.9710	0.9710	0.9710	724

Overall, the classification report suggested that the model achieved high precision, recall, and F1-score for both COVID-19 and Normal classes, indicating strong performance in distinguishing between the two classes. The high accuracy further confirms the overall effectiveness of the model in classification. Fig. 5 shows the confusion matrix for the CNN model. There are 350 instances (True Negative) correctly predicted as Covid when they are actually Covid, 12 instances (False Positive) incorrectly predicted as Normal when they are actually Covid, 9 instances (False Negative) incorrectly predicted as Covid when they are actually Normal and 353 instances (True Positive) correctly predicted as Normal when they are actually Normal.

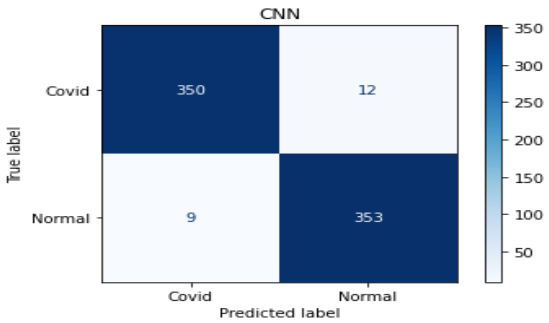

Fig. 5 Confusion Matrix for CNN

3.4 Machine Learning Models

The objective of the research is to use machine learning models and combine all to form an Ensemble model to predict whether a patient as Covid-19 or not using a software. The models used for the research are decision tree, Random forest, Support Vector Machine and Ada Boost.

3.4.1 Decision Tree

A decision tree is a non-linear supervised learning algorithm used for classification and regression tasks. It partitions the feature space into regions and makes predictions based on the majority class within each region. We can denote a decision tree as *T*. The prediction of a decision tree for a sample *x* can be represented as:

$$\hat{y}T = \mathrm{T}(x) \quad (5)$$

3.4.2 Random Forest

Random Forest aggregates predictions from multiple decision trees. Let T_i represent the *i-th* decision tree in the Random Forest. The prediction of the Random Forest for a sample *x* can be represented as:

$$\hat{y}\mathrm{RF} = \frac{1}{N} \sum_{i=1}^{N} T_i(x) \quad (6)$$

where N is the number of trees in the Random Forest.

3.4.3 Support Vector Machine (SVM)

Support Vector Machine finds the hyperplane that best separates the classes in the feature space. Let's denote an SVM model as SVM. The prediction of an SVM for a sample *x* can be represented as:

$$\hat{y}\mathrm{SVM} = \mathrm{SVM}\ (x) \quad (7)$$

3.4.4 AdaBoost

AdaBoost combines predictions from multiple weak learners. Let H_i represent the *i-th* weak learner. The prediction of AdaBoost for a sample *x* can be represented as:

$$\hat{y}_{AdaBoost} = \text{sign}\left(\sum_{i=1}^{N} \alpha_i H_i\ (x)\right) \quad (8)$$

3.4.5 Soft Voting Classifier (Ensemble Model)

In a Soft Voting Classifier, predictions from individual models are combined by averaging their class probabilities. Let's denote the Soft Voting Classifier as Voting. The prediction of the Soft Voting Classifier for a sample xx can be represented as:

$$\hat{y}\text{Voting} = \text{argmax}\left(\frac{1}{M}\sum_{j=1}^{M} p_j(x)\right) \quad (9)$$

where M is the number of individual models (in this case, 4) and $p_j(x)$ represents the probability estimates from the *j-th* model for each class.

To combine the predictions from the Decision Tree, Random Forest, SVM, and AdaBoost models into a Soft Voting Classifier, we computed the probability estimates $p_j(x)$ for each sample *x* using each model, and then average these probabilities across all models. We then chose the class with the highest average probability as the final prediction for the software.

IV. DISCUSSION OF FINDINGS

The experiment setup for deep learning was implemented on Google Colab with GPU runtime using the following packages: Tensorflow, Scikit-learn, pandas, numpy, matplotlib, flask in Python 3.7.12

4.1 Decision Tree Performance Metrics

For the decision tree, the confusion matrix is shown in Fig. 6 and classification table shown in Table 2. A precision of 0.9635 for Covid and 0.9484 for Normal indicated that the model had a high percentage of correct positive predictions for both classes. It had a recall of 0.9475 for Covid and 0.9641 for Normal indicates that the model captured a high percentage of actual positive cases for both classes. A high F1-score (0.9554 for Covid and 0.9562 for Normal) indicated a good balance between precision and recall for both classes. There are 362 samples for both Covid and Normal classes. An accuracy of 0.9558 indicated that the model correctly predicted the class for approximately 95.58% of the samples. In this case, the macro average for precision, recall, and F1-score was 0.9559, indicating balanced performance across classes. The weighted average for precision, recall, and F1-score was 0.9559, indicating balanced performance considering the class distribution. Overall, the classification report suggested that the model achieved high precision, recall, and F1-score for both Covid and Normal classes, indicating strong performance in distinguishing between the two classes.

4.2 Support Vector Machine (SVM) classifier Performance Metrics

A precision of 0.9670 for Covid and 0.9722 for Normal indicated that the model had a high percentage of correct positive predictions for both classes. A recall of 0.9724 for Covid and 0.9669 for Normal indicated that the model captured a high percentage of actual positive instances for both classes. A high F1-score (0.9697 for Covid and 0.9695 for Normal) indicated a good balance between precision and recall for both classes. There are 362 instances for both Covid and Normal classes. An accuracy of 0.9696 indicated that the model correctly predicted the class for approximately 96.96% of the instances. In this case, the macro average for precision, recall, and F1-score was 0.9696, which indicated balanced performance across classes. In this case, the weighted average for precision, recall, and F1-score was 0.9696, indicating balanced performance considering the class distribution. In general, the classification report suggested that the SVM model achieved high precision, recall, and F1-score for both Covid and Normal classes, indicating strong performance in distinguishing between the two classes. The confusion matrix is as shown in Fig. 7 and classification table in Table 3.

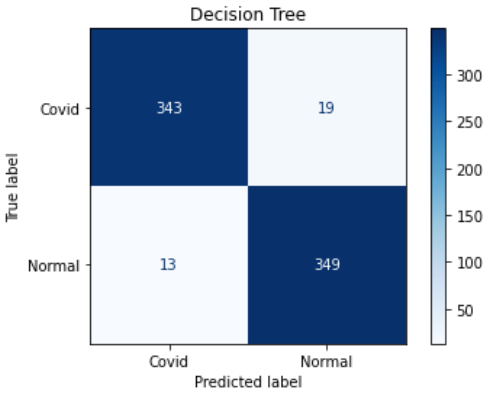

Fig. 6 Confusion Matrix for Decision Tree

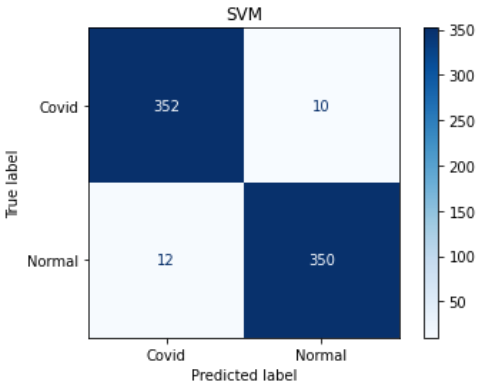

Fig. 7 Confusion Matrix for SVM

Table 2 Classification report for Decision Tree Model

	Precision	Recall	F1-score	Support
Covid	0.9635	0.9475	0.9558	362
Normal	0.9484	0.9641	0.9562	362
Accuracy			0.9558	724
Macro avg	0.9559	0.9558	0.9558	724
Weighted avg	0.9559	0.9558	0.9558	724

Table 3 Classification report for SVM Model

	precision	recall	F1-score	Support
Covid	0.9670	0.9724	0.9697	362
Normal	0.9722	0.9669	0.9695	362
Accuracy			0.9696	724
Macro avg	0.9696	0.9696	0.9696	724
Weighted avg	0.9696	0.9696	0.9696	724

4.3 Random Forest Performance Metrics

A precision of 0.9721 for Covid and 0.9617 for Normal indicated that the model had a high percentage of correct positive predictions for both classes. A recall of 0.9613 for Covid and 0.9724 for Normal indicated that the model captured a high percentage of actual positive instances for both classes. A high F1-score (0.9667 for Covid and 0.9670 for Normal) indicated a good balance between precision and recall for both classes. There are 362 instances for both Covid and Normal classes. An accuracy of 0.9669 indicated that the model correctly predicted the class for approximately 96.69% of the instances. In this case, the macro average for precision, recall, and F1-score is approximately 0.9669, indicating balanced performance across classes. The weighted average for precision, recall, and F1-score is approximately 0.9669, indicating balanced performance considering the class distribution. The classification report suggested that the Random Forest model achieved high precision, recall, and F1-score for both Covid and Normal classes, indicating strong performance in distinguishing between the two classes. The confusion matrix is as shown in Fig. 8 and classification table in Table 4.

4.4 Ada Boost Performance Metrics

A precision of 0.9614 for Covid and 0.9640 for Normal indicated that the AdaBoost model had a high percentage of correct positive predictions for both classes. A recall of 0.9641 for Covid and 0.9613 for Normal indicated that the model captured a high percentage of actual positive instances for both classes. A high F1-score (0.9628 for Covid and 0.9627 for Normal) indicated a good balance between precision and recall for both classes. There are 362 instances for both Covid and Normal classes. An accuracy of 0.9627 indicated that the AdaBoost model correctly predicted the class for approximately 96.27% of the instances. In this case, the macro average for precision, recall, and F1-score was approximately

0.9627, indicating balanced performance across classes. The weighted average for precision, recall, and F1-score was approximately 0.9627, indicating balanced performance considering the class distribution. The classification report suggested that the AdaBoost model achieved high precision, recall, and F1-score for both Covid and Normal classes, indicating strong performance in distinguishing between the two classes. The confusion matrix is as shown in Fig. 9 and classification table in Table 5.

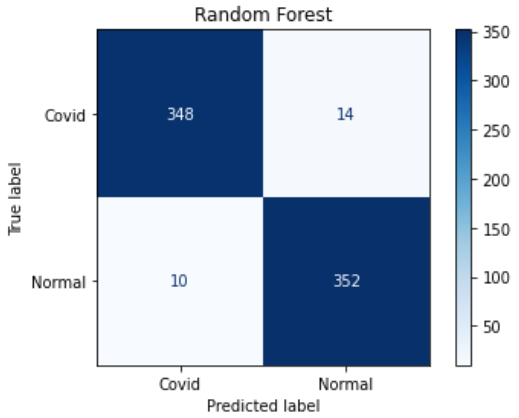

Fig. 8 Confusion Matrix for Random Forest

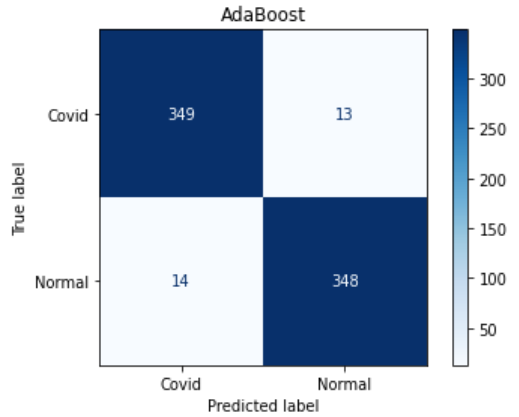

Fig. 9 Confusion Matrix for Ada Boost

Table 4 Classification report for Random Forest Model

	Precision	Recall	F1-score	Support
Covid	0.9721	0.9613	0.9667	362
Normal	0.9617	0.9724	0.9670	362
Accuracy			0.9669	724
Macro avg	0.9669	0.9669	0.9668	724
Weighted avg	0.9669	0.9669	0.9668	724

Table 5 Classification report for Ada Boost Model

	precision	Recall	F1-score	Support
Covid	0.9614	0.9724	0.9628	362
Normal	0.9640	0.9669	0.9627	362
Accuracy			0.9627	724
Macro avg	0.9627	0.9627	0.9627	724
Weighted avg	0.9627	0.9627	0.9627	724

4.5 Ensemble Model Performance Metrics

A precision of 0.9751 for both Covid and Normal classes indicated that the ensemble model had a high percentage of correct positive predictions for both classes. A recall of 0.9751 for both Covid and Normal classes indicated that the model captured a high percentage of actual positive instances for both classes. A high F1-score (0.9751 for both Covid and Normal) indicated a good balance between precision and recall for both classes. There are 362 instances for both Covid and Normal classes. An accuracy of 0.9751 indicated that the ensemble model correctly predicted the class for approximately 97.51% of the instances. The macro average for precision, recall, and F1-score is approximately 0.9751, indicating balanced performance across classes. In this case, the weighted average for precision, recall, and F1-score was approximately 0.9751, indicating balanced performance considering the class distribution. Overall, the classification report suggested that the ensemble model, which combined predictions from multiple individual models (Decision Tree, Random Forest, Support Vector Machine, and AdaBoost), achieved exceptional performance in distinguishing between Covid and Normal classes, with high precision, recall, F1-score, and accuracy. The confusion matrix is as shown in Fig. 10 and classification table in Table 6.

Table 6 Classification report for Ensemble Model

	precision	Recall	F1-score	Support
Covid	0.9751	0.9751	0.9751	362
Normal	0. 9751	0.9751	0.9751	362
Accuracy			0.9751	724
Macro avg	0. 9751	0.9751	0.9751	724
Weighted avg	0. 9751	0.9751	0.9751	724

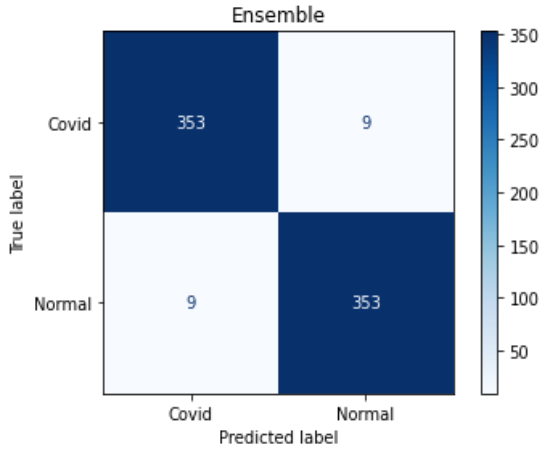

Fig. 10 Confusion Matrix for Ensemble

4.6 COVID-19 Diagnosis System

The COVID-19 Diagnosis System is a software application designed using flask API for medical practitioners to upload chest X-ray images of patients. Using machine learning techniques, the system can predict whether the image indicates COVID-19 infection or shows a normal condition. Upon image upload, the system processes the image, generates a pie chart displaying the likelihood of COVID-19 or normalcy, and allows medical practitioners to interpret the results for informed decision-making. The system prioritizes user-friendliness, security, and compliance with healthcare regulations to ensure effective and confidential diagnosis assistance. A screenshot of the system is shown in Fig. 11 and sample X-ray image uploaded into the system.

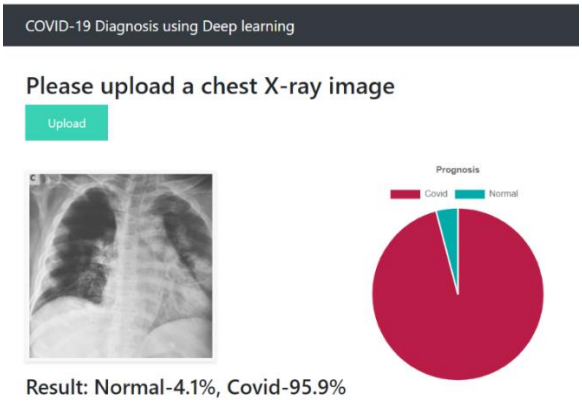

Fig. 11 Screenshot of the Proposed Covid-19 Diagnosis System

CONCLUSION

This study provides a comprehensive evaluation of a deep learning-based COVID-19 diagnosis software, employing convolutional neural networks (CNNs) and ensemble machine learning techniques. One of the main issues with prediction of Covid-19 using machine learning models was the failure to provide medical personnel with a simple and easy to use system that can quickly identity covid cases. The system shed light on the efficacy and potential of utilizing advanced machine learning algorithms in medical diagnostics. One of the key observations from this study is the remarkable accuracy achieved by the software in distinguishing between COVID-19 and normal cases. The CNNs demonstrated exceptional capability in capturing intricate patterns and features indicative of COVID-19 infection. Furthermore, the ensemble machine learning approach, combining multiple models (Decision Trees, Random Forests, Support Vector Machines, and AdaBoost) yielded robust and reliable predictions. The synergy among these models contributed to improved diagnostic accuracy and robustness. However, it is essential to acknowledge certain limitations and areas for future exploration. While the current study focused on CNNs and ensemble machine learning, there exist alternative approaches and techniques that warrant investigation. For instance, future research could explore the integration of additional deep learning architectures, such as recurrent neural networks (RNNs) or transformer models, to further enhance diagnostic performance. Additionally, the incorporation of advanced feature engineering techniques and data

augmentation strategies may contribute to better generalization and scalability of the diagnostic software.

Funding There was no outside funding for the study.
Data availability the data that support the findings of this study are openly available in COVID-QU-Ex Dataset Retrieved from https://www.kaggle.com/datasets/anasmohammedtahir/covidqu

REFERENCES

[1] Kohli A, Hande PC, Chugh S. (2021). Role of chest radiography in the management of COVID-19 pneumonia: An overview and correlation with pathophysiologic changes. Indian Journal of Radiology Imaging. doi: 10.4103/ijri.IJRI_967_20.

[2] Sarki R, Ahmed K, Wang H, Zhang Y, Wang K. (2022). Automated detection of COVID-19 through convolutional neural network using chest x-ray images. PLoS One. 2022 Jan 21;17(1):e0262052. doi: 10.1371/journal.pone.0262052.

[3] LeCun, Yann & Bengio, Y. & Hinton, Geoffrey. (2015). Deep Learning. Nature. 521. 436-44. 10.1038/nature14539.

[4] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, (2016). A. Learning Deep Features for Discriminative Localization, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 2921-2929, doi: 10.1109/CVPR.2016.319.

[5] Dietterich, T.G. (2000). Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

[6] *Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556.*

[7] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D, Bagul, A, Langlotz, C., Shpanskaya, K., Lungren, M.P., Chexnet (2017). Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225

[8] Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., & Muller, K.R. (2021). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Journal Proceedings of the IEEE, Volume 109, Number 3, Page 247. DOI: 10.1109/JPROC.2021.3060483

[9] Zeiler, M.D., Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham. https://doi.org/10.1007/978-3-319-10590-1_53

[10] COVID-QU-Ex Dataset Retrieved from https://www.kaggle.com/datasets/anasmohammedtahir/covidqu

[11] Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., & Tang, Z. (2020). Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19. IEEE Reviews in Biomedical Engineering, 14, 4-15.

[12] Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, Cai M, Yang J, Li Y, Meng X, Xu B. (2021). A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). Eur Radiol. Aug;31(8):6096-6104. doi: 10.1007/s00330-021-07715-1. Epub 2021 Feb 24. PMID: 33629156; PMCID: PMC7904034.

[13] Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, Bai J, Lu Y, Fang Z, Song Q, Cao K, Liu D, Wang G, Xu Q, Fang X, Zhang S, Xia J, Xia J. (2020). Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. Radiology. 2020 Aug;296(2):E65-E71. doi: 10.1148/radiol.2020200905. Epub 2020 Mar 19. PMID: 32191588; PMCID: PMC7233473.

[14] Abbas A., Abdelsamea M.M., Gaber M.M. (2021). Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. Appl Intell (Dordr). 2021;51(2):854-864. doi: 10.1007/s10489-020-

01829-7. Epub 2020 Sep 5. PMID: 34764548; PMCID: PMC7474514.

[15] Mohanty S, Harun Ai Rashid M, Mridul M, Mohanty C, Swayamsiddha S. Application of Artificial Intelligence in COVID-19 drug repurposing. Diabetes Metab Syndr. 2020 Sep-Oct;14(5):1027-1031. doi: 10.1016/j.dsx.2020.06.068. Epub 2020 Jul 3. PMID: 32634717; PMCID: PMC7332938.

[16] KC, K., Yin, Z., Wu, M., Zhilu Wu. *(2021).* Evaluation of deep learning-based approaches for COVID-19 classification based on chest X-ray images. *SIViP* 15, 959–966 (2021). https://doi.org/10.1007/s11760-020-01820-2

[17] Apostolopoulos I.D., Mpesiana T.A. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med. 2020 Jun;43(2):635-640. doi: 10.1007/s13246-020-00865-4. Epub 2020 Apr 3. PMID: 32524445; PMCID: PMC7118364.

[18] Renuka D.S.M., Rose, S.B., Akshitha, S., Niharika, P. Comparison of COVID-19 Diagnosis by CNN Model and ResNet Using Chest X-Ray," *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, Theni, India, 2023, pp. 1569-1574, doi: 10.1109/ICSCNA58489.2023.10370248.

[19] Indra, Z., Elfizar, Bahri, Z. and Alfirman. (2023). Covid-19 Early Diagnosis Based on Transfer Learning and Modified CNN Architecture, *2023 Sixth International Conference on Vocational Education and Electrical Engineering (ICVEE)*, Surabaya, Indonesia, 2023, pp. 150-155, doi: 10.1109/ICVEE59738.2023.10348185.

[20] Kundu, A., Mishra, C. & Bilgaiyan, S. (2021). COVID-SEGNET: Diagnosis of Covid-19 Cases on Radiological Images using Mask R-CNN," *2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII)*, Chennai, India, 2021, pp. 1-5, doi: 10.1109/ICBSII51839.2021.9445190.

[21] Carvalho, E. D., Carvalho, E. D., De Carvalho Filho, A. O., de Araújo, F. H. D. & Andrade Lira Rabêlo, R. D. (2020). Diagnosis of COVID-19 in CT image using CNN and XGBoost. *2020 IEEE Symposium on Computers and Communications (ISCC)*, Rennes, France, 2020, pp. 1-6, doi: 10.1109/ISCC50000.2020.9219726.