Information Dynamics and Applications
https://www.acadlore.com/journals/IDA

Enhanced Detection of COVID-19 in Chest X-ray Images: A Comparative Analysis of CNNs and the DL+ Ensemble Technique

Bwanali Haji Ntaibu Jereni[1*], Iota Sundire[2]

[1] Adjunct Lecturer, Department of Medicine, University of Botswana, 0061 Gaborone, Botswana
[2] Fuels and Energy Department, Midlands State University, ZW170407 Gweru, Zimbabwe

* Correspondence: Bwanali Haji Ntaibu Jereni (202006488@ub.ac.bw)

Received: 10-10-2023 **Revised:** 11-18-2023 **Accepted:** 11-23-2023

Citation: B. H. N. Jereni and I. Sundire, "Enhanced detection of COVID-19 in chest X-ray images: A comparative analysis of CNNs and the DL+ ensemble technique," *Inf. Dyn. Appl.*, vol. 2, no. 4, pp. 186–199, 2023. https://doi.org/10.56578/ida020403.

Abstract: The swift global spread of Corona Virus Disease 2019 (COVID-19), identified merely four months prior, necessitates rapid and precise diagnostic methods. Currently, the diagnosis largely depends on computed tomography (CT) image interpretation by medical professionals, a process susceptible to human error. This research delves into the utility of Convolutional Neural Networks (CNNs) in automating the classification of COVID-19 from medical images. An exhaustive evaluation and comparison of prominent CNN architectures, namely Visual Geometry Group (VGG), Residual Network (ResNet), MobileNet, Inception, and Xception, are conducted. Furthermore, the study investigates ensemble approaches to harness the combined strengths of these models. Findings demonstrate the distinct advantage of ensemble models, with the novel deep learning (DL)+ ensemble technique notably surpassing the accuracy, precision, recall, and F-score of individual CNNs, achieving an exceptional rate of 99.5%. This remarkable performance accentuates the transformative potential of CNNs in COVID-19 diagnostics. The significance of this advancement lies not only in its reliability and automated nature, surpassing traditional, subjective human interpretation but also in its contribution to accelerating the diagnostic process. This acceleration is pivotal for the effective implementation of containment and mitigation strategies against the pandemic. The abstract delineates the methodological choices, highlights the unparalleled efficacy of the DL+ ensemble technique, and underscores the far-reaching implications of employing CNNs for COVID-19 detection.

Keywords: Computed tomography images; Diagnostic methods; Ensemble techniques; Visual Geometry Group (VGG); Residual Network (ResNet); MobileNet; Inception; Xception; Deep learning+ ensemble technique

1 Introduction

Coronaviruses, a diverse family of viruses, are recognized for their capacity to instigate respiratory infections in humans and animals. In December 2019, a novel coronavirus, designated SARS-CoV-2, emerged in Wuhan, China. This virus rapidly disseminated globally, precipitating a widespread pandemic of COVID-19 [1]. Characterized as RNA viruses, coronaviruses possess the most extensive viral RNA genome known, predominantly hosted by bats, yet capable of zoonotic transmission to humans. As of August 24, 2020, over 23 million cases of coronavirus have been documented worldwide, resulting in approximately 800,000 fatalities. Notably, about five million individuals have recuperated, with the United States, Brazil, India, and Russia reporting the highest incidence rates [2].

Diagnosis of COVID-19 necessitates the identification of the SARS-CoV-2 virus. Current diagnostic approaches bifurcate into laboratory-based and point-of-care methodologies. Laboratory-based diagnostics, though more precise, demand specialized apparatus and skilled personnel. These encompass nucleic acid testing (NAT), antigen testing, and serology testing [3]. NAT is renowned for its high accuracy in detecting SARS-CoV-2 from nasal or throat samples but suffers from protracted result turnaround times. In contrast, antigen tests, while faster, offer diminished accuracy. Serology tests, detecting antibodies against SARS-CoV-2, can indicate past infections but are not viable for diagnosing active cases. Point-of-care diagnostics, though expedient and more user-friendly, compromise on accuracy. These include rapid antigen tests and lateral flow immunoassays [3]. Predominant limitations of current COVID-19 diagnostics encompass cost, speed, accuracy, and accessibility, particularly in resource-limited settings. In the domain of image processing, CNNs, a subset of artificial intelligence (AI), have gained prominence. Their

application in medical imaging for tasks such as image segmentation and target recognition is well-documented [4]. CNNs have demonstrated efficacy in classifying CT images, including those of COVID-19 patients. The proposed methodology in this study aims to mitigate the constraints of existing COVID-19 diagnostics by developing a CNN-based rapid, accurate, and cost-effective test, suitable for point-of-care settings such as clinics and pharmacies [5]. Advantages of CNN-based diagnostics include affordability, rapidity, and potentially equal or greater accuracy compared to laboratory-based tests, thereby enhancing accessibility.

The advent of CNN-based diagnostic methods could substantially influence the battle against the pandemic. By offering quicker, more precise, and affordable testing solutions, these methods promise to enhance testing efficiency and accessibility across diverse communities. Consequently, this could contribute to a reduction in COVID-19 transmission and associated mortalities [6].

1.1 Deep CNN

In recent advancements, CNNs have been recognized as the most extensively researched machine learning methodologies for the diagnosis of medical conditions through imaging. The efficacy of CNNs is attributed to their capability to retain complex features while scanning input data. This characteristic is particularly crucial in radiology, where spatial relationships, such as the interfaces between bones and muscles or the transition from healthy to diseased lung tissue, are pivotal. As depicted in Figure 1, the architecture of the system under study is detailed.

Each chest image, presented as a tensor of dimensions 244×244, is processed through a CNN structure comprising five convolutional layers. In the initial convolutional layer, 553 kernel filters are utilized with a stride of one, generating a maximum of 64 filters. This is followed by a max-pooling layer, which receives the output of the first layer and reduces the input dimensions by half to 112×112, employing a stride of two. The outcome of this pooling layer undergoes the Rectified Linear Unit (ReLU) activation function across all levels [7]. The processed nonlinear data then enters the subsequent convolutional layer, equipped with 55×64×128 filters and maintaining the same stride value. This output is again subject to max-pooling with identical strides of 2×2, further reducing dimensions to 56×56.

After ReLU activation, the output proceeds to the third convolutional layer, which houses 256 filters with a 5×5×128 kernel size and a stride of 1×1. The resulting output is channelled to a max-pooling layer, yielding a tensor of 28×28 dimensions. Post-ReLU activation, the signal enters the fourth convolutional layer, consisting of 512 filters with a 5×5×256 kernel size and a 1×1 stride. The fourth convolutional operation's output undergoes max-pooling, diminishing its dimensions to 14×14. Following ReLU activation, this data feeds into the fifth convolutional layer, which features 512 filters with a kernel size designed to accommodate the output from the preceding layers. This layer's output is then subjected to max-pooling with a 2×2 stride, maintaining an output size of 14×14. The resultant tensor assumes a form of 7×7×512. A compression of this tensor results in 25,088 neurons. The weighted values produced by these neurons are indicative of their correlation with COVID-19 symptoms. To prevent system overfitting, a dropout layer is employed, selectively omitting information.

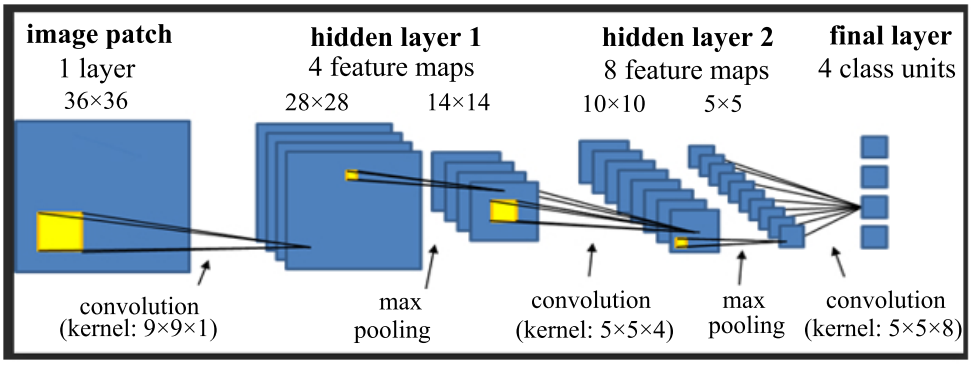

Figure 1. Deep CNN [7]

• Convolutional layer: This layer incorporates numerous filters, also known as kernels, each of which is applied to the input image to extract features and construct a new layer. These layers encapsulate key characteristics of the input image. The convolution operation, denoted by the symbol *, is a mathematical process where an input *In (t)* is convolved with a kernel *f(a)*, resulting in a feature map *F(t)* represented as:

$$\mathrm{F(t)} = (\mathrm{I_n}^{*}\mathrm{f})\,(\mathrm{t}) \tag{1}$$

For discrete convolution, where *t* is restricted to integer values, the process is defined as:

$$\mathrm{F(t)} = \sum_{n} \quad I_{\mathrm{n}}(\mathrm{a}).\mathrm{f}(\mathrm{t}-\mathrm{a}) \tag{2}$$

In the context of CNNs as applied in this study, a key aspect of the methodology involves the implementation of a two-dimensional convolution procedure. This process entails the application of an input matrix, denoted as $in(m, n)$, and a convolution kernel, represented as $f(a, b)$. The convolution operation is defined mathematically as:

$$\mathrm{F(t)} = \sum_{a}\sum_{b} I_{\mathrm{n}}(\mathrm{a},\mathrm{b}).\mathrm{f}(\mathrm{m}-\mathrm{a},\mathrm{n}-\mathrm{b}) \tag{3}$$

It is pertinent to note that the convolution operation adheres to the commutative law. Consequently, this law allows for the reversal of the kernel, rendering an equivalent expression as:

$$\mathrm{F(t)} = \sum_{a}\sum_{b} I_{\mathrm{n}}(\mathrm{m}-\mathrm{a},\mathrm{n}-\mathrm{b}).\mathrm{f}(\mathrm{a},\mathrm{b}) \tag{4}$$

While convolution typically involves flipping the kernel, neural networks often use the cross-correlation formula, which is similar to convolution but does not involve this flipping. In this case, the operation is defined as:

$$\mathrm{F(t)} = \sum_{a}\sum_{b} I_{\mathrm{n}}(\mathrm{m}+\mathrm{a},\mathrm{n}+\mathrm{b}) \cdot \mathrm{f}(\mathrm{a},\mathrm{b}) \tag{5}$$

• ReLU layer: The ReLU layer serves as an activation function, turning negative input values to zero. This operation, mathematically expressed as the following Eq. (6), optimizes learning by accelerating convergence and mitigating the issue of gradient vanishing.

$$R(x) = \max(0, x) \tag{6}$$

• Max pooling layer: Employing a sample-based discretization process, this layer aims to reduce the spatial dimensions of the input (such as an image or output from previous layers), thereby lowering the number of parameters and computation in the network. A typical implementation in this study used a kernel size of 33%, reducing the dimensions of the output from the last convolutional block.

• Batch normalization: This technique normalizes the output from the previous layer by subtracting the batch mean and dividing by the batch standard deviation. It enhances network stability and allows each layer to learn more independently from others.

• Fully connected layer: This layer uses the output from the previous layer to create a probability function for classification into different categories.

• Loss function: At this stage, the input data sample is subjected to a softmax function. This level is critical for making the final prediction. The loss function, essential for model training, is formulated accordingly.

$$\mathrm{L_i} = -\log\left(\frac{e\beta\gamma}{\sum_{j}^{c} e\beta j}\right) \tag{7}$$

• Regularization: To prevent overfitting, a dropout technique, as proposed by Srivastava et al. [8], is employed. During training, this involves randomly setting a proportion of neurons to zero, effectively thinning the network.

1.2 Transfer Learning

Transfer learning has emerged as a pivotal technique in specialized fields where acquiring extensive, high-quality data poses a challenge. In such contexts, the transfer of knowledge from a source to a target task often becomes a vital strategy [9]. This method leverages pre-trained models, optimizing for categorical sensitivities. It is observed that the initial layers of a CNN are trained to discern common features such as edges, textures, and shapes, while the deeper layers are adept at identifying more complex and unique characteristics of the image, such as pathological lesions.

In the field of Computer-Aided Diagnosis (CAD), a prevalent approach involves training only the top layer of the model with the target dataset, while utilizing the pre-initialized values of the base layers. This methodology reduces the likelihood of overfitting, a significant concern in neural network training cycles, particularly when limited by the number of model parameters. In this study, S. Pal and K. Simonyan utilized the ImageNet database, incorporating the VGG16, InceptionV3, and Xception models equipped with pre-trained weights. The ImageNet database, structured based on the WorldNet hierarchy, comprises over 3.2 million meticulously annotated images across 5,247 categories [10].

1.2.1 VGG16

The VGG16 architecture, also known as VGGNet, is a pre-trained deep CNN proficient in extracting visual features for class differentiation, thereby enhancing outcome accuracy. This architecture includes 16 convolutional layers with a significant receptive field of 3×3 and five max-pooling layers, each of size 2×2, for spatial pooling [11]. The model's proficiency in distinguishing visual differences between images is attributed to its depth. It also incorporates three fully connected layers, with the terminal layer being a softmax layer. ReLU activation functions are applied across all hidden units, and dropout regularization is integrated into the fully connected layers [10]. When the densely interconnected classifier is detached from the pre-trained VGG16 model, it can serve as an effective feature vector generator. The VGG16 architecture was employed as a pre-trained classifier, with SoftMax utilized for classification.

1.2.2 InceptionV3

The 'Inception' micro-architecture, introduced by Szegedy et al. [12], represents a deep CNN that utilizes varied filters for critical feature identification within images. Functioning as a multi-level feature extractor, the Inception model calculates 1×1, 3×3, and 5×5 convolutional layers within a unified system. The input undergoes processing through these filters, with the outcomes being concatenated along the channel dimension before proceeding to the next layer. The InceptionV3 architecture, as delineated in "Rethinking the Inception Architecture for Computer Vision" by Szegedy et al. [12], was employed. This version evaluated the Inception-v3 classifier using images from 1000 labeled classes in the ImageNet benchmark datasets, comprising both a CNN for feature extraction and fully connected & softmax layers for classification.

1.2.3 Xception

The Xception architecture, introduced by Chollet [13], represents a significant advancement in deep learning. It is predicated on depth-wise separable convolution, distinguishing it from conventional CNNs. Comprising 36 layers, this model excels in feature extraction due to its unique architecture. The publicly available version of Xception, developed using Keras and TensorFlow under the MIT license, includes these 36 convolutional layers, with exceptions only at the beginning and end of the network. In performance comparisons on the ImageNet database, Xception has demonstrated superior results over models like InceptionV3, various ResNet architectures, and VGG. These models have shown exceptional efficacy in classifying medical images, indicating their potential in aiding COVID-19 detection.

In the realm of medical imaging, CNNs have emerged as invaluable tools, particularly in the rapid and efficient analysis of large volumes of images. The utility of CNNs lies in their capability to discern critical features within medical images, a function crucial for accurate diagnosis. However, it is imperative to select the appropriate type of CNN for specific diagnostic tasks, as different architectures possess unique strengths and limitations. This study undertook a comprehensive evaluation of various CNN models to determine their efficacy in detecting COVID-19 from medical images. The ImageNet dataset was employed to provide these models with an initial learning base, leveraging its extensive collection of images. This approach ensures the models are robust and effective in their diagnostic capabilities. The performance of these models was rigorously assessed using a range of metrics to confirm their accuracy and reliability. The structure of the study is organized as follows: The subsequent section elucidates the related work undertaken for COVID-19 identification, along with the objectives and approach of this study. Section 3 delineates the CNN topologies utilized, while Section 4 presents the results of simulations comparing the proposed models. Finally, Section 5 encapsulates the findings of this study.

Recent studies have extensively explored COVID-19 detection using machine learning methodologies. A notable challenge faced by researchers, attributable to the scarcity of specialized databases, has been the application of CNN algorithms to CT images [14]. Although various studies have presented outcomes in tabular formats, direct comparisons often prove challenging due to the heterogeneity and varying complexity of the utilized databases. This section further details the endeavours related to COVID-19 detection on CT images, emphasizing the employed methodologies.

Anu et al. (2021) implemented a deep ensemble-based technique for detecting COVID-19 related fake news, incorporating Support Vector Machine (SVM), dense neural networks, and CNN within the ensemble classifier. The study conducted exhaustive testing using character and word n-gram term frequency–inverse document frequency (TF-IDF) features, comparing the ensemble architecture against eight traditional machine learning classifiers. The findings indicated that character n-gram features were superior to word n-gram features, with the ensemble classifier achieving a weighted F1-score of 0.97.

Table 1. Comparative analysis of several surveyed studies and their respective methodologies

Author Name	Year	Method	Parameter Used	Dataset Used	Weaknesses	Strengths/Conclusions
Muhammad et al.	2020	DT, SVM, NB, LR, RF, K-NN	Accuracy	Epidemiological database of COVID-19 patients in South Korea	The absence of data points reduces predictive accuracy and can lead to biased conclusions.	The DT model exhibited the highest accuracy at 99.85%, followed by RF (99.60%), SVM (98.85%), K $-$ NN(98.06%), NB(97.52 %), and LR (97.49 %). These models are instrumental in advancing healthcare strategies against COVID-19.
Anu et al.	2021	SVM, DNN, and CNN.	Precision, Recall, ROC-AUC, F1 score	COVID-19-related fake news	The omission of character-level features in detecting COVID-19 fake news warrants further investigation.	The study demonstrates superior performance using character-level features over word-level features with the classifiers.
Ali et al.	2020	DeepLabV3+	F1 Score	Publicly available dataset provided by Shenzhen Hospital, that contains 566 CRs with manually segmented lungs (ground truth)	The lack of automated lung segmentation hinders efficiency in medical imaging.	The model achieves an Intersection-Over-Union (IoU, Jaccard Index) score of 0.97 on the test set, illustrating its effectiveness.
Saha et al.	2020	Several traditional CNN architectures are tested and finally in the ensemble operation, MobileNet, InceptionV3, DenseNet201, DenseNet121 and Xception are used.	Accuracy	The Kaggle dataset is composed of 1,583 nomal and 4273 pneumonia images of pediatric patients from Guangzhou Women and Children's Medical Center, Guangzhou	Exploration of other models and ensemble techniques with a more extensive dataset is needed.	The model achieves 96% accuracy in 3-class (COVID-19 inormal/pneumonia) diagnosis and 89.21% in 4 -class (COVID-19/normal/viral pneumonia/bacterial pneumonia) diagnosis.
Tewari et al.	2020	VGG-16, ResNet-50 and MobileNetV2	Accuracy, precision, recall and F1-Score	The final augmented dataset consists a total of 7,585 images with varying resolutions in which 2,255 are of COVID- $19, 2, 614$ are of viral pneumonia and 2,716 are of normal category.	Limitations in training and evaluating the model on a larger dataset.	The proposed model attained an overall accuracy of 96.34%. For the COVID- 19 class, precision, recall, and F1-Score were recorded at $100\%, 96\%$, and 98%, respectively.
Deep Deb et al.	2020	DCNN	Accuracy	The chest X-ray dataset used for our experimentation. The same was acquired on 17 April 2020	Slight delay in processing is acceptable for achieving higher accuracy.	An accuracy of 91.99% was achieved, marginally surpassing current state-of-the-art performances.

Ali et al. [15] showcased the application of a lung segmentation ensemble deep network, based on an advanced iteration of DeepLabV3, known as DeepLabV3+. This system utilized diverse topologies including ResNet18, ResNet50, Mobilenetv2, Xception, and inceptionresnetv2. Enhancements were made to the spatial pyramid pooling's receptive field within the DeepLabV3+ encoder module. The method underwent testing on a publicly available dataset from Shenzhen Hospital, comprising 566 chest radiographs with manually segmented lungs, achieving an Intersection-Over-Union (IoU) score of 0.97 on the test dataset. In recent advancements in AI-assisted medical diagnostics, notable strides have been made in the application of CNNs for COVID-19 detection. Saha et al. [16] developed a novel approach that integrates multiple CNN architectures in an ensemble framework. This methodology encompassed two distinct strategies: feature-level fusion and decision-level ensemble techniques. Prior to the assembly phase, several standard CNN designs, including MobileNet, InceptionV3, DenseNet201, DenseNet121, and Inception, were evaluated. The transfer learning approach was implemented, utilizing ImageNet pre-trained weights to manage the computational demands of multiple networks. The convolutional feature maps from various layers were globally averaged, followed by passage through fully connected layers, facilitating joint optimization in feature-level ensemble technique. Additionally, the decision-level ensemble method employed majority voting to combine final predictions from multiple networks. Despite the integration of various techniques, neither strategy exceeded the performance of individual models. Nonetheless, testing on the COVID-CT database demonstrated remarkable results, with a 96% accuracy in 3-class identification (COVID-19/normal/pneumonia) and 89.21% accuracy in 4-class identification [16]. Tewari et al. [17] explored a deep learning approach incorporating fuzzy image enhancement, offline data augmentation, image segmentation, and CNN-based classification for detecting COVID-19 in chest X-ray images. The proposed model integrated features from VGG-16, ResNet-50, and MobileNetV2, achieving an overall accuracy of 96.34%. Notably, the precision, recall, and F1-score for the COVID-19 class were 100%, 96%, and 98%, respectively. Deb et al. [18] formulated an ensemble architecture based on deep CNN for feature extraction from chest X-ray images, categorizing them into three classes: CAP, healthy, and COVID-19. The ensemble system incorporated NASNet, MobileNet, and DenseNet, combining low-level features extracted from these frameworks for classification. This method yielded a precision of 91.99%, marginally surpassing existing state-of-the-art performances. Muhammad et al. [19] utilized epidemiological databases of COVID-19 patients in South Korea to develop data mining techniques for predicting patient recovery. The implementation involved various methods, including decision trees, SVM, Naïve Bayes, logistic regression, random forests, and K-nearest neighbors, using Python programming. The decision tree-based model exhibited the highest level of effectiveness with an average accuracy of 99.85%, followed by random forests, SVM, K-nearest neighbors, Naïve Bayes, and logistic regression in descending order of accuracy. These findings hold substantial promise in enhancing healthcare strategies against COVID-19.

The COVID-19 pandemic has emerged as a formidable global health crisis, exerting unprecedented strain on societal and healthcare frameworks. This is attributed to the escalating rates of infection and mortality. Early identification of patients is a critical intervention in mitigating the pandemic's impact and alleviating the burden on healthcare systems. The delayed diagnostic process has been identified as a key factor in the rapid spread of COVID-19. Acceleration of the diagnostic procedure can be facilitated through imaging techniques such as chest X-rays. Numerous studies have been conducted with the aim of enhancing the speed and accuracy of COVID-19 detection using imaging. A prominent approach in recent research is the application of machine learning techniques for classification purposes. An emerging method involves the utilization of deep transfer learning algorithms as classifiers. Despite the effectiveness of these models, detailed analysis reveals potential areas for improvement. A notable deficiency in the current systems is the lack of feature extraction or preprocessing techniques, which are critical in analyzing X-ray images where textures and patterns constitute significant diagnostic information. Consequently, the objectives of this study are outlined as follows:

- To develop and implement a model for feature extraction based on Principal Component Analysis (PCA).
- To devise and apply a method for COVID-19 detection using chest X-rays, incorporating data preprocessing and augmentation.
- To enhance accuracy through the implementation of advanced deep learning algorithms, employing multiple transfer learning techniques, and ensemble and novel ensemble classifiers.
- To conduct a comprehensive analysis and comparison of the proposed model with existing methodologies.

Several surveyed studies and their respective methodologies are compared in Table 1.

2 Methodology

The methodology adopted for this study encompasses distinct phases, as illustrated in Figure 2.

Phase 1: Data collection

Data was sourced from online repositories containing key attributes such as age, gender, and symptoms (cough, sore throat, shortness of breath, fever, headache), along with additional data regarding contact with COVID-19 patients. The dataset includes:

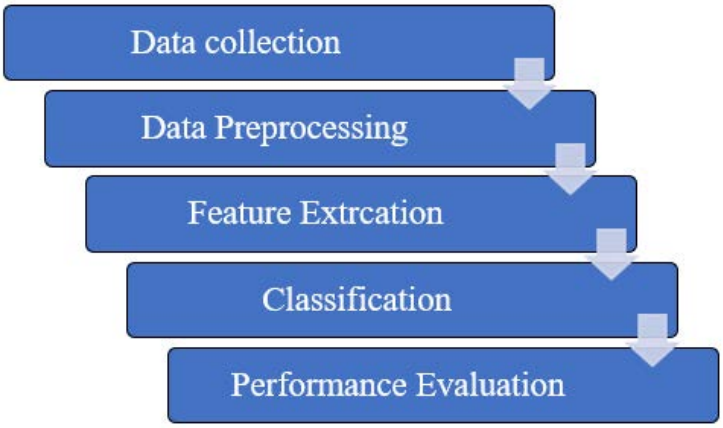

Figure 2. Different phases of the methodology

Basic information:gender (male/female), age $\geq$ 60 years (yes/no);
Symptoms: cough (yes/no), fever (yes/no), sore throat (yes/no), shortness of breath (yes/no), headache (yes/no);
Additional data: confirmed interaction with a COVID-19 positive individual (true/false).

Phase 2: Data preprocessing or cleaning

The raw data obtained required preprocessing and cleaning. This phase involved a series of filtration processes, such as replacing missing values, eliminating entries without a class index, converting data types, and normalizing the dataset. In instances where attribute values were absent, a missing value filter was applied, substituting the missing entries with the attribute's average value. Furthermore, PCA was utilized for feature extraction, aiming to distill the data into a more concise and informative representation. PCA facilitates the identification of patterns and reduces dimensionality while preserving critical information and eliminating redundant features. This procedure is expected to augment the efficiency of the subsequent machine learning models by concentrating on the most pertinent aspects of the input data.

Phase 3: Feature selection

Feature selection involves the meticulous selection of pivotal attributes from a dataset, emphasizing those crucial for the specific objective. This process eliminates superfluous or irrelevant features, focusing on detailed selection, attribute selection, or variable selection. Features with a broad range of values are typically preferred, though care is taken to avoid overfitting. In this study, Random Forest was employed for feature selection due to its proficiency in handling diverse features and resilience against noisy data. Random rorest, an ensemble machine learning method, constructs numerous decision trees during training and outputs the mode of the classes for classification purposes. Its ability to handle complex data relationships and mitigate overfitting makes it an ideal choice for feature selection, classification, and regression tasks. Random forest's versatility in handling various data types enhances the performance of the classification model by identifying and prioritizing the most influential features.

Phase 4: Implementation of hybrid classification algorithm

In this study, a hybrid classification algorithm, integrating random forest and gradient boosting, was employed. This amalgamation of deep learning predictions with ensemble classification forecasts is designed to enhance predictive accuracy, surpassing the capabilities of individual models. In sentiment analysis, where classification is crucial, the use of such a combination is advantageous, as each method compensates for the limitations of the other. A voting mechanism is incorporated to refine the model's architecture further. Despite the increased complexity associated with using multiple algorithms, the primary objective is the development of a model with superior performance. The most effective algorithm is determined through an average probability phase, ensuring a robust classification approach.

Phase 5: Model evaluation techniques

Various model evaluation strategies were considered, with four options provided by the Weka machine learning workbench selected for this study:

a) Percentage split: The dataset is randomly divided into training and testing partitions. This approach offers a rapid performance approximation and is recommended for large datasets.

b) Cross-validation: The data is divided into k-folds, with each fold serving as a test set in turn, while the model is trained on the remaining folds. This method is a standard for performance assessment, though it requires the generation of multiple model variants.

c) Training dataset: The model is trained and tested on the same dataset. This approach can be misleading as a perfect algorithm might memorize training patterns, resulting in inflated performance metrics.

d) Provided test set: The dataset is manually split into training and testing sets. The model is trained on the entire training set and evaluated on a separate test set. This method is suitable for datasets with a large number of instances.

Phase 6: Performance metrics

The performance of the two classification algorithms was compared using metrics such as accuracy, recall, precision, F-score, and error rate. The algorithm exhibiting the highest accuracy, recall, precision, F-measure, and the lowest error rate was deemed superior. Performance parameters are calculated using the following terms:

TP (True Positive): Correctly classified positive instance.

TN (True Negative): Correctly classified negative instance.

FP (False Positive): Incorrectly classified positive instance.

FN (False Negative): Incorrectly classified negative instance.

These terms are incorporated into the confusion matrix for analytical purposes. A confusion matrix, also known as an error matrix, displays the TP, FN, TP, and TN values in a two-row, two-column table. Table 2 illustrates the confusion matrix utilized in this study for machine learning purposes.

Table 2. Confusion matrix for machine learning

		Correct Labels	
		Positive	Negative
Classified Labels	Positive	TP	FP
	Negative	FN	TN

2.1 Precision and Recall

Precision and recall are critical metrics employed in the evaluation of performance in fields such as text mining and information extraction. These metrics are instrumental in assessing the exactitude and comprehensiveness of a dataset. Precision is defined as the ratio of true positive instances to the sum of true positive and false positive instances. It reflects the accuracy of positive classifications. Recall, on the other hand, is the ratio of true positive instances to the sum of true positive and false negative instances, indicating the completeness of the positive classifications.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{8}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{9}$$

2.2 F-Score

The F-score, is the harmonic mean of precision and recall. This metric provides a balanced measure that considers both the precision and recall of a classification system. The F-score is particularly useful when seeking an equilibrium between precision and recall, ensuring neither is disproportionately emphasized.

$$\text{F-measure} = \frac{2 * \text{recall} * \text{precision}}{\text{precision} + \text{recall}} \tag{10}$$

2.3 Accuracy

Accuracy stands as one of the most commonly utilized metrics for evaluating classification performance. It is calculated as the ratio of correctly classified instances (both true positive and true negative) to the total number of instances in the dataset. Conversely, the error rate is determined by the proportion of incorrectly classified instances (both false positive and false negative) to the total dataset.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} \tag{11}$$

In the aforementioned equations:

True Positive denotes instances where the prediction is positive, and the actual outcome is also positive.

True Negative represents instances where the prediction is negative, and the actual outcome is negative.

False Positive signifies instances where the prediction is positive, but the actual outcome is negative.

False Negative refers to instances where the prediction is negative, but the actual outcome is positive.

3 Result and Discussion

The integration of DL algorithms in advanced medical diagnostic systems, particularly for COVID-19 detection through medical image processing, is increasingly pivotal. The research methodology for COVID-19 detection in this study is characterized by three principal phases:

(a) Development of DL methods and ensemble procedure

• Three distinct DL methods, along with an ensemble procedure, were developed.

• Models 1 and 3 consist of three convolutional blocks, each succeeded by a max pooling layer, while Model 2 comprises four convolutional blocks with subsequent max pooling layers.

• Batch normalization follows the first convolutional block to accelerate the learning process of the CNN model.

• Dropout layers are incorporated to mitigate overfitting in deep CNN architectures.

• Each method includes fully connected layers to facilitate comprehensive data analysis.

(b) Validation using CT images

• CT images were utilized to validate the DL framework. The compilation of a balanced database was prioritized, encompassing CT scans of both COVID-19 and non-COVID pulmonary infections.

(c) Experimental evaluation

• The DL + ensemble technique underwent evaluation using the COVID-CT database, with a focus on smaller-scale databases for robust analysis.

• The fully connected layer, subsequent to the fusion layer, comprised 256 neurons with ReLU activation and parameter regularization.

• A dropout rate of 0.2 was applied post each fully connected layer.

• Accuracy, precision, recall, and F-score of the proposed 5-Clf, proposed 8-Clf, and DL with ensemble approach were compared, as depicted in Figure 3.

The results of this study underscore the efficacy of the proposed innovative ensemble method in detecting COVID-19 infection from chest CT images. Experimental analyses were conducted using the COVID-CT database, evaluating the performance of the suggested DL + ensemble method, particularly in the context of smaller-scale databases.

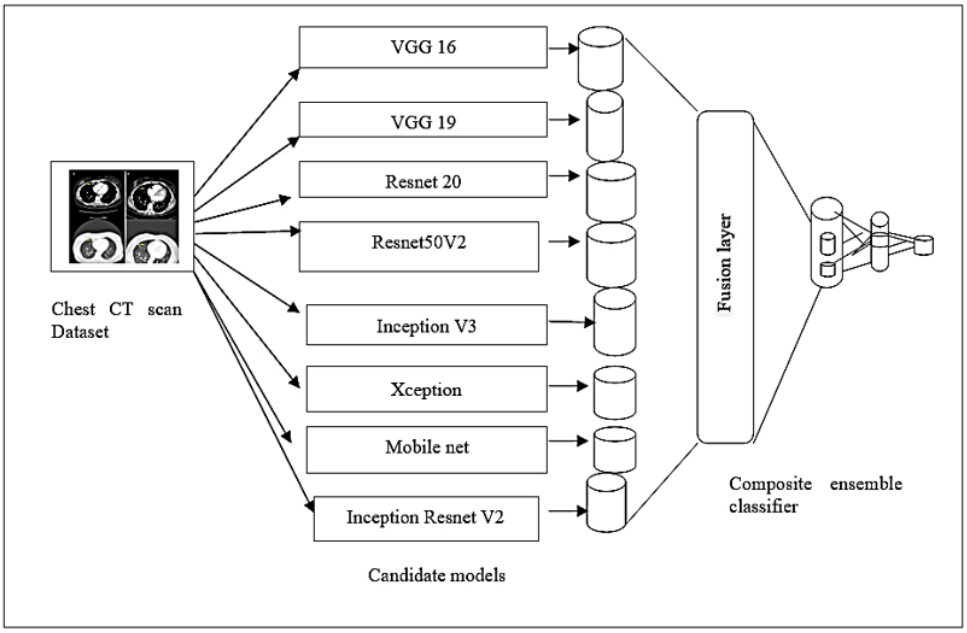

Figure 3. Framework of the proposed novel ensemble classifier

Each fully connected layer, following the fusion layer, contained 256 neurons with ReLU activation and regularization of parameters. Furthermore, a dropout rate of 0.2 was implemented following each fully connected layer. Figure 4 shows the training, validation of accuracy and loss in case of CNN design. Figure 5 and Figure 6 exhibit the training and validation of accuracy and loss for VGG-19, respectively.

Figures 4- 8 illustrate the training and validation of accuracy and loss for various CNN designs, including VGG-19 and ResNet-50. These figures depict the effectiveness of each model in the context of training and validation phases.

Figure 9 presents a comparative analysis of key performance metrics, namely, accuracy, precision, recall, and F-score, for the proposed 5-Clf, 8-Clf models, and the DL with ensemble approach.

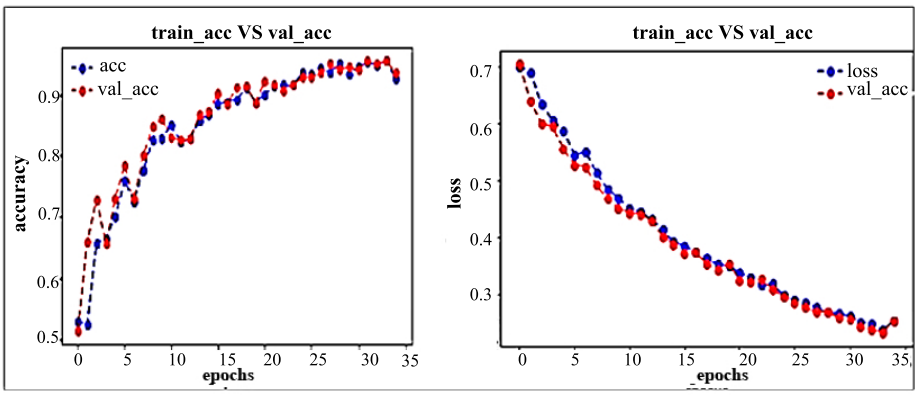

Figure 4. Training, validation of accuracy and loss (CNN design)

Figure 5. Training and validation of loss (VGG-19)

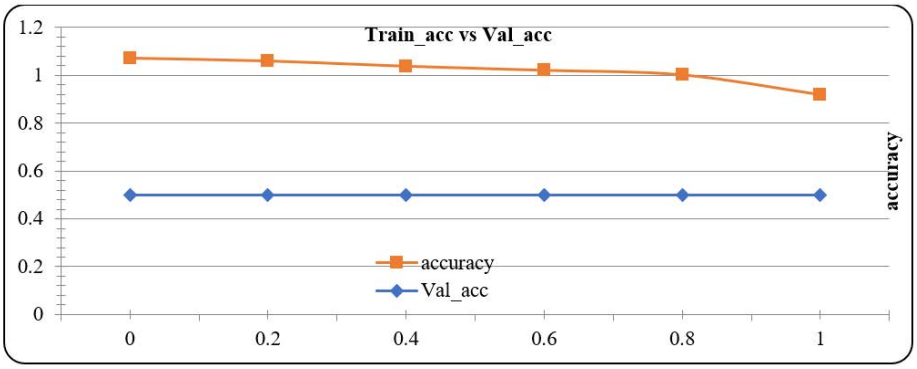

Figure 6. Training and validation of accuracy (VGG-19)

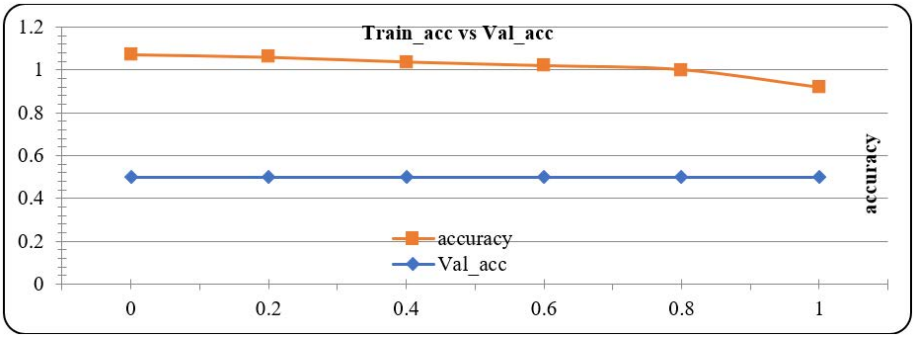

Figure 7. Training and validation of accuracy (ResNet 50)

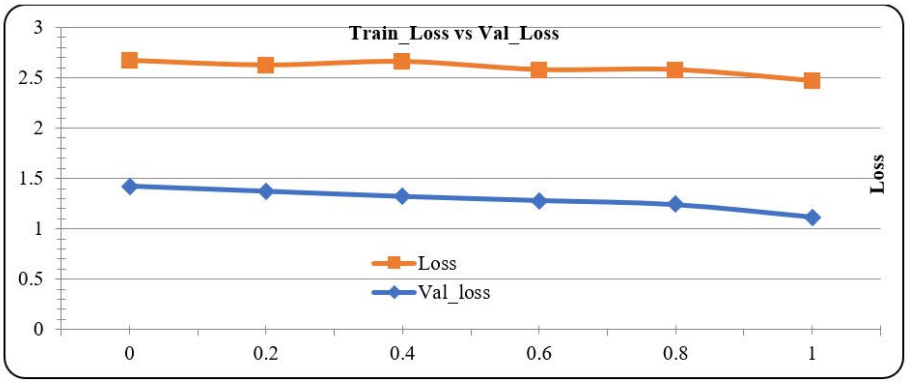

Figure 8. Training and validation of loss (ResNet 50)

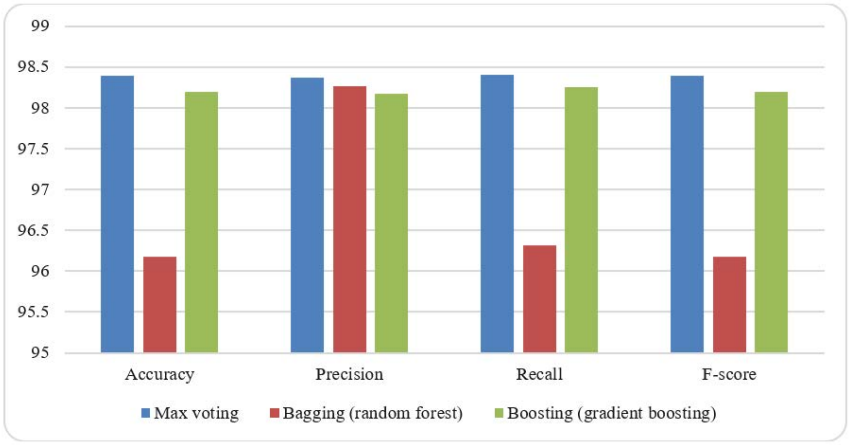

Figure 9. Comparison of performance parameter

The efficacy of the proposed DL+ ensemble technique was evaluated against the top 5 designs (5-Clf) and all models (8-Clf) to ascertain whether lower-performing models adversely impact the effectiveness of stronger models in ensemble algorithms. As shown in Table 3, the DL+ ensemble approach demonstrated the highest precision across all parameters, with accuracy at 99.50%, precision at 99.50%, recall at 99.50%, and F-score at 99.50%. Furthermore, Figure 10 showcases the performance of various ensemble designs juxtaposed with the proposed technique, providing insights into the relative effectiveness of each approach.

Table 4 presents a comparative analysis of several ensemble designs, highlighting the effectiveness of the proposed DL+ ensemble technique. In this approach, predictions from three distinct deep learning models were integrated using a weighted average, where the weight assigned to each model correlated with its accuracy on the validation set.

The superior performance of the DL+ ensemble method compared to other models is attributed to its integration of varied deep learning architectures. Each model within the ensemble possesses unique strengths and compensates for the limitations of others, contributing to a reduction in the overall error rate. The application of weighted averaging in aggregating predictions further enhances the model's accuracy, allocating greater significance to predictions from better-performing models. Ensemble methods, by amalgamating the forecasts of multiple models, effectively decrease the incidence of misclassifications. This reduction is achieved as different models in the ensemble may commit disparate errors, but their collective predictions lead to a diminished error rate. Moreover, ensemble methods exhibit enhanced generalizability compared to individual models, thereby reducing the propensity to overfit to training data.

The DL+ ensemble model outperformed the other ensemble methods due to its integration of diverse deep learning models. Each model contributes unique strengths and mitigates individual weaknesses, collectively reducing the overall error rate. The application of weighted averaging in combining predictions further accentuates the influence of better-performing models. Ensemble methods, by amalgamating multiple model predictions, reduce misclassifications more effectively than individual models. These methods offer enhanced generalizability, thus reducing the likelihood of overfitting to training data.

The implications of these findings extend to both future research and practical applications. The results demonstrate that ensemble methods are a promising avenue for COVID-19 detection from chest CT images. The diversity

of models within the ensemble underscores the importance of varied approaches. Additionally, the study offers insights into optimizing ensemble models for improved diagnostic accuracy.

Table 3. Comparative analysis of performance metrics

Parameters	Proposed 5-Clf	Proposed 8-Clf	DL+ Ensemble
Accuracy	98.99	98.99	99.500000
Precision	99.02	98.98	99.500000
Recall	98.97	99.00	99.504950
F-score	98.99	98.99	99.497487

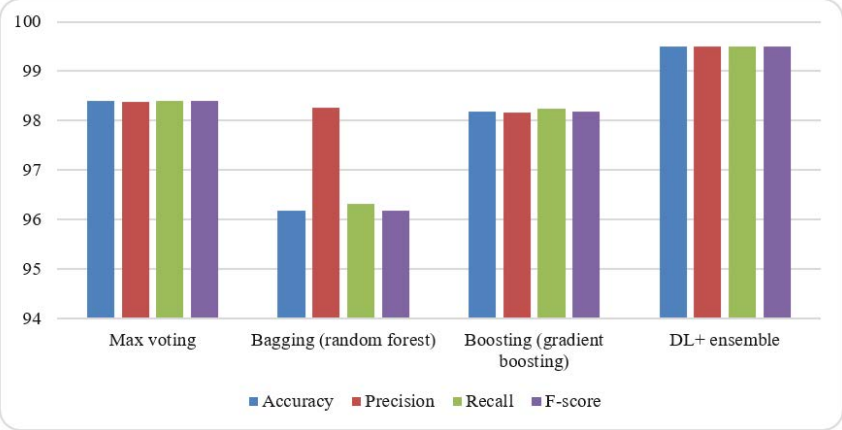

Figure 10. Comparison of performance parameter

Table 4. Comparative analysis of ensemble techniques

Parameters	Max Voting	Bagging (Random Forest)	Boosting (Gradient Boosting)	DL+ Ensemble
Accuracy	98.39	96.18	98.19	99.500000
Precision	98.37	98.26	98.17	99.500000
Recall	98.40	96.32	98.25	99.504950
F-score	98.39	96.18	98.19	99.497487

In practical terms, the DL+ ensemble approach holds potential for the development of a CAD system for COVID-19 detection. Such a system could significantly aid radiologists in rendering more accurate and expedient diagnoses.

4 Conclusions

The ensemble approach proposed in this study for the detection of COVID-19 from chest CT scan images has been demonstrated to surpass other models in terms of accuracy, precision, recall, and F-score. This approach integrates the deep pre-trained convolutional bases from VGGNet, ResNet, MobileNet, InceptionV3, Xception, and IRV2, enabling the extraction of discriminative features from CT scan images. Several advantages have been identified in the proposed ensemble method compared to traditional diagnostic approaches. Firstly, the integration of predictions from multiple models contributes to a reduction in the overall error rate. Secondly, the utilization of deep pre-trained convolutional bases facilitates the extraction of key features from CT scan images, obviating the need for extensive model training. Thirdly, the implementation of a regularized categorization head aids in preventing overfitting to the training data.

The enhanced diagnostic accuracy afforded by this ensemble approach holds significant potential for impacting the diagnosis and treatment of COVID-19 patients, particularly in real-world scenarios. Moreover, the potential of this approach in developing a CAD system for COVID-19 detection is considerable. Such a system could assist radiologists, especially in settings where resources are limited. Future research directions include evaluating the performance of the proposed ensemble approach with a more diverse dataset. The intention is to incorporate CT scan images from patients with various types of pneumonia and other medical conditions that may present similar characteristics to COVID-19 on CT scans. Additionally, the incorporation of other medical data types, such as clinical and laboratory data, will be explored to further enhance the ensemble approach's performance.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, and D. Shen, "Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 4–15, 2021. https://doi.org/10.1109/RBME.2020.2987975

[2] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, and W. Ji, "Sensitivity of chest CT for COVID-19: Comparison to RT-PCR," *Radiology*, vol. 296, no. 2, pp. E115–E117, 2020. https://doi.org/10.1148/radiol.2020200432

[3] S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, A. Rovira, A. Oliver, and X. Lladó, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3D CNN approach," *Neuroimage*, vol. 155, pp. 159–168, 2017. https://doi.org/10.1016/j.neuroimage.2017.04.034

[4] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. San Tan, "A deep CNN model to classify heartbeats," *Comput. Biol. Med.*, vol. 89, pp. 389–396, 2017. https://doi.org/10.1016/j.compbiomed.2017.08.022

[5] Z. Wang, J. Li, and M. Enoh, "Removing ring artifacts in CBCT images via generative adversarial networks with unidirectional relative total variation loss," *Neural Comput. Appl.*, vol. 31, no. 9, pp. 5147–5158, 2019. https://doi.org/10.1007/s00521-018-04007-6

[6] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang, "Low-Dose CT with a residual encoder-decoder CNN," *IEEE Trans. Med. Imaging*, vol. 36, no. 12, pp. 2524–2535, 2017. https://doi.org/10.1109/TMI.2017.2715284

[7] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017. https://doi.org/10.1016/j.media.2017.07.005

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[9] M. T. Letsatsi, A. Agarwal, and O. M. Seretse, "The battle towards skill-based competency integration to knowledge-based competency in the sustainable development of growing economy," *Int. J. Recent Technol. Eng.*, vol. 8, no. 1, 2019.

[10] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, USA, 2009, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[11] S. Pal, "Transfer learning and fine tuning for cross domain image classification with keras," 2016. https://www.slideshare.net/sujitpal/transfer-learning-and-fine-tuning-for-cross-domain-image-classification-with-keras.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA, 2016, pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[13] F. Chollet, "keras," 2015. https://github.com/fchollet/keras.

[14] W. Alawad, B. Alburaidi, A. Alzahrani, and F. Alflaj, "A comparative study of stand-alone and hybrid CNN models for COVID-19 detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, 2021. https://doi.org/10.14569/IJACSA.2021.01206102

[15] R. Ali, R. C. Hardie, and H. K. Ragb, "Ensemble lung segmentation system using deep neural networks," in *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. Chennai, India, 2020, pp. 1–5. https://doi.org/10.1109/AIPR50011.2020.9425311

[16] O. Saha, J. Tasnim, M. T. Raihan, T. Mahmud, I. Ahmmed, and S. A. Fattah, "A multi-model based ensembling approach to detect COVID-19 from chest X-ray images," in *2020 IEEE Region 10 Conference (TENCON)*. Washington DC, USA, 2020, pp. 591–595. https://doi.org/10.1109/TENCON50793.2020.9293802

[17] S. Tewari, U. Agrawal, S. Verma, S. Kumar, and S. Jeevaraj, "Ensemble model for COVID-19 detection from chest X-ray scans using image segmentation, fuzzy color and stacking approaches," in *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*. Chennai, India, 2020, pp. 1–6. https://doi.org/10.1109/CICT51604.2020.9312076

[18] S. D. Deb, R. K. Jha, R. Kumar, P. S. Tripathi, Y. Talera, and M. Kumar, "CoVSeverity-Net: An efficient deep learning model for COVID-19 severity estimation from Chest X-ray images," *Res. Biomed. Eng.*, vol. 39, pp. 85–98, 2023. https://doi.org/10.1007/s42600-022-00254-8

[19] L. J. Muhammad, M. M. Islam, S. S. Usman, and S. I. Ayon, "Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery," *SN Comput. Sci.*, vol. 1, no. 4, p. 206, 2020. https://doi.org/10.1007/s42979-020-00216-w