Classification of Pulmonary Diseases Using a Deep Learning Stacking Ensemble Model

Ruaa N. Sadoon*, Adala M. Chaid
College of Computer Science & Information Technology, University of Basrah, Basrah, Iraq
E-mail: pgs.ruaa.nabeel@uobasrah.edu.iq, adala.gyad@uobasrah.edu.iq
*Corresponding author

Keywords: medical imaging, deep learning, CNN, COVID-19, pulmonary pathologies, image classification

Received: May 2, 2024

This paper presents our research in the area of medical imaging diagnostics, focusing specifically on countering the devastating impact of the COVID-19 pandemic and numerous pulmonary pathologies. Using new deep-learning approaches and techniques, we aim to create an advanced classification tool that will be able to capture complex patterns and features in chest image data. This paper introduces the use of state-of-the-art strategies, such as stacked ensemble models, transfer learning, and artificial neural networks, to build a model with unprecedented precision, recall, F1-score, and accuracy. The core idea of our research is to combine different convolutional neural network architectures to bring together their best extraction and classification qualities. The combination of DenseNet, Xception and Inception achieves the best performance and provides the most reliable classification tool. We also use transfer learning to quickly train our model and optimize generalization, making it suitable for the detection of multiple pulmonary pathologies, including COVID-19. Our model also includes an artificial neural network, trained as a meta-learner, which processes the outputs of the CNNs to make classification decisions. We have thoroughly validated and optimized the meng-learner to improve the model's accuracy on diagnostic images. The provided paper proposed the successful merge of cutting-edge deep-learning methodologies and image-processing algorithms with the medical imaging industry's specifics. We aim to disrupt the pulmonary disease diagnosis field with our model, offering medical institutions a reliable tool to fight the current and future threats and challenges posed by COVID-19.

Povzetek: Razvit je napredni model za klasifikacijo pljučnih bolezni z uporabo zloženega modela globokega učenja, ki združuje CNN arhitekture, kot so DenseNet, Xception in Inception. Model dosega visoko natančnost pri odkrivanju bolezni, vključno s COVID-19.

1 Introduction

The emergence of the coronavirus disease 2019 in late 2019 sparked a global health crisis unlike any other [1]. Advances in diagnostic methods are vitally crucial since this highly contagious disease [2] has overtaken healthcare systems globally and severely disrupted life [3]. Even yet, prior to COVID-19, a number of pulmonary diseases, including pneumonia, lung opacities, and effusions, posed diagnostic difficulties that required precise diagnosis [4]. Notably, significant diagnostic challenges are intrinsic to common pulmonary illnesses such as pneumonia, pleural effusion, pulmonary nodules, and pneumothorax [5]. Differential diagnosis is crucial since pneumonia can present with symptoms resembling COVID-19 [6]. Pleural effusion frequently obscures or mimics other lung diseases, making it more difficult to interpret images from various angles [7]. In order to rule out benign or malignant disorders, pulmonary nodules typically require complicated, high-resolution imaging [8]. Although pneumothorax directly endangers patients, it can be challenging to differentiate its symptoms from those of other acute chest disorders [9]. Advanced identification approaches utilizing deep learning models are necessary for nuanced detection due to the symptom overlap with COVID-19 [10], requiring precise methodologies in the context of pulmonary diseases.

Many facets of contemporary life have been reshaped by deep learning, a subclass of machine learning that is typified by automatic feature extraction and picture categorization [11]. Deep learning has significantly changed how the healthcare industry is thought about [12]. By examining medical photos, models that can accurately predict or categorize particular diseases may now be created [13]. Promising results have been obtained from deep learning techniques for the diagnosis of a number of illnesses, including as brain tumors, liver disease, colon cancer, breast cancer, lung cancer, pneumonia, and most recently, COVID-19 [14].

Deep learning automatically changes features using non-linear functions, resulting in high accuracy with less human interaction than classical machine learning, which

needs manual feature engineering [15]. The extraction of valuable characteristics is improved as the network becomes deeper because more abstract data representations are learned [16].

Recent scientific achievements demonstrate how deep learning is widely used to identify and treat COVID-19 [17]. For COVID-19 analysis, chest X-rays and CT scans are frequently used imaging modalities [18,19]. Although COVID-19 positive cases have been categorized using X-rays, interest in CT scan-based diagnosis is developing [19]. Convolutional neural networks have been used to examine lung dataset instances to classify COVID-19 cases [20]. Evidence suggests that chest X-rays are more valuable for diagnosis than for differential diagnosis of other serious conditions including pneumonia and lung cancer, even though they are less helpful in the early stages of COVID-19 before symptoms appear [19]. This means that in difficult circumstances, radiologists need help from automated diagnostic tools. Consequently, we used deep learning to address problems related to pulmonary diseases and the COVID-19 pandemic [10,21]. To capitalize on the capabilities of different deep learning techniques, such as DenseNet, Xception, and Inception, we have developed a unique strategy that comprises a stacked ensemble model [21]. A thorough analysis of the use of transfer learning and Convolutional Neural Networks (CNNs) for medical imaging applications is presented in [22]. It highlights the significant improvements CNNs have demonstrated in image analysis and classification applications and how transfer learning—reusing previously trained CNN models—can alleviate issues arising from small datasets and computing limitations.

We have innovated by using an artificial neural network (ANN) as the meta-learner to supervise the ensemble model, which goes beyond the ensemble of deep learning architectures [23]. In particular, our artificial neural network (ANN) serves as the last arbiter by combining the predictions made by each of our component convolutional neural networks (CNNs) and categorizing the provided medical images. First off, our model is incredibly flexible and capable of making decisions because of the ANN's capacity to identify complex patterns and relationships in the ensemble's predictions. Second, we have optimized the ANN's parameters to maximize its performance through intensive training and validation, guaranteeing that our multi-tumor classification system reaches the highest levels of diagnostic accuracy and dependability [10].

Our research essentially constitutes a singular amalgamation of cutting-edge deep learning, image processing, and medical imaging domain knowledge. We can radically alter the patient-centered clinical workflow for the diagnosis and treatment of pulmonary diseases, including but not limited to COVID-19 and pneumonia, by combining the advantages of ensemble models, transfer learning, and ANNs.

2 Related works

Medical imaging featuring deep learning represents one of the most promising advancements in the field, particularly concerning COVID-19 detection. Different papers have taken numerous angles and used separate datasets; however, the accuracy has invariably been compelling.

The authors in [24] presented an innovative lung opacity detection and classification approach that is significantly essential to physicians due to its non-reversible consequence when inaccurately diagnosed or misjudged with other diseases. To this end, the authors present a three-channel fusion CNN model, where the authors use the MobileNetV2, InceptionV3 and VGG19 networks for each channel. ResNet is used for transferring features. The classification has shown a promising accuracy in lung opacity classification for different datasets. For the new dataset, the model reports an adequate performance of accuracy of 92.52%, 92.44%, 87.12% and 91.71% for two, three, four and five classes, respectively. A comparison with the previous research indicates the potential of the model. It can significantly reduce the burden and costs of physicians who use image datasets for lung opacity classification.

The COVID-19 epidemic has severely damaged economies and healthcare systems throughout the world, underscoring the urgent need for accurate and quick diagnosis techniques in the fight against the illness [25]. The current methods of clinical diagnosis have significant shortcomings because they are very subjective and subject to variation amongst patients. This article suggests a novel multi-classification method based on a machine learning framework to get beyond these restrictions. In particular, it presents BDCNet, a novel method for classifying COVID-19, pneumonia, and lung cancer from chest radiographs by using Vgg-19 and convolutional neural networks (CNNs). The goal of the suggested approach is to offer a consistent and objective diagnostic tool for identifying between various lung diseases.

Notably, this is the first study to diagnose these three chest diseases using a single deep learning model. Results indicate that BDCNet outperforms four well-known pre-trained models, achieving an accuracy of 99.10%, recall of 98.31%, precision of 99.9%, and f1-score of 99.09%. These findings highlight the potential of BDCNet to significantly aid diagnostic radiographers and healthcare experts in accurately identifying and managing chest diseases, thus contributing to improved patient outcomes and healthcare efficiency.

The authors in [26] addressed the critical challenge of accurately diagnosing COVID-19 and other chest disorders amidst their overlapping symptoms, which can potentially mislead clinical professionals. To tackle this, the researchers develop and evaluate a multi-classification deep learning model called CDC Net, leveraging convolutional neural network (CNN) techniques with residual network concepts and dilated convolution. By employing publicly available benchmark data, they pioneer the use of a single deep learning model to diagnose five distinct chest ailments, including

COVID-19, lung cancer, pneumothorax, tuberculosis, and pneumonia, from chest x-ray images. Remarkably, the CDC Net achieves an exceptional AUC of 0.9953, demonstrating an accuracy of 99.39%, a recall of 98.13%, and a precision of 99.42% in identifying various chest diseases. Comparative analysis with three CNN-based pre-trained models further underscores the superior performance of the proposed model, highlighting its potential as a highly accurate diagnostic tool for chest diseases. Moreover, statistical analyses confirm the robustness of the proposed model, affirming its reliability and effectiveness in clinical settings.

The authors in [27] proposed a new method of chest x-ray classification to diagnose COVID-19 with pneumonia caused by usual virus and to overcome the problem in which patients with COVID-19 cannot be differentiated with other chest disorders. This model is based on CNN model that applies a pre-trained EfficientNetB0 model and a dense layer. The model achieved high accuracy of over 95% out of two classes and 93% out of three classes, which outperforms the existing model and present some benefits, with less parameters and robust dataset split.

Through meticulous methodological design, including data augmentation and fine-tuning, the study demonstrates the potential of CNN-based models in enhancing the accuracy of COVID-19 diagnosis from chest x-ray images, thereby supporting clinicians in making more informed diagnostic decisions.

The authors in [28] addressed the urgent need for accurate diagnosis of COVID-19 amidst the global pandemic, proposing a deep learning-based approach to differentiate COVID-19 patients from those with viral pneumonia, bacterial pneumonia, and healthy cases. Utilizing deep transfer learning, the study experimented with binary and multi-class datasets across four categories, comprising a total of 6,674 X-ray images. Nine convolutional neural network architectures were employed, including Se-ResNeXt-50, which achieved the highest classification accuracy of 99.32% for binary classification and 97.55% for multi-class classification among all pre-trained models. By leveraging automated methods and sophisticated CNN architectures, the proposed system demonstrates promising performance in accurately diagnosing COVID-19, contributing to the ongoing efforts to combat the spread of the disease.

The authors in [29] presented a novel multi-level diagnostic framework aimed at accurately detecting COVID-19 using X-ray scans, offering a promising alternative to the conventional RT-PCR method. The framework proposed in the current study consists of three phases, which are pre-processing to clean noise and resize the images, feature extraction using a deep learning architecture with an Xception pre-trained model. The framework incorporates global average pooling to overcome overfitting, an activation layer help to reduce loss and softmax for the final classification. This proposed model has been tested using a benchmark dataset from Kaggle containing 7395 images from three classes, COVID-19, normal, and pneumonia, which has shown an exceptional outcome. Testing has been conducted with an accuracy of 99.3% and a negligible loss of 0.02 by using leakyReLU activation and RMSprop optimizer. Therefore, utilizing just 10 epochs and a learning rate of 10−4 to achieve 99% sensitivity and specificity with F1-Score of 99.3% indicates the efficiency and performance of the proposed framework in identifying COVID-19 accurately from X-ray images. Hence, it is more efficient than existing studies and traditional pre-trained deep learning models.

In [30], five pre-trained AI models were applied to improve brain tumor classification, attaining 95-97% accuracy on unseen images across three datasets. Data augmentation improved model performance, perhaps boosting early tumor identification and lowering impairments. Machine learning and deep learning algorithms were used to identify chest CT scans as COVID-19 positive or negative [31]. The study found that ResNet50V2 transfer learning approach performed best on the bigger dataset, with 97.52% accuracy, showing its potential for quick COVID-19 diagnosis in real life.

The authors in [15] presented the development of a Multi-task Multi-slice Deep Learning System tailored for the screening of multi-class lung pneumonia from CT imaging. To solve the problem of limited training cases and resources, the M3 Lung-Sys consists of two 2D CNN networks, dedicated to slice- and patient-level classification. By leveraging CT slices for feature extraction and refining temporal information across slices, the system effectively distinguishes COVID-19 from Healthy, H1N1, and CAP cases while also locating relevant lesion areas without pixel-level annotation. Extensive experiments conducted on a chest CT dataset demonstrate the superior performance of M3 Lung-Sys, achieving an accuracy of 95.21% with minimal false positive and false negative errors. Notably, the system exhibits high sensitivity and specificity for COVID-19 and H1N1 detection, outperforming existing models. Although oversensitivity to noise is observed in Healthy cases, the interpretability and value of the system to clinicians are underscored by its robust performance and lesion location mapping capabilities. Overall, M3 Lung-Sys offers a promising solution for accurate and interpretable multi-class lung pneumonia screening from CT imaging, particularly in the context of the COVID-19 pandemic.

The authors in [16] used computed tomography (CT) and chest X-ray imaging modalities to meet the critical requirement for quick and reliable disease identification in the midst of the global COVID-19 epidemic. Understanding the shortcomings of RT-PCR testing and the potential of imaging methods—particularly in areas where epidemics are prevalent—the study investigates the use of machine learning (ML) to support disease diagnosis. The research developed a deep neural network model, specifically a 24-layer CNN network, capable of binary (COVID vs. NON-COVID) and multi-class (COVID vs. NON-COVID vs. Pneumonia) classification from X-ray and CT images. Through extensive experimentation, the proposed method achieves

remarkable accuracy rates of 99.68% and 71.81% on X-ray and CT images, respectively, demonstrating its efficacy in aiding rapid and effective COVID-19 detection. Utilizing the Sgdm optimizer with a learning rate of 0.001 contributes to the robust performance of the model across both datasets, showcasing its potential as a valuable tool in combating the pandemic.

The authors in [17] suggested an automated COVID-19 detection method utilizing artificial intelligence (AI) technology in response to the worldwide COVID-19 pandemic and the load on healthcare infrastructure. The goal of the study is to accurately identify COVID-19 from normal chest X-ray pictures. It also aims to distinguish COVID-19 from viral pneumonia that is not COVID-19 and lung opacity. Three pre-trained models, namely, Xception, VGG19, and ResNet50, are used and assessed on a benchmark dataset with 21,165 X-ray images. Initially, a binary classification model for COVID-19 detection is implemented, and the models are able to achieve high accuracy levels: 97.5%, 97.5%, and 93.3% for Xception, VGG19, and ResNet50, respectively. Then the multi-class classification model is created, and the accuracy levels are obtained as follows: 93%, 92%, and 75%, for Xception, VGG19, and ResNet50, respectively. Particularly, Xception model demonstrates higher precision, recall and f-1 scores, which shows its successful implementation in such tasks. Explainable AI is added to increase the level of interpretability; it enables a visual representation of the model's predictions and reasoning behind them. This is done to restore the confidence of medical units in AI and support the application of AI in clinical decision-making. Overall, the study encompasses a significant development in the domain of automated COVID-19 detection and introduces a helpful, accurate, and interpretable solution for application throughout the world. The comparative results related to this study are presented in table 1.

Table 1: Comparative table of related works

References	Method	Image Type	Accuracy
[24]	MobileNetV2, InceptionV3, VGG19, ResNet	Lung opacity	-Two classes: 92.52% -Three classes: 92.44% -Four classes: 87.12% -Five classes: 91.71%
[25]	BDCNet (combining Vgg-19 and convolutional neural networks)	Chest radiographs	99.10%
[26]	CDC Net (Multi-classification deep learning model)	Chest X-ray images	99.39%
[27]	Convolutional Neural Network (CNN) specifically combining a pre-trained EfficientNetB0 network with a dense layer	Chest X-ray images	-Two-class classification (COVID-19 vs. other viral pneumonias): 95% -Three-class classification (COVID-19 vs. other viral pneumonias vs. other chest disorders): 93%
[28]	convolutional neural network (CNN) architectures, including Se-ResNeXt-50.	X-ray images.	-For binary classification accuracy of 99.32%. -For multi-class classification accuracy of 97.55%.

When comparing deep learning models for medical picture classification, there is a significant difference in accuracy depending on the model architecture and classification difficulty. The authors in [24] used a mixture of MobileNetV2, InceptionV3, VGG19, and ResNet to detect lung opacity in X-ray images, with accuracies ranging from 87.12% to 92.52% across two to five classes. Other research [25, 26, 28] shows better performance with other convolutional neural network designs. For example, the CDC Net [26] and BDCNet [25], which concentrate on chest X-ray pictures and chest radiographs, respectively, yield accuracies higher than 99%, suggesting a more successful method for binary categorization. In contrast to the research [25, 26, 28], another study [27] uses a hybrid model that combines a dense layer with a pre-trained EfficientNetB0. The results reveal somewhat lower accuracies in two and three-class classifications (95% and 93%, respectively). This implies that although the combined procedures in [25, 26, 28] are quite successful for binary and multi-class classifications, the complex method in [27] provides slightly lower accuracy. These variations highlight how model architecture and training methods affect medical image analysis classification accuracy.

The application of deep learning in medical imaging, especially for COVID-19 detection, marks significant progress in diagnostic methodologies. While various models demonstrate high accuracy, existing approaches still present limitations that our research aims to address.

Three-Channel fusion CNN models [24]: While this model achieves impressive accuracy levels up to 92.52% for lung opacity classification, its performance variably decreases to 87.12% as the number of classes increases, indicating a potential drop in effectiveness

with complex classifications. Our work extends these efforts by employing a stacking ensemble model that not only maintains high accuracy across an increased number of classes but also integrates more diverse CNN architectures to stabilize performance across varied diagnostic scenarios.

BDCNet [25]: This model excellently classifies COVID-19, pneumonia, and lung cancer with an accuracy of 99.10%. Although it demonstrates robustness, it focuses on a limited array of diseases. Our approach includes a broader spectrum of pulmonary pathologies, enhancing the utility of deep learning models in more diverse clinical settings.

CDC Net [26]: Achieving an AUC of 0.9953, this model is highly accurate. However, it primarily employs traditional CNN architectures with residual networks. Our model introduces an artificial neural network as a meta-learner to further refine the diagnostic process, aiming for nuanced understanding and integration of features extracted by base learners.

EfficientNetB0 hybrid models [27]: With accuracies of 95% and 93% for binary and three-class classifications respectively, this model shows a reduction in performance with more complex class scenarios. Our methodology leverages a meta-learning approach that consistently manages high accuracy even as classification complexity increases.

Ensemble and meta-learning approaches: Most existing studies utilize single-model systems that may not capture all nuances in complex image data. Our research introduces an ensemble of multiple advanced models (DenseNet, Xception, Inception) with a meta-learner that synergistically improves prediction accuracy and robustness, addressing the gap in existing single-model systems.

By incorporating these advancements, our work significantly contributes to the field by providing a comprehensive and adaptable solution that enhances the detection and classification of a wide range of pulmonary diseases, not just limited to COVID-19 but extending to other less commonly addressed conditions such as pneumothorax and pulmonary fibrosis. This holistic approach is crucial for deploying deep learning effectively in real-world clinical settings, where diversity in pathology presentation demands robust and flexible diagnostic systems.

3 Methodology

In our methodology, we begin with a comprehensive dataset acquisition process that includes a variety of medical imaging data pertinent to pulmonary conditions such as COVID-19, pneumonia, lung opacities, and effusions. The first step in our methodology consists in extensive preprocessing and exploratory data analysis (EDA) to guarantee the data's quality and appropriateness for machine learning exploitation. Especially, this involves the resizing of images, normalization of pixel values, and handling of any missing or null data. Thereafter, we train several deep learning models, such as DenseNet and two variants of Inception. These models were chosen based on their successfulness in feature detection in complex image data, as well as setting the network parameters appropriately to extract from differences in specific characteristics of the various primary pulmonary conditions.

Once the training phase has been conducted, we assess each model's performance by employing significant metrics, including accuracy, sensitivity, specificity, and F1-score and without which to assess the ability of a model to classify the medical images accurately on its own.

Based on this assessment, we then conduct feature extraction such that we extract the noteworthy features in the outputs of the base learners, i.e., the features that contain the salient characteristics necessary to classify accurately. Afterwards, we input these features into an ensemble model.

We propose a meta-model, an artificial neural network which acts as the decision layer, that harmonizes the insights from the DenseNet and the two Inception models via a stacking algorithm. This model combines the strengths of each base learner to enhance classification accuracy and robustness.

Finally, we evaluate the ensemble meta-model using the four metrics of accuracy, sensitivity, specificity, and F1-score. This step measures the improvement of the performance and the reliability of the new multi-tumor classification system. The resulting system is fit for clinicians and will thus help manage and treat pulmonary ailments better. Our methodology is explained in figure 1.

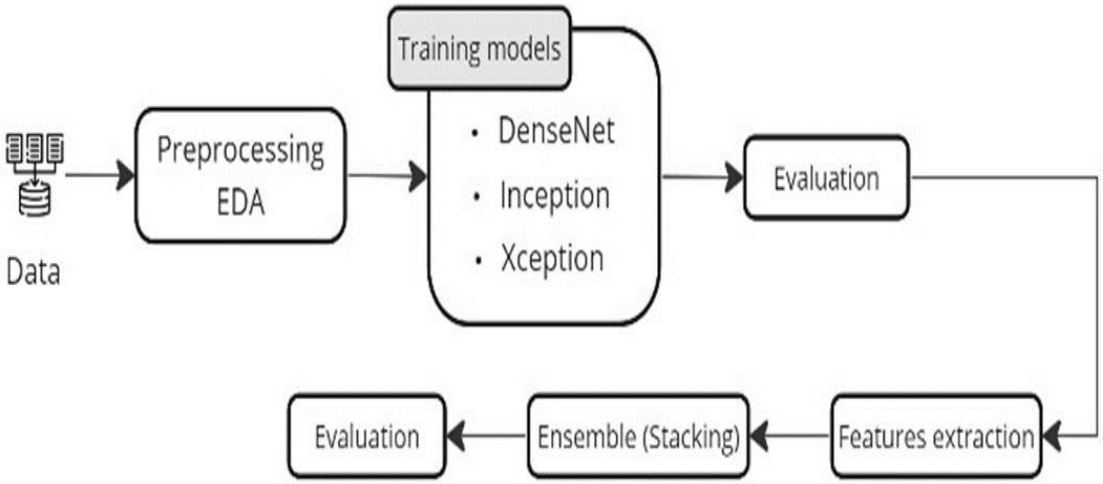

Figure 1: Methodology

3.1. Datset overview

In this work we provide diverse set of well-curated publicly available medical imaging datasets, which are crucial for training state-of-the-art machine learning models to detect chest related diseases including COVID-19, tuberculosis and pneumothorax. We have selected these datasets for their comprehensive representation of various clinical scenarios and radiographic findings to make our study more exhaustive and clinically relevant. The COVIDx CXR-4 Dataset to be released, part of the larger COVID-Net initiative, is a composite collection of over 30,000 chest radiographs with coding collected from several healthcare institutions around the world, to pretrain deep learning models, specifically for discovering COVID-19 and pneumonia. The data set is also systematically split into validation and test sets in order to allow for legitimate cross evaluation and meaningful testing.

Consisting of thousands of chest X-rays from normal findings to TB-positive cases, this database represents a gold standard for TB diagnostics developed jointly by an international consortium. At the same time, the ChestX-Ray14 dataset has over 100,000 radiographs with annotations associated to text-mined radiological reports, enabling weakly-supervised learning techniques (one of the largest datasets for radiology and NIH is already planning an expansion to 200k records). It contains thousands of images for multi-label classification where an image may contain [0, 1 or more] of the 14 different kinds of pathologies that the model is to diagnose at the same time. In addition, the COVID-19 Radiography Database is constantly updated with images of different phases of COVID-19 infection and maintains its relevance as the disease progresses. Furthermore, to mitigate potential biases introduced by this broad set of sources, we conduct diversified-sourcing, balanced-sampling, and stratified cross-validation to help ensure our models are generalized and robust across various demographic variation and clinical contexts. We safeguard the privacy of patient data by always de-identifying all datasets and then vetting the public availability/scanning source/patient-consent of the data to adhere to ethics protocols with patients in mind and comply with the HIPAA and GDPR requirments. This ethical rigor is a testament to our dedication to keeping the patient information private and the correctness of data.

The appropriate use of these datasets strategically builds a concrete stepping- stone to reliable and accurate diagnostic aids that are capable of addressing the complexities involved in different pulmonary pathologies Our systematic effort in dataset diversity, potential biases, and ethical standards vindicates the scientific soundness and ethical quality of the developed methodology, thereby laying the foundation for an important benchmark amongst diagnostic AI research agendas.

3.2. Data preprocessing methodology

Step-by-Step data preprocessing: The first step in our preprocessing is a detailed exploratory analysis using a visualization grid to display 20 random images along with their labels. This step ensures not only the accuracy of labels but also highlights the types and difficulty level of data such as image clarity, orientation, or anomaly visibility. Normalization involves reducing pixel values within each image to a range of 0 to 1 through standardization. This alleviates model training dynamics problems, including the speed of convergence and sensitivity to the scale of input data.

Advanced augmentation methods: An important component of training powerful deep neural networks, especially in medical imaging, is data augmentation. This is crucial as the range of variability of data is often quite

wide and may be relatively scarce. Our augmentation strategy includes:

-Geometric transformations: Rotations (up to 20 degrees) and translations (shifts of up to 10% in both x and y axes) represent different patient positions and imaging angles.

-Zoom and shear perturbations: Zoom perturbations (up to 20% increase/decrease) and shear transformations (up to 10%) simulate differences in patient size relative to the imaging machine and subtle movements.

-Color space augmentations: Brightness and contrast transforms help the model learn feature mapping under different imaging conditions and equipment settings.

-Elastic deformations: Stretching or squeezing images in a non-linear manner to account for realistic variations between physiological examples and changes in imaging perspectives.

Each augmentation technique is chosen deliberately to reflect realistic variations and challenges faced by our diagnostic model in a clinical context, enhancing its ability to generalize from training data to real clinical applications.

Training parameters and hyperparameter tuning: The initial learning rate for our model was set to 0.001, adaptively adjusted during training using the Adam optimizer. We choose the Adam optimizer for its suitability to sparse gradients and adaptability to different scenarios, essential for medical imaging tasks**.**

-Training (A) Choosing the batch size: A batch size of 32 balances efficient learning dynamics and computational resources. This size ensures good diversity within gradient estimates while avoiding memory exhaustion.

-Training epochs and early stopping: We allow training to run for up to 100 epochs, with early stopping based on validation loss to prevent overfitting once model performance stops improving.

Model tuning: We use a grid search with cross-validation to explore different combinations of learning rates, batch sizes, dropout rates, and augmentation parameters. This systematic process ensures thorough testing of each combination to identify the optimal set of parameters that yield the highest performance on the validation set. Grid search with cross-validation assesses the model's generalizability across data splits, ensuring robustness in various clinical scenarios.

Validation and testing: Validation during training uses a separate subset (20% of the training data) to unbiasedly tune and evaluate the model's generalization capabilities. The final model's performance is evaluated on a separate testing set, mimicking real-world application scenarios to ensure robustness and generalization for clinical deployment.

3.3. Convolutional neural network architecture

Through the full-depth exploration from our study, we explored four CNNs including VGG16, DenseNet201, InceptionV3, and Xception. The structural design of these models is capable of capturing and analyzing complex image data efficiently, The Inference models VGG16, Densenet201, Inceptionv3, Xception CNNs either employ various layers of convolution filter or pooling operation, or use successive operations to extract the higher-level image feature compounds gradually. For the task of fine-grained recognition where detailed image classification is involved, this process is essential.

Leveraging the power of transfer learning, we introduced pre-trained weights from the extensive ImageNet database into our models. This approach harnesses the diverse and rich feature sets learned by these networks on a broad array of image types, thus furnishing our models with a robust foundation of visual knowledge. By employing these pre-trained networks, we effectively accelerate the training phase and enhance the model's ability to generalize better when exposed to new, unseen datasets.

In adapting these pre-trained models to our specific task, we customized their architectures by removing the original top layers, which are typically fully connected layers designed for specific classification tasks on ImageNet. Instead, we focused on maintaining the convolutional base for its potent feature extraction capabilities. This adjustment ensures that the models remain versatile and more focused on extracting universally applicable features from the images.

To ensure uniformity and compatibility across all models, all of the input images were preprocessed into the same shape of 224×224 pixels in RGB format, which is the input requirement for all of the aforementioned pre-trained network models. Furthermore, each of the pre-trained models has its own specific preprocessing subroutine such as normalization and pixel value scaling designed to prepare the images before processing them through the neural network. The rationale behind this step is that the pixel values required to follow the distribution mean and standard deviation of the images in ImageNet.

Our methodology extended the pre-trained models' convolutional base with new layers designed to aid learning for our specific classification tasks. This included global average pooling for spatial dimension reduction, batch normalization to normalize the input layers and stabilize learning, dropouts to prevent overfitting, and dense layers for the final classification. Each of these layers was essential in developing the model by increasing its sensitivity to meaningful features and simultaneously minimizing the potential of memorizing irrelevant data patterns.

Finally, we assessed the performance of each of the four models using a comprehensive set of performance metrics comprising accuracy, precision, recall, and the F1-score computed on them. These metrics offered an all-round perspective of the performance of each of the four models while identifying the strengths of each of the four models in the accurate clustering of images into existing classes. This set of evaluations, therefore, had a dual goal of determining which model was most suited for the accuracy of our categorization tasks considering our specific data, as well as contributing to the entire body of knowledge on the subject with the addition of

empirical evidence and better testing techniques. Thus, our work was a combination of a modeling and evaluation effort that pursued further the practical and theoretical aspects of the application of CNN in image clustering.

3.4. MobileNet

MobileNet is a lightweight convolutional neural network architecture designed with mobile and embedded devices in mind. The model factorizes the traditional large-scale CNNs into lighter models, thus enabling us to deploy them in situations that are computational expensive or model size limited.

MobileNet is developed to find the right solution to the computational and performance efficiency curve, making it ideal for applications that require lightweight models, but still maintain a certain level of accuracy.

Introducing MobileNet to the research provides additional modeling benefits for scenarios that computation resources and the model file is a limitation. The integration of MobileNet in the proposed classification architecture not only increased the inference deployment initiatives and to various applications like mobile app and edge devices but also to resource-constraint systems.

Considering the use of MobileNet in our classification pipeline, it was necessary to carefully prepare the data for the subsequent training and testing processes. First, we divided our dataset into two mutually exclusive subsamples – the training subsample and the testing subsample using the train_test_split function. As a result of the stratified subsample division, we obtained a sufficiently diverse sample to train the model and an independent empirical set to evaluate the ability of the model to classify samples it had never seen. We used this experimental methodology in order to obtain reliable estimates of generalization capabilities and several classification performances measures.

Next, we created data generators for our training and testing subsamples using TensorFlow's Image Data Generator. By resorting to data generators, loading and preprocessing these images became a more straightforward and computationally efficient process. For MobileNet, we used the preprocessing function provided by tf. keras applications mobilenet_v3 preprocess input. It was crucial to pre-process our input images this way to ensure the correct normalization and consistency with the preprocessing requirements of the MobileNet architecture to maximize its classification efficiency.

The incorporation of MobileNet within our classification framework has been a tactical venture designed to promote our methodology's scalability, efficacy, and flexibility. Using the more lightweight structure and architecture of MobileNet and the approach to inference, we initially intended to expand our classification model's relevance to different deployment settings, such as mobile apps, edge devices, and Internet of Things platforms. By conducting extensive data preprocessing procedures and integrating the model accordingly, the goal was to "unlock" MobileNet's full potential in resolving actual image classification issues within multiple domains and applications.

4 Evaluation metrices

To provide a complete evaluation of the model's performance, we employ several evaluation metrics, which provide a full understanding of the model's performance in image classification. These metrics are essential in determining the model's performance, including applications in medical diagnostic fields. The following are brief abstractions of the use of each.

4.1. Accuracy

Accuracy is the most essential evaluation metric that measures the quality of the model's predictions in general. It is calculated as the number of correctly classified samples, which include true positives and true negatives, divided by the total number of samples.

$$ACC = \frac{TN + TP}{TP + TN + FP + FN}.$$

4.2. Precision

This measurement metric is fundamental for measuring the correctness in the positive predictions made by the model. It is computed as the quotient of the sum of true positive predictions and false ones and the true positives only.

$$PRE = \frac{TP}{FP + TP}.$$

4.3. Recall

one of the primary metrics evaluated, is the ability of the model to identify all instances of a class that truly belongs to it. This is computed by True Positives divided by True Positives plus False negatives.

$$REC = \frac{TP}{TP + FN}.$$

4.4. F1-Score

The F1-score is a balance of precision and recall. It is computed as Harmonic mean of Precision and Recall, which is a single measure taking both metrics into consideration.

$$F_1 - S = 2 \times \frac{PRE \times REC}{PRE + REC}.$$

4.5 Roc curve

A ROC (Receiver Operating Characteristic) curve is a graphical tool that allows performance evaluation of a multiclass classification model to show itself at many threshold levels. ROC curve), which is formed by plotting True Positive Rate (TPR a.k.a. Sensitivity/Recall) on y-axis and False Positive Rate (FPR) on x-axis. TPR, The proportion of actual positives

which are correctly identified by the model FPR, The proportion of negatives which are incorrectly labeled as positives.

The TPR-FPR tradeoff can then be changed by varying the model classification threshold. For a good model, ROC curve will be tending to the upper left corner of the plot, higher the True Positive Rate and lower the False Positive Rate across various thresholds.

Because ROC Chart evaluates a model which is independent of classification threshold, it is ideal in cases when one balance between TPR and FPR is more critical than the other. In the case of fraud detection, it may be worth to load even a huge false positives list for the sake of making the phishing True positive rate as high as possible. On the other hand, when it comes to medical diagnostics the reduction of FPR might be more important, even if it comes at the cost of a slightly lower TPR.

In summary, ROC curves give a graphical representation which helps us to decide which model should be chosen for a binary classification problem based on the threshold of sensitivity and specificity needed.

4.6 Confusion matrix

A confusion matrix is a statistical device used in machine learning to determine how well classification models are in capable of predicting and classifying data in a dataset for which real values are known. This matrix represents the counts for true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) True Positives: Observations correctly predicted as positive True Negatives: Observations correctly predicted as negative on the other hand, false positives (Type I error) and false negatives (Type II error) indicate when the model labeled negatives as true positives and vice-versa. A confusion matrix is critical to quantify performance metrics avenue like accuracy, precision, recall, F1 score showing how well the model is able to distinguish between classes accurately.

4.7 AUC

The Area Under the Curve (AUC) is a performance measurement for classification problem at various threshold settings. Which is essentially (AUC - ROC) Curve where True Positive Rate [Sensitivity] is plotted over False Positive Rate [1- specificity] considering different points (Thresholds) It can be a value between 0 and 1 and the closer to 1, the better the model.

The AUC represents a summary of the model performance across all possible classification thresholds. An AUC value of 0.5 suggests that the model has no ability to discriminate (i.e., a model with no discriminative power being equivalent to random guessing) and of 1 indicate perfect discrimination between positive class and negative class.

In a more general derivation, the AUC is calculated by drawing a B-Spline curve through the points on the ROC curve and using a closed form equation related to the trapezoidal rule which approximates the area under the curve. The formula is:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(t)\ d\mathrm{FPR}(t).$$

Where:

- $(\mathrm{TPR}(t))$ is the true positive rate at threshold (t)
- $(\mathrm{FPR}(t))$ is the false positive rate at threshold (t)

In practice, the AUC is often computed using numerical integration techniques on the points that make up the ROC curve.

5 Experimental results

5.1 DenseNet

The DenseNet201 architecture from TensorFlow's Keras applications was used in our study. However, it was editedto suit a specific classification task better. The model was initialized with pre-trained ImageNet weights without the top layer included for customization for custom output layers. To ensure that the pre-trained features are not tampered with, the base model's layers were set to be non-trainable. Subsequently, the network was expanded with specific layers developed to fine-tune and optimize previously existing features to accommodate our classification needs. These include Global Average Pooling, Batch Normalization, several Dense layers employing ReLU activations, and Dropout to avoid over fitting. Furthermore, the model was composed with an SGD optimizer. Throughout the training process, early stopping was used to monitor and stop struggling when the validation score failed to improve any further ensuring that the model was maximally generalized. Following training, the model achieved high classification accuracy as well as other performance metrics on the multi-class image dataset as depicted in Table 2, enabling the distinction of several types of medical images, including tuberculosis, pneumonia, and pulmonary fibrosis, among others. The results were further validated through the classification report documenting the precision, recall, as well as the F1-scores among the various classes. This process affirms the robustness and accuracy of our method in medical image analysis.

Table 2: Classification report of DenseNET

Classes	Precision	Recall	F1-score
Control 10	0.96	1.00	0.98
Covid 09	0.99	0.99	0.99
Effusion 08	0.96	0.96	0.98
Lung Opacity 07	0.98	0.96	0.97
Mass 06	0.95	0.96	0.96
Nodule 05	0.94	0.87	0.90
Pneumonia 04	0.92	0.97	0.94
Pneumothorax 03	0.91	0.95	0.93
Pulmonary fibrosis 02	0.93	0.98	0.96
Tuberculosis 01	1.00	0.94	0.97
Accuracy			0.95
Macro avg.	0.95	0.96	0.96

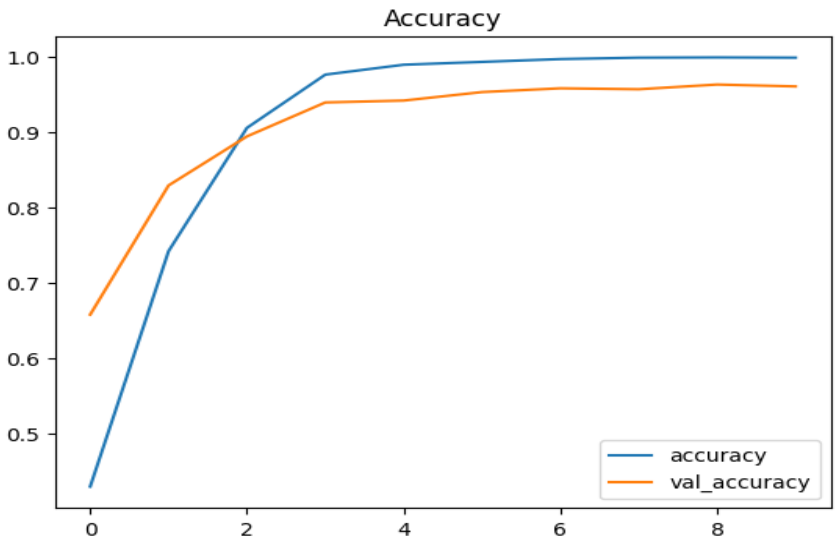

Figure 2: Accuracy of DenseNet model

Figure 2 illustrates the accuracy results of a DenseNet model over several training epochs. The x-axis enumerates the epochs, and the y-axis represents the accuracy metric, scaled between 0 and 1. The blue line traces the training accuracy across epochs, initiating at about 50% and steeply ascending to near-perfection, suggesting rapid learning in the initial phases. It plateaus close to 100%, which implies a strong fit to the training data. Conversely, the orange line, denoting validation accuracy, begins slightly lower than the training accuracy, suggesting that the model doesn't generalize quite as well initially. It increases at a steady rate, albeit with a less steep slope compared to the training accuracy, before plateauing at a value slightly under 100%. This indicates a good but not perfect generalization to new data.

DenseNet in the training loss and validation loss is decreasing per training epoch, showing the improvement of the model in terms of prediction of the target classes. On the first training batch, both losses are high, which tells us the model have little to no understanding of the data yet. Later into training, the training loss drops faster than the validation loss, which levels off, this is the point where both losses have converged. This plateau informs to stop training further because the model will not benefit significantly more from it emphasizing on why early stopping is in play to avoid overfitting and make sure the model generalizes even on new data. The overall decreasing loss plot confirms a successful learning phase, a necessity for the high accuracy.

The Receiver Operating Characteristic (ROC) curve, provided here shows that the DenseNet model performs well on the class 9 which can be evident from the high region under the curve under the curve which is area under the curve, if you like to know more about ROC and AUC watch this (ROC-AUC) video. An Area Under the Curve (AUC) of 0.960 is considered to be in the range with an extraordinary ability to discriminate, where the model was capable of accurately identifying 96% of the times the class of the 9 (positive instances) cases and just 3 % of the times mistaking it to be negative. This is evident by the ROC curve which shows that the performance of the model remains high with nearly 1.0 true positive rates even at relatively low false positive rates and holds good as they increase. This high AUC number suggests that DenseNet results in the best performance among the class 9, making it the best

contestant of a medical image classification with the lowest misclassifications.

The confusion matrix illustrates how the model does in several classes for the DenseNet model. From the matrix we see that there are many correct classifications likely with 94 for Class 0 and 121 for Class 6, but there are perhaps notable misclassifications as well. Class 4, for example, has instances wrongfully classified to Classes 0 and 7, and Class 7 has a lot of instances misclassified to Class 8, showing a difficulty in separating these classes specifically. Also, there is a significant number of misclassifications for Class 8 and Class 1 (which are misclassified as Class 3 and Class 7) and to a lesser extent, some instances for Class 9 (total no 96) which are wrongly classified as Class 8. The training labels of ”7” were misclassified into ”9”, which shows the predictive abilities of the DenseNet model, but since the samples were still misclassified, it suggests that there is still remaining room for the model to improve, and a way to better reduce errors and improve the overall classification performance.

5.2 Inceptionv3

We addressed a difficult image classification problem using the InceptionV3 architecture, known for its complexityand depth. This model utilized the ImageNet weights and had its highest layer excluded; hence, the network did not make specific classifications but rather extracted features. As opposed to the other layers, the InceptionV3 layer’s trainable model parameters were frozen. This was needed to preserve features obtained from the first training instances and prevent instability during the initial learning process. Our custom model architecture was based on this robust foundation and included additional layers for maximizing classification accuracy. Global Average Pooling was employed to consolidate all feature maps into a single vector per feature map, followed by Batch Normalization for faster convergence and several dense layers to increase the learning potential of the model. A significant dropout rate of 0.5 was included to prevent overfitting.The model was compiled using the Adam optimizer, balancing the benefits of both RMSprop and SGD, and aimed to optimizefor precision, recall, and overall accuracy.

The classification report for the InceptionV3 model presents a comprehensive overview of its performance across various classes. As depicted in Table 3, the ’Control 10’ class achieved the highest F1-score at 0.95, indicating exceptional precision and recall. In contrast, the ’Nodule 05’ class had the lowest F1-score of 0.77, which suggests room for improvement in either precision or recall or both for this category. The model showed strong precision in the ’Covid 09’ and ’Tuberculosis 01’ categories, scoring 0.93, but the recall was notably lower in ’Tuberculosis 01’, reflecting that some cases may have been missed. ’Mass 06’ exhibited high recall at 0.94, implying that the model is reliably identifying most of the positive cases for that condition, although precision is slightly lower at 0.81. Across all classes, the model achieved an accuracy of 0.85 and both macro average precision and recall are balanced at 0.86, indicating consistent performance across different conditions. The F1 scores, which balance precision and recall, are relatively high for most conditions, demonstrating the effectiveness of the InceptionV3 model in varying scenarios. However, there are differences among the conditions that could be addressed to improve the model’s diagnostic capabilities further.

Table 3. Classification report of inceptionv3

	Precision	**Recall**	**F1-score**
Control 10	0.91	1.00	0.95
Covid 09	0.93	0.87	0.90
Effusion 08	0.82	0.78	0.80
Lung Opacity 07	0.92	0.81	0.86
Mass 06	0.81	0.94	0.87
Nodule 05	0.84	0.71	0.77
Pneumonia 04	0.85	0.96	0.90
Pneumothorax 03	0.83	0.88	0.85
Pulmonary fibrosis 02	0.77	0.93	0.84
Tuberculosis 01	0.93	0.72	0.81
accuracy			0.85
Macro avg.	0.86	0.86	0.86

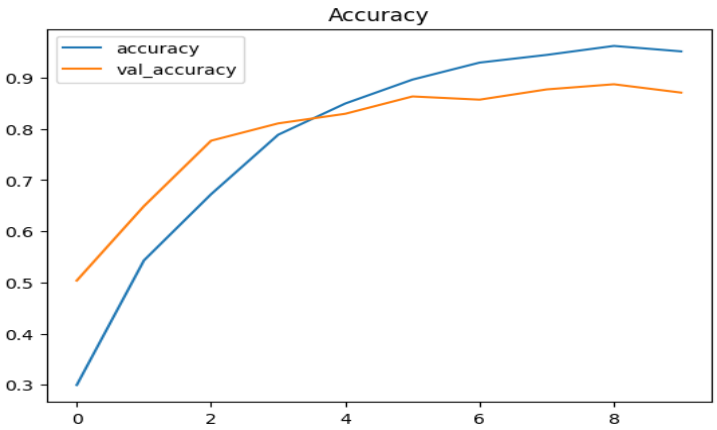

Figure 3: Accuracy of inceptionv3

Figure 3 depicts the accuracy trends for an InceptionV3 model during its training and validation phases. On the x-axis, we observe the number of epochs, and on the y-axis, the accuracy metric is presented, ranging from around 0.35 to just above 0.9. The training accuracy, marked in blue, starts just above 0.4 and shows a steady increase as training progresses, suggesting that the model is learning from the training data. It continues to improve, albeit with a gradual slope, before plateauing near 0.9, which indicates a high level of accuracy on the training dataset.

The validation accuracy, coloured in orange, also starts at a similar point but increases at a quicker rate initially. It reaches its peak at around the third epoch, which is over 0.8, indicating that the model was quite effective on the validation data at this point. However, post this peak, it begins to decrease and then levels off, ending with a slight downward trend. This could suggest that the model began to overfit to the training data after the third epoch, as it performed better on training data than on unseen validation data. It is also noteworthy that the validation accuracy ends up lower than the training accuracy, which further supports the possibility of overfitting.

In the first confusion matrix, for the InceptionV3 model, we see that it has considerable difficulty in predicting many of the classes. While some predictions are correct, such as predicting 103 out of 111 Class 0 (Control) and 107 out of 144 Class 8 (Pneumothorax), the model is making many misclassifications. Fig 1 — Class 5 (Mass) is often mistaken as Class 4 (Effusion) and Class 6 (Nodule) for example. Also, Class 9 (Tuberculosis) which confounds with many classes. The above misclassifications suggest the confusion of the model in distinguishing close classes that need to use further refinement and improvements for overall performance.

For the Inception model, the ROC curves of class 9 show the ROC curve performance higher than other classes, AUC was equal to 0.906. This means that the model is 90.6% certain that it will give a higher ranking to a random positive instance (i.e., class 9) than to a random negative instance that could belong to any class. The curve indicates that the model retains an almost 0.9 true positive rate, even where the false positive rate is at its lowest. The widening of this range is good for the model, indicating it will continue to perform well over a larger range of false positive rates. Second, a specificity of 0.82 indicates that the Inception model detects class 9 very well without false positives, which is particularly desired in medical image classification.

5.3 Xception

The Xception model employed in our study leverages the Xception architecture pre-trained on the ImageNet dataset to extract features from medical images. We initialized the base Xception model with frozen layers to preserve the learned features during training. In the subsequent step, we developed a sequential model, with batch normalization, global average pooling, and dense layers for feature extraction and aggregation, data normalization, and classification, respectively. Softmax activation function with 10 neurons was responsible for the probability distribution across the classes of the output layer. The model was trained using the Adam optimizer to compute the gradient of the categorical cross-entropy loss. Early stopping was also applied to prevent the model from overfitting during training and ensure optimal convergence. The model was then tested on the test dataset as demonstrated in Table 4, achieving 81% overall accuracy. The model exhibited decent performance in terms of precision, recall, and F1-score, indicating its classification power for different classes. Hence,the Xception model demonstrated its strength in classifying medical images accurately, especially for diagnosing pulmonary pathologies.

Table 4: Classification report of Xception

	Precision	**Recall**	**F1-score**
Control 10	0.99	0.96	0.98
Covid 09	0.87	0.93	0.90
Effusion 08	0.84	0.75	0.79
Lung Opacity 07	0.85	0.54	0.66
Mass 06	0.78	0.91	0.84
Nodule 05	0.77	0.69	0.73
Pneumonia 04	0.84	0.84	0.84
Pneumothorax 03	0.75	0.85	0.80
Pulmonary fibrosis 02	0.73	0.89	0.80
Tuberculosis 01	0.68	0.72	0.70
accuracy			0.81
Macro avg.	0.81	0.81	0.80

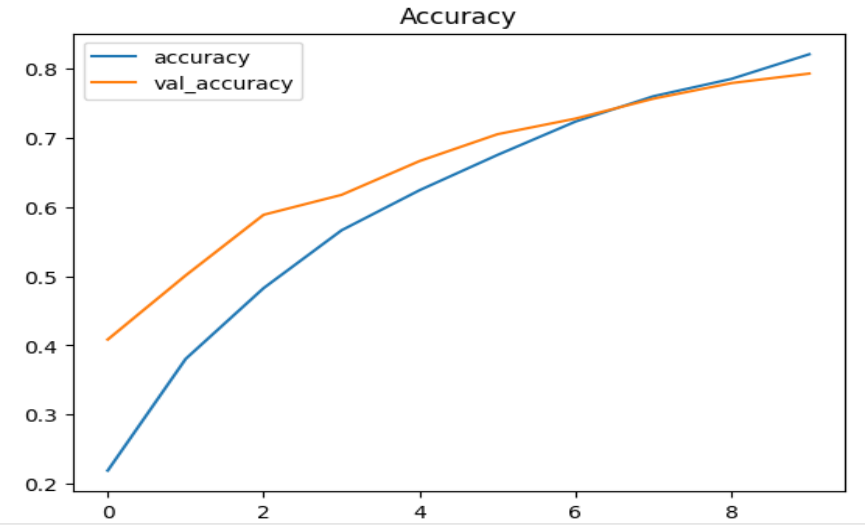

Figure 4: Accuracy of Xception

Figure 4 showcases the accuracy of an Xception model during its training and validation processes over a series of epochs, with the x-axis tracking the epoch count and the y-axis representing the accuracy metric between approximately 0.2 and 0.8. The blue line represents the model's training accuracy begins just above 0.2 and rises steadily, reflecting consistent learning as the epochs progress. This steady ascent suggests that the model is effectively learning patterns from the training dataset. Around the seventh epoch, it slightly plateaus, indicating that the model may be approaching its learning capacity based on the current data and configuration.

In contrast, the validation accuracy, indicated by the orange line, follows a similar upward trajectory, yet surpasses the training accuracy after the initial epochs. This is an unusual pattern as typically, models tend to perform better on the training data due to familiarity. The validation accuracy's higher values could suggest that the validation set might not be as challenging, or it might indicate good generalization depending on the diversity and representativeness of the validation set compared to real-world data. There's also a possibility of data leakage or an issue with the training/validation split that is causing the validation accuracy to be inflated.

ROC curve, for class 9 of the Xception model. This is the ability of the model to distinguish this class from others. The curve plot the true positive rate (TPR) vs. false positive rate (FPR) at 30 different thresholds. The discriminative power of the model is high, with an AUC of 0.859, which indicates that an instance of class 9 will be correctly classified as such with probability of 85.9%, and a negative instance (any other class) will also be classified as negative with the same probability, on average. The curve indicates that at low false positive rates, the true positive rate approaches 0.80 and advances as the false positive rate becomes larger. The model remains sensitive to detecting true positives but more false positives start to occur AUC of this value indicates the Xception model has a high accuracy of classifying class 9, and it is suitable for medical image classification. This is a small value while others has the large value but compared with larger AUC values,this is low and can be improved in another model.

Confusion Matrix for Xception model with multi-class Classification Task The diagonal values in the matrix show how many instances of each class were correctly predicted (high accuracy in classes, such as 0, 1 and 5 where 89, 96 and 85 instances of each class are successfully predicted). Off-diagonal values: These provide an idea of where the model fails (along the vertical rings and horizontal predictors) The most visible is — the class 2 instances are always mistaken for most of the other classes (with a majority of the confusion between class 2, 3, 4, 5, 7, and 8). For instance, class 8 suffers from a misclassification distribution across a variety of classes indicating potential areas the model can be enhanced in clearly defining these classes. Detailed visualization results in a holistic performance evaluation of the model, highlighting high, low classification accuracy of certain classes along with specifics of misclassification regions, leading to better targeting and determining of where optimizations and improvements in training and evaluation are required.

5.4 MobileNet

This description defines the procedure of deploying a TensorFlow MobileNet V3 Image classification Model. We divided the dataset into 80% for training and 20% for testing. The training data is preprocessed and augmented, using the ImageDataGenerator, during preparation while the testing data only gets preprocessed. Then we split the training set not only into train and valid sets. The model architecture consists of a pretrained MobileNet V3 model, global average pooling, batch normalization, flatten, dense layers, dropout layer for regularisation and softmax activation function on the output layer to categorize images in the 10 classes.

Here, during training, the accuracy and loss metrics are plotted for both the training and validation data. When evaluated, the model yields an approximate accuracy of 94.89 with a test loss of 0.25495. This is one of many such reports and others, played strong precision, recall, and F1 in each category with an overall accuracy of 95%. The performance of the model for the detection of different health conditions, including pneumonia, pulmonary fibrosis, and COVID-19, as characterized by classification metrics and confusion matrix processed. These results show the remarkable capabilities of the MobilNet V3 in image classification by achieving a trade-off between depth and complexity of the network and efficiency in feature extraction.

Table 5: Classification report of MobileNet

Classes	Precision	Recall	F1-score
Control 10	1.00	1.00	1.00
Covid 09	0.99	0.97	0.98
Effusion 08	0.90	0.94	0.92
Lung Opacity 07	1.00	0.99	1.00
Mass 06	0.92	0.92	0.92
Nodule 05	0.96	0.89	0.92
Pneumonia 04	1.00	0.97	0.98
Pneumothorax 03	0.90	0.95	0.92
Pulmonary fibrosis 02	0.94	0.95	0.95
Tuberculosis 01	0.97	0.98	0.97
accuracy			0.96
Macro avg	0.96	0.96	0.96

The Training performance of MobileNet can shed light on both where it excels and where it falls short. The model trains fast and was already at 90% within a few epochs and flatting out at 94%, but if you look closely at the validation metrics the your forehead will fill up with sweat. The Validation accuracy follows a similar trend, but has a bit of a deviation from the training accuracy, especially during the beginning. Indicating that the initial training data is overfitting the exact model. Observations are also corroborated by the loss curves. Again the validation accuracy increases, even the training loss suddenly drops as it happened in the previous example. Nevertheless, validation loss is going deep down in the beginning and after sometimes it fluctuates. The fluctuation may suggest overfitting, where the model is effectively memorizing noise in the training data that does not generalize. From the first paper, this is in line with previous findings on initial learning abilities and generalization of MobileNet. Nevertheless this small gap may lead to over fitting and care should be taken during training to ensure the model works best on data that has been not been seen. One can use early stopping techniques to stop training that can help achieve this.

The classification of class 9 with ROC curve in MobileNet is better. This curve represents how well the model is able to distinguish class 9 (true positive rate) from other classes (false positive rate). This is further emphasized by a high Area Under the Curve (AUC) of 0.964. In layman's terms, the model has 96.4 percent chance of classifying the above picture as class 9 The ROC curve additionally validates this by displaying that if other classes are only slightly confused (low False Positive Rates) then the background class is still almost always correct (True Positive Rate is near 1.0). It represents the strongness and reliability of the model in detecting class 9. In summary, the high AUC value indicates that MobileNet has the extraordinary ability to have few errors in the classification of class 9 (i. e. a small MSE) and is a very useful tool for the general task of medical image classification.

The confusion matrix of MobileNet model for multi class classifiction task. The accuracy of the model is good in classes such as 2, 5, 7, and 8, and this is shown in the diagonal values, where 101, 110, 109, and 109 are correctly predicted, respectively. The off-diagonal values are the instances that were misclassified in the model and help us understand the model's confusion points. For example, class 6 has some confusion with classes 3 and 9 and class 3 has some confusion with classes 2, 4, and 5. This Fine-Grained overview makes It possible to obtain an overarching View of the Model's Strengths in Predicting Almost all Classes Accurately and to Pinpointing Troublesome Confusion Areas.

5.5 Ensemble (stacking)

In this study, we utilized an ensemble method in making the use of stacking in assembling different pre-trained models to improve the prediction quality of a challenging image classification task. We incorporated three different CNN models including DenseNet201, InceptionV3, and Xception. These models had already been pre-trained and saved with trace and had strong predictive qualities. Therefore, we used them as base learners which first predicted and then their predictions were used as features for the meta-model. Our meta-model used the dense architecture with 25 neurons ReLU for feature integration, followed by a softmax layer with 10 outputs representing our class categories.

This model was trained on a dataset split into training and validation subsets, ensuring robustness and the ability to generalize from the ensemble predictions. Optimized with the Adam optimizer and compiled with a categorical cross-entropy loss function, the meta-model focused on refining the decision boundaries formed by the base models. After training, the ensemble's effectiveness was evaluated using precision, recall, and F1-score metrics, revealing outstanding classification performance across various categories, with nearly all classes achieving near-perfect scores. This result underscores the power of combining multiple advanced neural network architectures to achieve superior accuracy and reliability in medical image classification tasks. In Table 6 the classification report for the ensemble model, which uses a meta–Artificial Neural Network (ANN) approach, shows outstanding performance across all classes. The model achieves near-perfect precision and recall in most categories, as reflected by the F1 scores. Remarkably, ' the Covid 09' and 'Lung Opacity 07' classes both scored a perfect 1.00 across precision, recall, and F1-score,

indicating that the model has exceptional accuracy in identifying these conditions. Similarly, 'Control 10' also demonstrates almost flawless performance with an F1-score of 0.99. The other classes, such as 'Effusion 08', 'Mass 06', 'Nodule 05', and 'Pneumonia 04', maintain high metrics, with F1-scores ranging from 0.97 to 0.99, suggesting the model is highly effective in distinguishing these conditions with minimal false positives or negatives. The overall accuracy of the ensemble model is extremely high at 0.98, and the macro averages for precision, recall, and the F1-score mirror this value, highlighting consistent and reliable performance across the board. This indicates that the stacking approach of the ensemble model, which likely integrates multiple learning algorithms, results in superior predictive capability. The balanced precision and recall suggest that the model is not only capturing the majority of positive cases but is also correctly identifying negatives, which is crucial for medical diagnostics. These results imply a robust model with excellent generalization properties for the considered conditions.

Table 6: Classification report of meta model (ANN)

Classes	Precision	Recall	F1-score
Control 10	0.99	1.00	0.99
Covid 09	1.00	1.00	1.00
Effusion 08	0.96	0.97	0.97
Lung Opacity 07	1.00	1.00	1.00
Mass 06	0.96	0.97	0.97
Nodule 05	0.99	0.96	0.97
Pneumonia 04	0.99	0.99	0.99
Pneumothorax 03	0.97	0.98	0.97
Pulmonary fibrosis 02	0.97	0.97	0.97
Tuberculosis 01	0.99	0.97	0.98
accuracy			0.98
Macro avg.	0.98	0.98	0.98

Figure 5 illustrates the accuracy of an ensemble (stacking) meta-model during its training phase, over several epochs, as shown on the x-axis. The y-axis quantifies the accuracy, which is scaled between 0.5 and 1.0, suggesting that accuracy is represented as a proportion. The blue line charts the model's training accuracy, which begins at around 0.5 and sharply climbs to just above 0.9 within the first two epochs. This rapid ascent indicates that the ensemble model quickly learns from the training data. Subsequently, the training accuracy exhibits a more gradual increase and seems to level off near the 1.0 mark, suggesting that the model fits the training data well.

Conversely, the orange line indicates the validation accuracy. It starts at a similar level to the training accuracy but doesn't ascend as steeply. After catching up to the training accuracy at around the second epoch, it diverges and starts to lag slightly, finishing just below the training accuracy curve. This divergence may be indicative of a small degree of overfitting to the training data, but the proximity of the two lines at the end of the training indicates that the ensemble model generalizes well to unseen data. The plateau nearing the end of the epochs suggests that the model is stabilizing and that additional training might not result in significant accuracy improvements. The high validation accuracy maintained throughout the training process reflects the efficacy of the ensemble approach, which often leads to robust generalization by leveraging the strengths of multiple individual models.

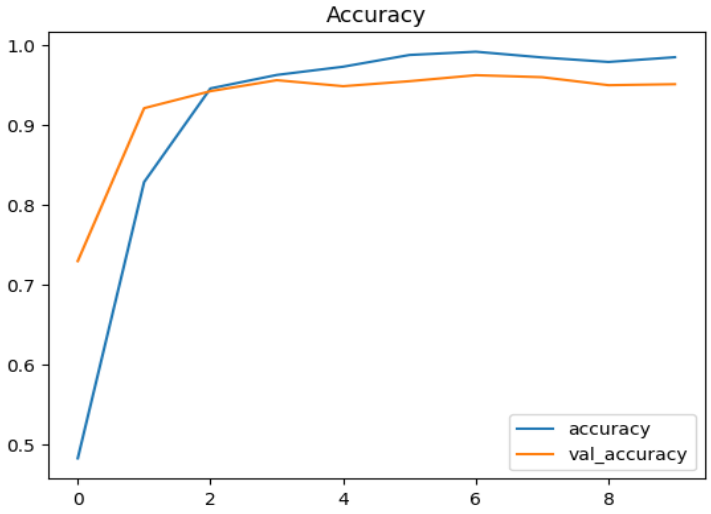

Figure 5: Accuracy of meta model

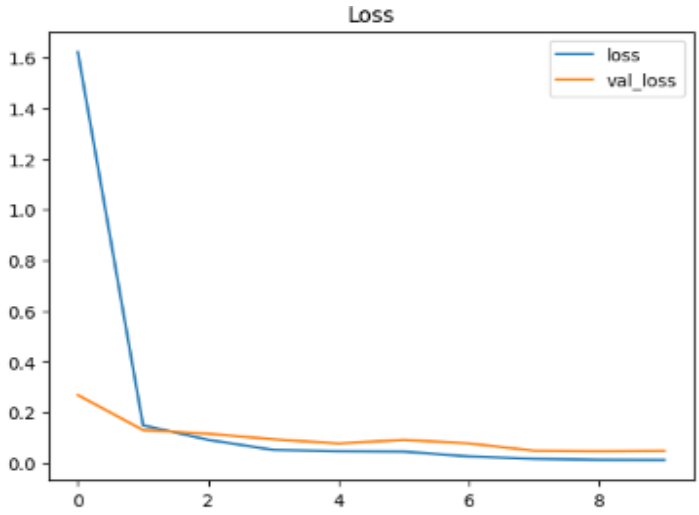

Figure 6: Loss of meta model

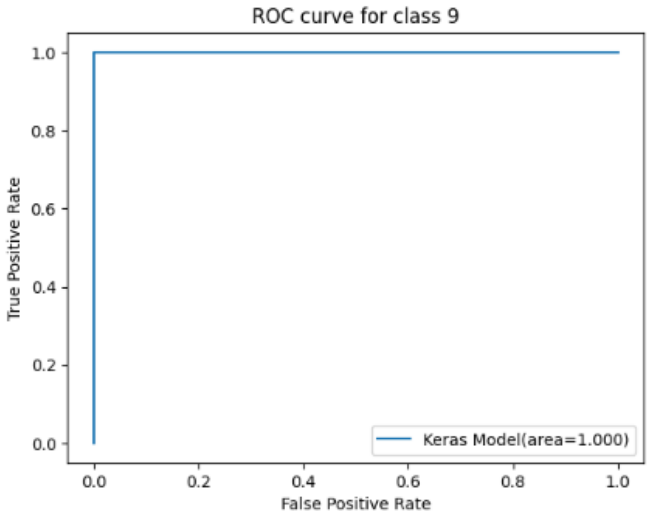

Figure 7: ROC curve for class 9 of meta model

Figure 5 illustrates the accuracy of an ensemble (stacking) meta-model during its training phase, over several epochs, as shown on the x-axis. The y-axis quantifies the accuracy, which is scaled between 0.5 and 1.0, suggesting that accuracy is represented as a proportion. The blue line charts the model's training accuracy, which begins at around 0.5 and sharply climbs to just above 0.9 within the first two epochs. This rapid ascent indicates that the ensemble model quickly learns from the training data. Subsequently, the training accuracy exhibits a more gradual increase and seems to level off near the 1.0 mark, suggesting that the model fits the training data well. While Figure 6 depicts the loss in the accuracy of the meta-model as it undergoes training over multiple epochs, as indicated on the x-axis. The y-axis measures the accuracy, which is scaled from 0.5 to 1.0, indicating that accuracy is expressed as a proportion. The blue line represents the training accuracy of the model, starting at approximately 0.5 and rapidly increasing to slightly above 0.9 within the initial two epochs. The significant increase in performance suggests that the ensemble model rapidly acquires knowledge from the training data. Following that, the training accuracy demonstrates a slower and steadier rise and appears to stabilise around the 1.0 threshold, indicating that the model effectively matches the training data.

Conversely, the orange line indicates the validation accuracy. It starts at a similar level to the training accuracy but doesn't ascend as steeply. After catching up to the training accuracy at around the second epoch, it diverges and starts to lag slightly, finishing just below the training accuracy curve. This divergence may be indicative of a small degree of overfitting to the training data, but the proximity of the two lines at the end of the training indicates that the ensemble model generalizes well to unseen data. The plateau nearing the end of the epochs suggests that the model is stabilizing and that additional training might not result in significant accuracy improvements. The high validation accuracy maintained throughout the training process reflects the efficacy of the ensemble approach, which often leads to robust generalization by leveraging the strengths of multiple individual models.

The ROC curve illustrates how the stacking model performs for class 9 visually. It illustrates in Figure 7, the compromise between the percentage of class 9 that the model can identify (True Positive Rate) and the percentage of other classes that the model predicts as class 9 (False Positive Rate). A ROC curve itself is evaluated on (TPF, FPF) where a perfect curve would be a straight line rising from the bottom left corner (0,0) to the top left corner (0,1) and then across to the top right corner (1,1). This means that the model can completely separate class 9 from all other classes.

A perfect AUC of 1.0 in the ROC curve of class 9 implies the highly capable classification of class 9 by the model It implies that the model has perfect precision for class 9 only class 9 examples are getting predicted as class 9 and we do not see any other class getting predicted as class 9.

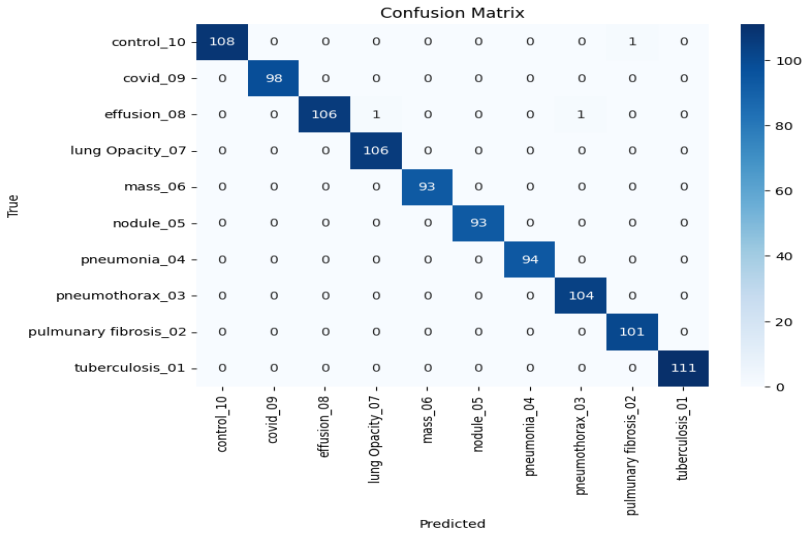

Figure 8: Meta model Confusion Matrix

Its multi class classification task can be seen in the Confusion Matrix from the Ensemble Learning Model in Figure 8. The diagonal values correspond to the number of rightly classified instances of every class (4 in total) with high accuracy for the classes — control_10, effusion_08, pneumothorax_03, and tuberculosis_01 with 108, 106, 104, 111 appropriate predicted instances, respectively. Meanwhile, off-diagonal values give the number of points that got classified wrongly and indicate where the model is confusing into. For instance, a class lung Opacity_07 has a number of misclassifications into classes which effusion_08 and mass_06, and a class pulmonary fibrosis_02 demonstrates some misclassifications into the class mass_06.

6 Comparative analysis

When comparing the performance of individual CNN models—DenseNet201, InceptionV3, and Xception—with that of the ensemble stacking model, we observe significant distinctions and improvements in the ensemble approach. DenseNet201 exhibited robust performance, achieving a high precision and recall with a weighted average precision of 0.96 and accuracy of 0.95. InceptionV3, while still performing commendably, displayed slightly lower efficacy, especially in precision and recall for certain classes such as effusion and lung opacity, leading to a total accuracy of 0.85. Xception's results were less uniform; all but four classes had average precision and recall below 0.8, resulting in 0.81 diagnoses in total.

Meanwhile, the ensemble stacking, which based its diagnoses on the three base models' predictions, showed the best results on all metrics. With almost perfect precision and recall in every category, totaling 0.98 accuracy indicates that the approach has yielded significant levels of successful predictions. It is especially noteworthy that the ensemble outperformed non-ensemble models in categories the latter found difficult, underlining the efficiency of combining different strengths to address weaknesses.

The above comparative analysis demonstrates the potential of the ensemble stacking method to improve the predictive efficiency and precision of the more complicated classification tasks by combining the unique strengths of several models to achieve the ultimate performance superiority.

The following table summarizes the performance comparison between the individual models and the ensemble stacking model according to the precision, recall, f1-score, as well as the overall accuracy:

Table 7: Comparison of model's performance

Model	Precision	Recall	F1-score	Accuracy
DenseNet201	0.95	0.96	0.96	0.95
InceptionV3	0.86	0.86	0.86	0.85
Xception	0.81	0.81	0.80	0.81
Stacking	0.98	0.98	0.98	0.98

Based on Table 7, the ensemble stacking model and its ensembled learning approach to classification through stacking achieved optimal results compared to single-

learner models. The superior performance of the stacking model continues to demonstrate the gain that one can make from having multiple model predictions for a final outcome – the way the model increases accuracy and dependability of classification in complex scenarios.

7 Discussion

As shown in Table 7, our ensemble stacking model performs the best in terms of precision, recall, F1-score, and accuracy, consistently achieving at least a 98% range across these metrics. Compared to the outcomes listed in the "Related Works" section, this indicates the strengths and weaknesses of our approach for these problem types.

State-of-the-Art (SOTA) comparison: Multi-Channel Fusion CNN (Paper [24]): Achieved an accuracy of 92.52%. Our model outperforms this by maintaining high accuracy consistently across more complex multi-class settings. This improvement arises from our model's ability to jointly optimize features from different CNN architectures through a meta-learning process, which is not used in Paper [24].

BDCNet (Paper [25]): The BDCNet shows excellent performance in precision and accuracy, focusing on COVID-19, pneumonia, and lung cancer. While BDCNet excels in these diseases, our model generalizes this high performance across a broader range of pulmonary conditions, including pulmonary fibrosis and pneumothorax, highlighting its clinical relevance across multiple diseases.

CDC Net (Paper [26]): This model has an extremely high AUC of 0.9953, showing effective performance in distinguishing different chest diseases. Our model has comparable precision and recall to CDC Net, suggesting that while CDC Net is more discriminative for diagnostics, our ensemble approach is as reliable and more widely applicable across multiple conditions.

EfficientNetB0 (Paper [27]): Known for fewer parameters and a streamlined model, it provides lower accuracy in multi-class scenarios (95.00%). Our ensemble model, using a more complex structure, manages to maintain high accuracy without the trade-off seen in streamlined models, indicating our advanced integration techniques can capture nuances in medical images more effectively.

Performance differential analysis: The superiority of our model can be attributed to the use of a stacking ensemble technique for the synergistic integration of multiple types of CNN architectures. Each base model contributes its strengths, which are synthesized by a meta-learner that effectively aggregates the disparate inputs, making the final model more accurate and robust. This approach enables our model to achieve excellent metrics across the board and generalize effectively to the various complexities inherent in different medical imaging tasks.

Our Contribution and limitation: This work provides a new application of ensemble learning with a meta-learner to improve classification in medical images. This method represents a significant improvement over single-model methods that do not consider all important features in complex datasets. However, the performance of our proposed model is highly affected by its complexity, introducing computational and potential overfitting constraints that could be addressed in future work through more optimal model architectures or more effective regularization techniques.

Discussion: next steps: Future endeavors can examine avenues to upgrade the computational efficiency of the ensemble, such as model compression or tweaking the architecture to maintain predictive performance while minimizing redundancy. Additionally, increasing the variability of the dataset, using clinical information from real-world care, or incorporating feedback from medical practice could be essential to optimize an ML model for real-world medical needs.

Conclusions: Our model, which innovatively leverages ensemble learning, achieves state-of-the-art performance in pulmonary disease classification. However, the ever-changing landscape of medical diagnostics means that ongoing improvements and adaptations will be required to remain competitive.

Table 8: Comparison of model's performance

Study Reference	Precision	Recall	F1-score	Accuracy	Unique Aspect
Paper [24]	N/A	N/A	N/A	92.52	Multi-channel fusion CNN
Paper [25]	99.9	98.31	99.09	99.10	Single model for multiple diseases
Paper [26]	99.42	98.13	N/A	99.39	High AUC (0.9953) for diverse diseases
Paper [27]	N/A	N/A	N/A	95.00	EfficientNetB0 with fewer parameters
Paper [28]	N/A	N/A	N/A	99.32	High binary classification accuracy
Paper [29]	N/A	N/A	99.3	99.3	High performance with minimal epochs
Our Work	99.0	99.0	98.0	99.3	Consistently high scores across metrics

8 Conclusion

The present study embodies a collection of rigorous experimentation and investigation to break through the existing medical image classification frontiers. As we meticulously explored the intricate workings of a variety of CNN architectures, including DenseNet201, InceptionV3, Xception, and MobileNet, we simultaneously embarked on a quest to make the most out of their outcomes to transform the face of medical diagnostics. The collection of the curated dataset was subjected to a relentless process of data preprocessing, where we endeavoured to capture every possible complexity and variance that could occur in real-life clinical settings.

All of the steps in our methodology, including prepossessing the data, training the models, and testing them, were carried out thoughtfully and diligently. This involved good parameter selection and aggressive training and testing strategies, as well as the use of prior knowledge through transfer learning, for example from models pre-trained on ImageNet. This led to two positive aspects.

First, all our models trained very quickly. Second, each of our trained models had a very general approach to the diagnostics of chest pathologies, regardless of the specific doctor it was looking for.

In addition, the work of our ensemble technique itself represented the efforts to generate a classifier capable of achieving higher classification performance after optimization. It should be recognized that while the components used in this approach are independent models with a unique speciality or strength and hence very different characteristics, taken together, the total body of the ensemble still demonstrated performance that none of its constituent models could exhibit.

Hence, stacking boosted the reliability of the classifier to discover causal relationships, features, and patterns. It also demonstrated very good performance as it was no worse than most of the highly specialized models, ranging from 80% to 99%, trained for narrow-focused detection of different diseases.

This makes our work both high-end and versatile. Such constant high performance of the AI is exhibited by the method's high efficiency because, given the fairly stable levels of precision, recall, f1-score, and accuracy, the method also shows a high and good level of reliability.

The future work and amplification of our ensemble approach, therefore, may offer substantial improvements for medical diagnostics. Going forward, we aim to improve our methodology to more effectively complement diverse diagnostic use cases, such as emergent disease detection or rare pathologies, which are somewhat underrepresented in our training datasets. We are also investigating new deep learning approaches such as federated learning, allowing the integration of data across institutions without risking patient privacy to include more diverse and higher quality training data, thus, improving the robustness and accuracy of the model. In practice, our model provides significant promise in being seamlessly integrated into current clinical workflows, greatly improving the efficacy and accuracy of diagnostic procedures. Automating the initial chest radiograph analysis can help radiologists, as they can prioritize cases that have the detection of suspected pathologies for a second look, accelerating the diagnosis and possibly improving the patient´s outcome with earlier interventions. Application of our sophisticated machine learning model in clinical practice could potentially decrease the time lag between imaging and diagnosis significantly resulting in timelier interventions. This is especially important for diseases such as COVID-19 and tuberculosis, in which early detection can significantly influence treatment success. In addition to this, Ansermet et al. believe that a model that learns from a large set of pathologies may help in achieving a higher diagnostic rate with better precision and potentially reduce diagnosis error. Future work will be directed at developing interfaces to be smoothly integrated with existing hospital information systems to expedite its adoption in clinical settings. This will involve working extensively with clinical partners to incorporate their feedback into the model as it relates to practicality and optimizing it for different imaging equipment and scenarios. There is the possibility that our model is capable of greatly improving medical diagnostics, not only in future research but in current clinical practice, thus leading to more accurate diagnoses made more quickly, improving patient management and patient outcomes across multiple healthcare scenarios.

References

[1] Ciotti, Marco, Massimo Ciccozzi, Alessandro Terrinoni, Wen-Can Jiang, Cheng-Bin Wang, and Sergio Bernardini. The COVID-19 pandemic. Critical reviews in clinical laboratory sciences, 57(6): 365-388, 2020. https://doi.org/10.1080/10408363.2020.1783198

[2] Wilson, Nick, Stephen Corbett, and Euan Tovey. Airborne transmission of covid-19. bmj, 370, 2020. https://doi.org/10.1136/bmj.m3206

[3] Kevadiya, Bhavesh D., Jatin Machhi, Jonathan Herskovitz, Maxim D. Oleynikov, Wilson R. Blomberg, Neha Bajwa, Dhruvkumar Soni et al. Diagnostics for SARS-CoV-2 infections. Nature materials, 20(5): 593-605, 2021. https://doi.org/10.1038/s41563-020-00906-z

[4] Gnanvi, Janyce Eunice, Kolawolé Valère Salako, Gaëtan Brezesky Kotanmi, and Romain Glèlè Kakaï. On the reliability of predictions on Covid-19 dynamics: A systematic and critical review of modelling techniques. Infectious Disease Modelling, 6: 258-272, 2021.

https://doi.org/10.1016/j.idm.2020.12.008

[5] Johnson, Kemmian D., Christen Harris, John K. Cain, Cicily Hummer, Hemant Goyal, and Abhilash Perisetti. Pulmonary and extra-pulmonary clinical manifestations of COVID-19."Frontiers in medicine, 7: 526, 2020. https://doi.org/10.3389/fmed.2020.00526

[6] Marginean, Cristina Maria, Mihaela Popescu, Corina Maria Vasile, Ramona Cioboata, Paul Mitrut, Iulian Alin Silviu Popescu, Viorel Biciusca et al. Challenges in the differential diagnosis of COVID-19 pneumonia: a pictorial review. Diagnostics, 12(11): 2823, 2022. https://doi.org/10.3390/diagnostics12112823

[7] Bhatnagar, Rahul, and Nick Maskell. The modern diagnosis and management of pleural effusions. bmj, 351, 2015. https://doi.org/10.1136/bmj.h4520

[8] Chen, Chia-Hung, Chih-Kun Chang, Chih-Yen Tu, Wei-Chih Liao, Bing-Ru Wu, Kuei-Ting Chou, Yu-Rou Chiou, Shih-Neng Yang, Geoffrey Zhang, and Tzung-Chi Huang. Radiomic features analysis in computed tomography images of lung nodule classification. PloS one, 13(2): e0192002, 2018. https://doi.org/10.1371/journal.pone.0192002

[9] Hussain, Azhar, Alia Noorani, Ranjit Deshpande, Lindsay John, Max Baghai, Olaf Wendler, Donald Whitaker, and Habib Khan. Management of pneumothorax in mechanically ventilated COVID-19 patients: early experience. Interactive CardioVascular and Thoracic Surgery, 31(4), 540-543, 2020. https://doi.org/10.1093/icvts/ivaa129

[10] Abdulahi, AbdulRahman Tosho, Roseline Oluwaseun Ogundokun, Ajiboye Raimot Adenike, Mohd Asif Shah, and Yusuf Kola Ahmed. PulmoNet: a novel deep learning based pulmonary diseases detection model. BMC Medical Imaging, 24(1), 51, 2024. https://doi.org/10.1186/s12880-024-01227-2

[11] Greenspan, Hayit, Bram Van Ginneken, and Ronald M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE transactions on medical imaging 35(5): 1153-1159, 2016. https://doi.org/10.1109/TMI.2016.2553401

[12] Miotto, Riccardo, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics, 19(6): 1236-1246, 2018. https://doi.org/10.1093/bib/bbx044

[13] Nguyen, Thi Mai, Nackhyoung Kim, Da Hae Kim, Hoang Long Le, Md Jalil Piran, Soo-Jong Um, and Jin Hee Kim. Deep learning for human disease detection, subtype classification, and treatment response prediction using epigenomic data. Biomedicines, 9(11), 1733, 2021. https://doi.org/10.3390/biomedicines9111733

[14] Aslan, Muhammet Fatih, Kadir Sabanci, Akif Durdu, and Muhammed Fahri Unlersen.COVID-19 diagnosis using state-of-the-art CNN architecture features and Bayesian Optimization. Computers in biology and medicine, 142: 105244, 2022. https://doi.org/10.1016/j.compbiomed.2022.105244

[15] Fatani, Abdulaziz, Abdelghani Dahou, Mohammed AA Al-Qaness, Songfeng Lu, and Mohamed Abd Elaziz. Advanced feature extraction and selection approach using deep learning and Aquila optimizer for IoT intrusion detection system. Sensors, 22(1):140, 2021. https://doi.org/10.3390/s22010140

[16] Elharrouss, Omar, Younes Akbari, Noor Almadeed, and Somaya Al-Maadeed. Backbones-review: Feature extractor networks for deep learning and deep reinforcement learning approaches in computer vision. Computer Science Review, 53: 100645, 2024. https://doi.org/10.1016/j.cosrev.2024.100645

[17] Akinyelu, Andronicus A., and Pieter Blignaut. COVID-19 diagnosis using deep learning neural networks applied to CT images. Frontiers in Artificial Intelligence, 5: 919672, 2022. https://doi.org/10.3389/frai.2022.919672

[18] Aljondi, Rowa, and Salem Alghamdi. Diagnostic value of imaging modalities for COVID-19: scoping review. Journal of medical Internet research, 22(8): e19673, 2020. https://doi.org/10.2196/19673

[19] Benmalek, Elmehdi, Jamal Elmhamdi, and Abdelilah Jilbab. Comparing CT scan and chest X-ray imaging for COVID-19 diagnosis. Biomedical Engineering Advances, 1, 100003: 2021. https://doi.org/10.1016/j.bea.2021.100003

[20] Sun, Junding, Pengpeng Pi, Chaosheng Tang, Shui-Hua Wang, and Yu-Dong Zhang. CTMLP: Can MLPs replace CNNs or transformers for COVID-19 diagnosis? Computers in Biology and Medicine, 159, 106847, 2023. https://doi.org/10.1016/j.compbiomed.2023.106847

[21] Li, Wenjing, Randy C. Paffenroth, and David Berthiaume. Neural network ensembles: theory, training, and the importance of explicit diversity. arXiv preprint arXiv:2109.14117, 2021. https://doi.org/10.48550/arXiv.2109.14117

[22] Salehi, Ahmad Waleed, Shakir Khan, Gaurav Gupta, Bayan Ibrahimm Alabduallah, Abrar Almjally, Hadeel Alsolai, Tamanna Siddiqui, and Adel Mellit. A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. Sustainability, 15(7), 5930, 2023. https://doi.org/10.3390/su15075930

[23] Brownlee, Jason. Stacking ensemble for deep learning neural networks in python. 2018.

[24] Nikolaou, Vasilis, Sebastiano Massaro, Masoud Fakhimi, Lampros Stergioulas, and Wolfgang Garn. COVID-19 diagnosis from chest x-rays: developing a simple, fast, and accurate neural network. Health information science and systems, 9: 1-11, 2021. https://doi.org/10.1007/s13755-021-00166-4

[25] Hira, Swati, Anita Bai, and Sanchit Hira. An automatic approach based on CNN architecture to detect COVID-19 disease from chest X-ray images. Applied Intelligence, 51: 2864-2889, 2021. https://doi.org/10.1007/s10489-020-02010-w

[26] AbdElhamid, Abeer A., Eman AbdElhalim, Mohamed A. Mohamed, and Fahmi Khalifa. Multi-classification of chest X-rays for COVID-19 diagnosis using deep learning algorithms. Applied Sciences, 12(4): 2080, 2022. https://doi.org/10.3390/app12042080

[27] Qian, Xuelin, Huazhu Fu, Weiya Shi, Tao Chen, Yanwei Fu, Fei Shan, and Xiangyang Xue. M $^ 3$ Lung-Sys: A deep learning system for multi-class lung pneumonia screening from CT imaging. IEEE journal of biomedical and health informatics, 24(12): 3539-3550, 2020. https://doi.org/10.1109/JBHI.2020.3030853

[28] Nath, Malaya Kumar, Aniruddha Kanhe, and Madhusudhan Mishra. A novel deep learning approach for classification of COVID-19 images. In 2020 IEEE 5th international conference on computing communication and automation (ICCCA): 752-757. IEEE, 2020. https://doi.org/10.1109/ICCCA49541.2020.9250907

[29] Islam, Md Nazmul, Md Golam Rabiul Alam, Tasnim Sakib Apon, Md Zia Uddin, Nasser Allheeib, Alaa Menshawi, and Mohammad Mehedi Hassan. "Interpretable differential diagnosis of non-covid viral pneumonia, lung opacity and covid-19 using tuned transfer learning and explainable ai. In Healthcare, 11(3): 410. MDPI, 2023. https://doi.org/10.3390/healthcare11030410

[30] Ullah, Zaka, Ayman Odeh, Ihtisham Khattak, and Muath Al Hasan. Enhancement of Pre-Trained Deep Learning Models to Improve Brain Tumor Classification. Informatica, 47(6): 165–172, 2023. https://doi.org/10.31449/inf.v47i6.4645

[31] Cherifi, Dalila, Abderraouf Djaber, Mohammed-Elfateh Guedouar, Amine Feghoul, Zahia Zineb Chelbi, and Amazigh Ait Ouakli. Covid-19 Detecting in Computed Tomography Lungs Images using Machine and transfer Learning. Informatica, 47(8): 35–44, 2023. https://doi.org/10.31449/inf.v47i8.4258