Decision Analytics Journal 12 (2024) 100499

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

An explainable artificial intelligence model for multiple lung diseases classification from chest X-ray images using fine-tuned transfer learning

Eram Mahamud [a], Nafiz Fahad [b,d], Md Assaduzzaman [a], S.M. Zain [a], Kah Ong Michael Goh [b,*], Md. Kishor Morol [c,d]

[a] Department of CSE, Daffodil International University, Dhaka, Bangladesh
[b] Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450, Bukit Beruang, Melaka, Malaysia
[c] Department of Computing and Information Science, Cornell University, NY, United States of America
[d] EliteLab.Ai, NY, United States of America

ARTICLE INFO

Keywords:
Transfer learning
X-ray diagnosis
Explainable artificial intelligence (XAI)
Data augmentation
Respiratory illnesses

ABSTRACT

Traditional deep learning models are often considered "black boxes" due to their lack of interpretability, which limits their therapeutic use despite their success in classification tasks. This study aims to improve the interpretability of diagnoses for COVID-19, pneumonia, and tuberculosis from X-ray images using an enhanced DenseNet201 model within a transfer learning framework. We incorporated Explainable Artificial Intelligence (XAI) techniques, including SHAP, LIME, Grad-CAM, and Grad-CAM++, to make the model's decisions more understandable. To enhance image clarity and detail, we applied preprocessing methods such as Denoising Autoencoder, Contrast Limited Adaptive Histogram Equalization (CLAHE), and Gamma Correction. An ablation study was conducted to identify the optimal parameters for the proposed approach. Our model's performance was compared with other transfer learning-based models like EfficientNetB0, InceptionV3, and LeNet using evaluation metrics. The model that included data augmentation techniques achieved the best results, with an accuracy of 99.20%, and precision and recall of 99%. This demonstrates the critical role of data augmentation in improving model performance. SHAP and LIME provided significant insights into the model's decision-making process, while Grad-CAM and Grad-CAM++ highlighted specific image features and regions influencing the model's classifications. These techniques enhanced transparency and trust in AI-assisted diagnoses. Finally, we developed an Android-based system using the most effective model to support medical specialists in their decision-making process.

1. Introduction

Given their widespread impact and high fatality rates, illnesses like pneumonia, tuberculosis (TB), and COVID-19 pose significant threats to global health. In 2019, chronic respiratory diseases caused 4.0 million deaths worldwide, highlighting the essential need for appropriate diagnostic solutions [1]. Despite significant breakthroughs in artificial intelligence and deep learning for medical diagnostics, accurately and promptly identifying lung diseases remains a challenging endeavor. The COVID-19 pandemic underscored the crucial importance of promptly and precisely classifying diseases to manage and control outbreaks effectively. It is crucial to quickly identify COVID-19 cases from other lung problems because of its fast spread and varied symptoms. Tuberculosis (TB) [2] and pneumonia remain significant health concerns, particularly in resource-limited areas. Predicting diseases accurately and timely is essential to avoid misdiagnoses, delayed treatment, and to prevent healthcare systems from becoming overwhelmed.

Medical professionals and computer experts have collaborated to develop advanced computer-aided diagnostic systems [3]. Utilizing AI to replicate human decision-making processes, these systems have the potential to improve diagnostic accuracy. Deep learning has shown significant success in classifying lung diseases from chest X-rays by learning from large datasets. For instance, a hybrid model using a modified capsule network for lung disease detection achieved an accuracy of 73%, and a multi-scale capsule network for detecting lung nodules achieved an accuracy of 83% on the LIDC datasets [4,5]. These results highlight the ongoing challenge and the potential for further improvements in using deep learning models for lung disease detection.

Early disease identification and appropriate intervention are pivotal elements in the successful management of lung disorders. Therefore, comprehending the origins and consequences of these ailments is crucial for an effective approach.

Fig. 1 illustrates the differences between normal lungs, lungs affected by COVID-19, TB-infected lungs, and lungs affected by pneumonia in chest X-ray images.

* Corresponding author.
E-mail address: michael.goh@mmu.edu.my (K.O.M. Goh).

https://doi.org/10.1016/j.dajour.2024.100499
Received 13 May 2024; Received in revised form 24 June 2024; Accepted 29 June 2024
Available online 2 July 2024

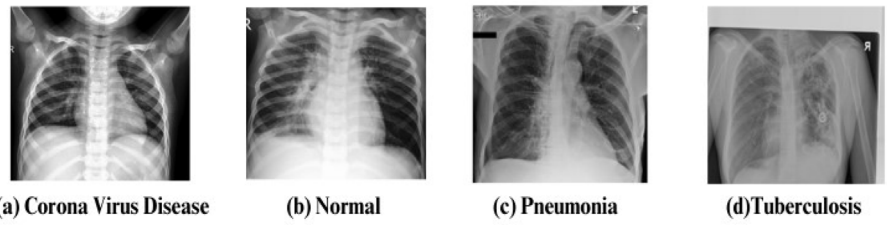


Fig. 1. Sample images (random) for each class of lungs disease.

1.0.1. Corona virus disease (Covid-19)

COVID, short for "Coronavirus Disease 2019", is caused by the highly contagious virus SARS-CoV-2. Primarily affecting the respiratory system, [6] COVID-19 can manifest various symptoms and complications. This acronym encapsulates the global health crisis that emerged in 2019, highlighting the severity of the respiratory syndrome caused by the novel coronavirus.

1.0.2. Normal

An unaltered, clean, well-defined lung field is visible on an X-ray of a healthy lung [7]. Here are some of things that requires to identify as a healthy lungs Clear Lung Fields, Normal Lung Size, Sharp Lung Borders, Visible Lung Markings, Normal Diaphragm, Absence of Abnormal Shadows, Healthy Ribs and Spine etc [8]. Overall, a healthy lung X-ray provides a clear view of the lung tissues, demonstrating normal lung structure, size, and function.

1.0.3. Pneumonia

Pneumonia is a lung infection that can have various causes and manifestations. It primarily leads to inflammation and fluid or pus accumulation in the lung's air sacs, [9] making breathing difficult and compromising oxygen exchange. The identification of pneumonia typically involves clinical assessment, medical history analysis, and imaging tests such as chest X-rays or computed tomography (CT) scans. Additionally, blood tests or respiratory secretions analysis may be performed to determine the specific underlying cause, [10] which could be viral or bacterial in nature. Accurate diagnosis of pneumonia is essential for effective treatment and forms a critical aspect of this research on deep learning-based prediction of lung diseases.

1.0.4. Tuberculosis

Mycobacterium tuberculosis, a bacterium renowned as the causative agent of the infectious and highly contagious disease tuberculosis (TB), primarily targets the lungs but can also affect vital organs such as the kidneys, spine, and brain [11]. Tuberculosis stands as one of the earliest documented human diseases and continues to pose a significant global health threat [12]. The detection of latent TB infection involves employing diagnostic tools such as the Tuberculin skin test or interferon-gamma release assay. Meanwhile, active TB disease diagnosis relies on the utilization of chest X-rays and sputum culture. It is crucial to note that a positive skin or blood test merely confirms exposure to TB germs [13], and it does not definitively establish the presence of active TB disease.

Therefore, this current study introduces an explainable transfer learning model for classifying lung diseases to address these challenges. We aim to enhance diagnostic accuracy and ensure transparency in the AI's decision-making process through our research. This study not only highlights the strength and reliability of the model, but also explores the field of explainable AI by utilizing techniques like SHAP, LIME, Grad-CAM, and Grad-CAM++ to clarify the model's decision-making process. Our goal is to connect advanced computational techniques with the practical needs of medical diagnostics, making a significant contribution to the field. Moreover, our contribution is mentioned below

I. Several image preprocessing techniques are utilized to eliminate artifacts and improve image quality, including Morphological Opening, Gamma Correction, CLAHE, Bilateral, and Spectrum.

II. Performed an ablation study to determine the most effective custom layers, fine-tuning techniques, and data augmentation strategies to improve the proposed model for detecting different lung diseases

III. Employed SHAP and LIME, two prominent interpretable AI techniques, to create saliency maps and identify the vital features that impact the model's decision.

IV. The proposed model's interpretability was enhanced by utilizing Grad-CAM and Grad-CAM++, which generated heatmaps that enhanced transparency and provided insights for better understanding.

V. The proposed model has a dropout layer to prevent overfitting and adjusted the last Dense layer for the four-class classification task. deployed in Android-based applications to aid healthcare professionals in the classification of lung diseases

However, this current study discussion is structured as follows: the paper begins with a review of existing literature on lung disease classification using deep learning techniques. Our paper begins with a literature overview of deep learning applications in medical imaging and the unique issues of lung disease categorization. We then organize our discussion according to this outline. After that, it describes our methodology section and shows our experimental findings which combine explainability with transfer learning.

2. Related works

Several studies have investigated the application of diverse deep learning models in the categorization and diagnosis of lung diseases. For instance, Hamal et al. [14] conducted a comparative analysis of traditional machine learning algorithms, such as K-Nearest Neighbor (KNN) and decision trees (DT), alongside advanced deep learning models to classify X-ray images into COVID-19 patients, pneumonia patients, and healthy individuals. While traditional models showed promise, deep learning models ultimately outperformed them, highlighting the superior pattern recognition capabilities of deep learning. However, the study primarily focused on traditional machine learning models, potentially limiting the capture of complex patterns as effectively as deep learning models. Additionally, the dataset size and diversity were not specified, which could affect the generalizability of the results. Similarly, Bondugula et al. [15] introduced a transfer learning-based deep neural architecture aimed at minimizing false negatives in infectious disease diagnosis, showcasing impressive accuracy. However, the model's complexity may hinder deployment in resource-constrained environments. Khanna et al. [16] developed a triage-prediction system for COVID-19 severity using Random Forest on Borderline SMOTE balanced data, achieving an 83% recall rate. While successful, the reliance on balanced data might not fully address real-world data imbalances, limiting its applicability. Tian et al. [17] developed a deep learning model using DenseNet-161 with a hybrid attention mechanism to classify pancreatic cystic tumors as either serous cystic neoplasms (SCN) or mucinous cystic neoplasms (MCN). The model, which incorporated clinical features with T2-weighted MRI images, achieved a classification accuracy of 92.44%, an AUC of 0.971,

precision of 0.956, recall of 0.919, specificity of 0.933, and an F1-score of 0.936. While innovative, the study's limitations include a small, specific dataset and increased computational complexity. Additionally, the model's interpretability for clinical decision-making could be improved. Nalluri et al. [18] proposed a hybrid classification model for early detection and differentiation of pneumonia from other lung diseases. This model integrated text data on symptoms and signs with X-ray images, achieving accuracy levels of 85.13%, 87.99%, 92.28%, and 95.28% for various learning percentages (60%, 70%, 80%, and 90%, respectively). The study highlighted the model's superior performance in terms of accuracy, sensitivity, specificity, and error rates. However, the model's performance depends on the quality and completeness of the text data, which can be variable in clinical settings. Chutia et al. [19] presented an advanced method for classifying lung diseases using an enhanced DenseNet201 model, achieving high accuracy with the CLAHE technique. However, the study's enhancements might increase computational complexity, potentially limiting real-time application. Malik et al. [20] introduced multi-modal deep learning methods for classifying chest diseases, demonstrating high accuracy but facing challenges in data integration. Mishra et al. [21] addressed the imbalanced class problem in lung cancer detection using convolutional neural networks (CNNs), achieving a high accuracy of 99.4% with DenseNet-121 utilizing SMOTE analysis. However, the study primarily focused on class imbalance, potentially neglecting other biases in the data and generalizability to diverse datasets. Bennour et al. [22] presented deep learning models for diagnosing pulmonary diseases, achieving high accuracies but encountering challenges related to generalization and dataset bias. Niño et al. [23] compared the performance of DenseNet, VGG19, and ResNet50 in pneumonia detection, with ResNet50 achieving the highest accuracy. However, limitations include a relatively small dataset size, potential shortcomings in evaluation metrics, and uncertainty regarding the model's generalizability. Building on this progress in artificial intelligence and machine learning for lung cancer detection, several studies explore advancements in deep learning architectures for improved accuracy and efficiency. Studies by [24,25] propose novel 3D CNN frameworks for lung nodule segmentation and classification in CT scans, achieving high Dice Similarity Coefficients (DSC) and accuracy rates. While these methods outperform traditional techniques, interpretability for clinical adoption remains a challenge [26]. Another approach investigates explainable deep learning for chest X-ray analysis, mimicking radiologist decision-making by identifying discriminative regions and achieving high accuracy. Furthermore, [27] explores the potential of traditional machine learning techniques like Support Vector Machines (SVM) and Random Forest (RF) in conjunction with CNNs for lung cancer diagnosis using CT scans. Their findings suggest that SVM and RF can achieve an accuracy of 95%, highlighting the continued role of these techniques alongside deep learning. Overall, deep learning offers promising avenues for lung cancer diagnosis, but further research is needed to address interpretability and generalizability for wider clinical use.

In comparison to the existing related works, our proposed deep learning model, stands out as the most robust and effective solution for lung disease prediction from chest X-ray images. Several key aspects differentiate our model and contribute to its superiority. The proposed deep learning model excels in lung disease prediction, showcasing superiority in accuracy, versatility, and resilience to data scarcity. Commitment to transparency and open access positions it as a premier choice in advancing respiratory health diagnostics. Augmentation techniques and ablation studies fine-tune key parameters, optimizing model training. Incorporation of interpretability tools, such as SHAP, LIME, Grad-CAM, and Grad-CAM++, enhances the understanding of model decisions. Deployment of the optimized model in an Android application empowers healthcare professionals for precise lung disease classification. This deployment empowers healthcare professionals by providing a robust tool for the precise classification of lung diseases, facilitating early interventions and ultimately contributing to improved patient outcomes. This new method fills an important research gap in quickly and easily detecting lung disease fractures. It also shows potential for reducing the serious consequences linked to medical emergencies related to the lung related issues.

Table 1
The lungs disease augmented dataset.

Class Name	Count	Testing	Training	Validation
Covid-19	5871	503	4867	501
Pneumonia	6427	512	5406	509
Normal	5905	508	4894	503
Tuberculosis	5851	504	4845	502

Table 2
The lungs disease original dataset.

Class Name	Count	Testing	Training	Validation
Covid-19	2510	503	1506	501
Pneumonia	2561	512	1540	509
Normal	2507	508	1495	503
Tuberculosis	2509	504	1503	502

3. Methodology

The proposed approach includes various steps, systematically represented in Fig. 2. The process begins with data acquisition, followed by key preprocessing steps [28]. These include image augmentation, resizing, and pixel normalization. The dataset undergoes segmentation and denoising through autoencoder-based methods. Contrast Limited Adaptive Histogram Equalization (CLAHE) is then applied, followed by gamma correction on the enhanced images. Subsequently, models, including InceptionV3, LeNet, EfficientNetB0, and the proposed model, are constructed and tested. The chosen models were selected based on superior accuracy. Training and testing were performed, and result evaluation involved classification, Univariate Feature Selection, and Recursive Feature Elimination, leading to the identification of the best classifier. Throughout this process, the proposed model, InceptionV3, LeNet, and EfficientNetB0 were employed.

3.1. Dataset

The study revolves around a dataset named "Lungs Disease Dataset" [29]. It involved the amalgamation of data from various sources, including renowned datasets available on platforms such as Kaggle. Through the augmentation of the training dataset, we have crafted a division wherein the training data encompasses roughly 80% of the total images, the validation data contributes approximately 10%, and the testing data mirrors this allocation with another 10%. As depicted in Table 1, the augmented training dataset comprises 20,012 sample images, while the test dataset comprises 2027 images, and the validation dataset holds 2015 images. This meticulous partitioning ensures a thorough evaluation, providing a holistic understanding of the model's performance across varied scenarios. This intentional approach not only enhances the depth of the study but also bolsters its reliability.

Table 4 in our study highlights a dataset centered on lung CXR images, featuring a collection of 10,000 images thoughtfully categorized into four distinct classes. Notably, this Table 4 serves as the original dataset from which we derived the content for Table 2. Within this context, one class showcases images of healthy lungs CXR, while the remaining classes spotlight various lung diseases. Adding to the dataset's richness, the images exhibit a diversity of formats, including jpg, png, and jpeg.

The train dataset contains 6024 sample images, and 2027 images are present in the test dataset, and 2015 images are in the validation dataset.

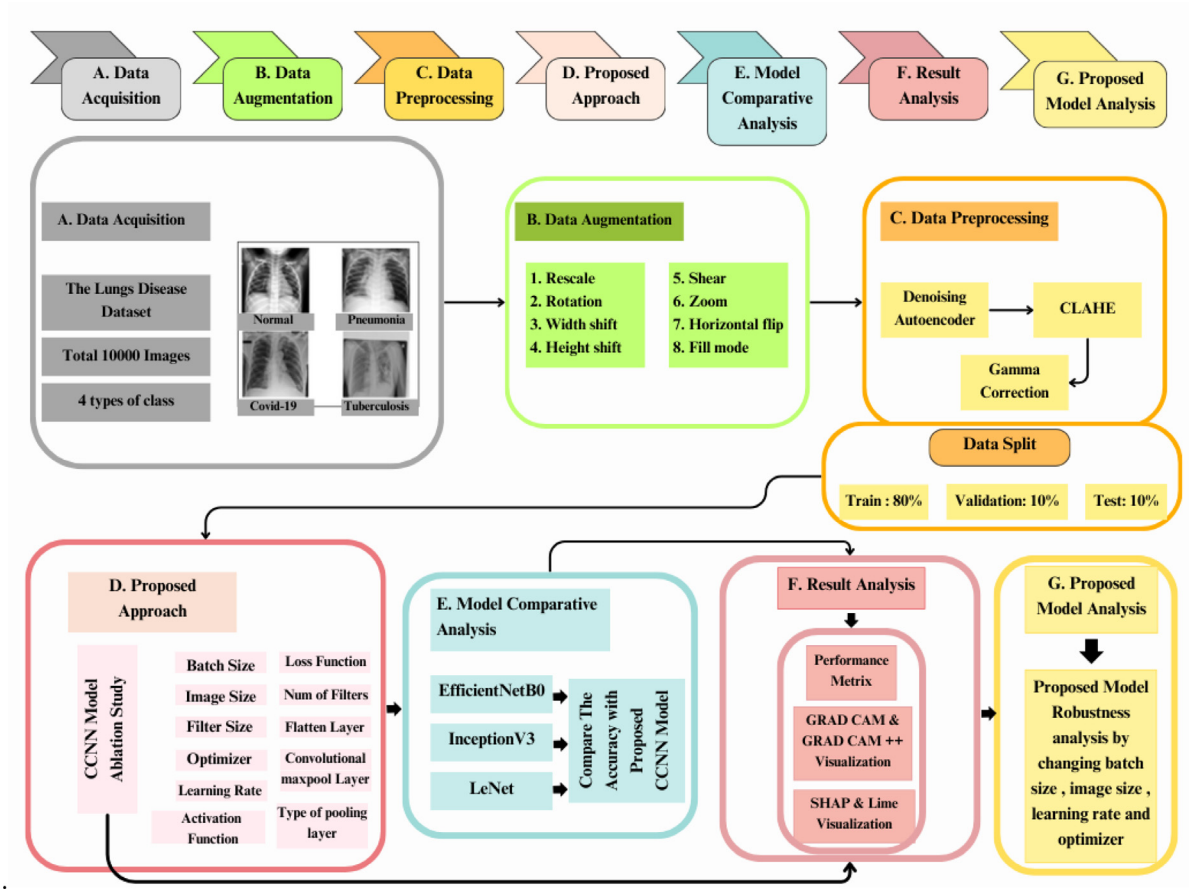


Fig. 2. Diagram of the implementation step for classifying lungs diseases.

Table 3
Hyperparameter for image data augmentation.

Techniques	Parameter
Shear range	0.2 or 20.00%
Zoom range	0.2 or zoom in and zoom out by 20.00%
Horizontal flip	True
Width shift range	0.1 or 10.00%
Height shift range	0.1 or 10.00%
Fill mode	nearest
Rescale	1/.255
Rotation	0.2 or 20.00%

3.2. Augmentation

The images in the dataset have been resized to a dimension of 224 × 224 pixels. This resizing expedites the training process and ensures computational feasibility. Image augmentation [30], a pivotal technique in deep learning, enhances the diversity of the training dataset by applying transformations to original images. This includes rotating images within a specified range (20 degrees), shifting horizontally and vertically (10% of image width and height), shearing (20 degrees), zooming (20%), and horizontal flipping. The ImageDataGenerator dynamically generates augmented images during model training [31]. Pixel values are rescaled to the range [0,1] (achieved through rescale=1./255) to ensure suitability for neural network training. Table 3 shows the hyperparameters employed in this study (see Fig. 3).

The augmentation techniques applied have contributed to several key benefits in the training process, such as improved generalization and reduced overfitting. This strategy has proven particularly advantageous for achieving better performance on imbalanced datasets [32, 33].

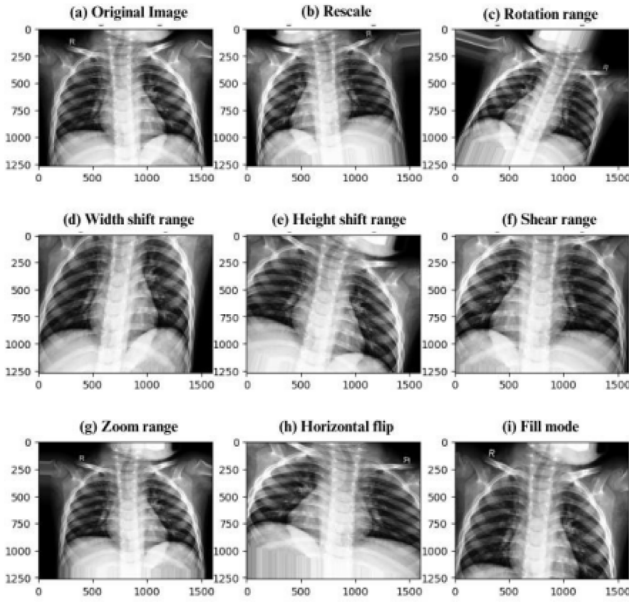


Fig. 3. Augmented Image.

3.3. Denoising autoencoder

Denoising Autoencoders (DAEs) [34] have been used for refining X-ray images for lung disease classification. DAEs act like digital image doctors, adept at noise reduction and feature enhancement. During

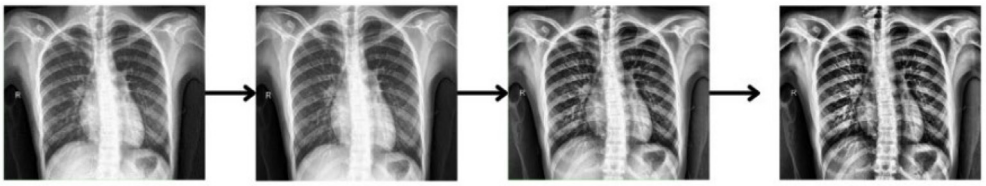


Fig. 4. Preprocessed Images.

training, random noise was introduced into the input data, making the model an expert in handling noisy input. The model aimed to minimize a loss function, improving performance. The validation loss evaluated the disparity between model predictions and actual ground truth, guiding the training process [35].

$$\text{process } J = 1/N \sum Ni = 1\mathfrak{L}(yi, yi) \tag{1}$$

The validation loss served as the scoreboard, evaluating the disparity between model predictions ("yi") and actual ground truth ("yi"), guiding the training process.

3.4. Image enhancement (CLAHE)

Contrast Limited Adaptive Histogram Equalization (CLAHE) [36] is a powerful image enhancement technique that improves image contrast. CLAHE divides an image into tiles, computes histograms for each, and adjusts pixel value distributions based on a clip limit parameter. The number of mean values of clipped pixels is calculated to redistribute leftover pixels effectively [37]. CLAHE is effective in enhancing local contrast and is valuable in various medical applications [38,39]. Developed for the total number of tiles, denoted as T, is calculated using Equation: The construction [40] of histograms for these tiles takes into account a clip limit, CL, as specified in Eq. (2):

$$T = N \times N/(n \times n) \tag{2}$$

Here, [41] NCL represents the Normalized Contrast Limit, and $NAVG$ is the average count of pixels in the image. The ratio of $NAVG$ is determined using Eq. (3):

$$NAVG = (Nx \times Ny)/Ng \tag{3}$$

Here, Ng, Nx, and Ny represent the number of grayscale levels, the pixel size in the x-dimension, and the pixel size in the y-dimension, respectively, within the tile. The number of mean values of clipped pixels, denoted as Ncp, is calculated using Eq. (4):

$$Ncp = N \sum CL/Ng \tag{4}$$

Here, CL represents clipped pixels, and $\sum CL$ is the overall number of clipped pixels. After the calculation of Ncp, some pixels remain unallocated. To redistribute these leftover pixels, Eq. (5) is employed:

$$R = (Ng)/Nr \tag{5}$$

3.5. Gamma correction

Gamma Correction [42] adjusts brightness and contrast, addressing exposure and luminance issues for accurate medical image interpretation.

$$I\_\sin = I\_\text{orig}\hat{}(1/\gamma) \tag{6}$$

Controlled by the gamma (γ) [43] value, Gamma Correction fine-tunes image tones. A gamma value greater than 1 enhances contrast and darkens the image, while a value less than 1 reduces contrast and brightens the image. Experimentation determined the optimal gamma value for the dataset, ensuring visually appealing and diagnostically informative images. Gamma Correction proves invaluable in enhancing X-ray images for precise analysis in lung disease classification. Carefully adjusting the gamma value enhances contrast, making images visually improved and conducive to accurate diagnosis (see Fig. 4).

Table 4
Training parameter for the proposed model.

Parameter	Description
Optimization algorithm	Adam optimizer
Learning rate (α)	0.000001
Weight initialization	"ImageNet""
Batch size	64
Number of epochs	100
Dropout rate	0.5 or 50%
Loss function	Categorical cross entropy
Activation function (Hidden layers)	Relu
Activation function (Output layer)	SoftMax

3.6. Proposed model

Fig. 5 presents a systematic depiction of our proposed model architecture. The input layer of the proposed model is designed to process images of a fixed size, specifically 224 × 224 pixels with 3 color channels (RGB). The proposed model is distinctive for its dense blocks, which comprise multiple interconnected convolutional layers. Unlike conventional convolutional networks, where each layer's output feeds into the next layer, the proposed model ensures that each layer within a dense block receives inputs from all preceding layers in the same block. This dense connectivity promotes feature reuse and gradient flow, enhancing the network's efficiency and training ease [44]. Within each dense block, a series of convolutional layers is employed, with each layer incorporating batch normalization and Rectified Linear Unit (ReLU) activation functions. To enhance computational efficiency, a 1 × 1 convolutional layer precedes every 3 × 3 convolutional layer within a dense block, reducing the number of input feature maps. Additionally, a dropout layer with a dropout rate of 0.5 is introduced after the flattening layer to prevent overfitting during training. Transition layers, featuring batch normalization, a 1 × 1 convolutional layer to reduce channel numbers, and a 2 × 2 average pooling layer for spatial dimension down sampling, separate dense blocks [45]. The growth rate parameter 'k' determines the number of feature maps produced by each layer in a dense block, influencing network capacity. Following the dense blocks, one or more fully connected layers process the flattened feature maps for classification. In this modified model, two dense layers with 512 and 256 neurons are introduced after the dense blocks. These layers contribute to the final output layer, utilizing a softmax activation function for multi-class classification. The output layer's neuron count aligns with the number of classes in the classification problem. The architecture of the proposed model, characterized by dense blocks, transition layers, a defined growth rate, and the incorporation of dropout layers, proves effective for feature extraction and classification tasks, especially in scenarios with limited data. The modifications introduced aim to enhance the model's generalization capability and mitigate overfitting during training.

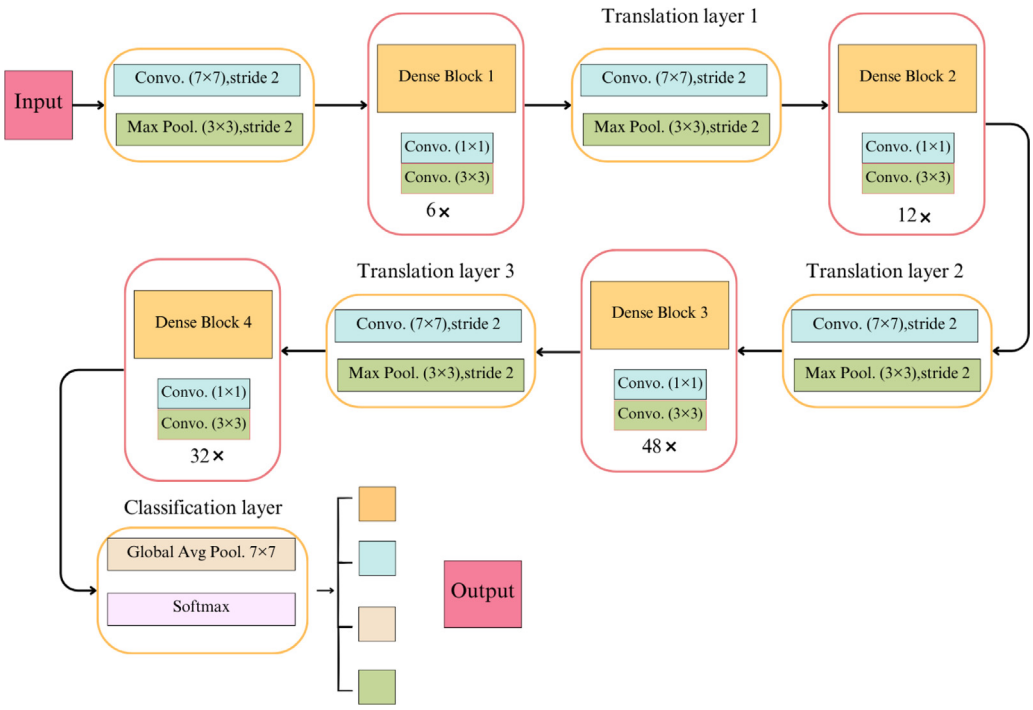


Fig. 5. Proposed model architecture (a systematic depiction of the proposed system architecture).

To quantify the disparity between predicted and actual probability distributions, the model utilized the Categorical Cross-Entropy loss function, apt for multi-class classification tasks.

$$\text{Cross-Entropy Loss} = -\sum_{i=1}^{n} y_i \log\left(p_i\right) \tag{7}$$

Additionally, the Max pooling layer was employed for down-sampling, reducing the spatial dimensions of the convolved feature maps while preserving salient features.

$$\text{Max Pooling}: \text{out}\,[i,j] = \underset{m,n}{max}\,\text{input}[i \times s + m, j \times s + n] \tag{8}$$

The model was trained for 100 epochs, handling 64 images at a time in each batch. This batch size was not picked randomly—it came from a careful study on how batch size impacts model performance. The goal was to find the best batch size that minimizes validation loss. To ensure we got the best model, a neat trick was used. We stored the model with the lowest validation loss using a handy callbacks function. This way, we secured the most effective model configuration observed during training. It is like saving the best version of the model after trying out different settings.

3.7. EfficientNetB0

The EfficientNetB0 model [46] is initialized with weights from the ImageNet dataset which was already pre-trained, and its layers are then set as non-trainable to retain the acquired knowledge. A distinctive classification head is appended, incorporating a Global Average Pooling layer for spatial dimension reduction, a Dropout layer with a 50% rate to curb overfitting, a dense layer featuring 512 units with ReLU activation, and a final dense layer with 4 units (presumed for a 4-class classification task) and a softmax activation for generating class probabilities. The resulting model, combining the frozen EfficientNetB0 base and the customized classification head, is compiled using the Adam optimizer, categorical crossentropy as the loss function, and accuracy as the evaluation metric. This configuration establishes a transfer learning framework, harnessing the pre-trained knowledge of EfficientNetB0 while tailoring the model for a specific image classification task.

3.8. InceptionV3

The InceptionV3 model [47] is preloaded with weights trained on the ImageNet dataset, and its layers are subsequently frozen to retain the learned features. The custom classification head is then appended, consisting of a flattening layer to reshape the output, a dropout layer with a 50% rate to mitigate overfitting, a dense layer with 512 units and ReLU activation, and a final dense layer with 4 units (assuming a 4-class classification task) and a softmax activation for generating class probabilities. The entire model, combining the frozen InceptionV3 base and the new classification head, is compiled using the Adam optimizer, categorical crossentropy as the loss function, and accuracy as the evaluation metric. This configuration establishes a transfer learning framework, leveraging InceptionV3's pre-trained knowledge while tailoring the model for a specific image classification task.

3.9. LeNet

The LeNet architecture [48] follows a classical pattern with convolutional and pooling layers, leading to a dense classification head. The model begins with a 3×3 convolutional layer with 32 filters and a ReLU activation function, followed by a max-pooling layer with a 2×2 window. This pattern is repeated twice with 64 and 128 filters, each followed by max-pooling. The spatial dimensions are progressively reduced, and the extracted features are flattened before entering two dense layers. The first dense layer has 512 units with a ReLU activation function. The final dense layer consists of 4 units with a softmax activation, suitable for multi-class classification tasks. The model is compiled using the Adam optimizer, categorical crossentropy as the loss function (indicating a multi-class classification task), and accuracy as the evaluation metric.

4. Environmental specification

On a computer running Windows 11, 11th Gen Intel(R) Core (TM) i3-1115G4 @ 3.00 GHz, and 8 GB of RAM are used to carry out our implementation. An open-source platform called Google Collab Notebook is used to execute the model. The Kera's and TensorFlow frameworks are utilized to put the model into action (see Table 5).

Table 5
Experimental setup.

Process name	S. N.	Action
Input	1.	Collected images of 4 classes of lungs disease CXR images, including 3 classes of disease.
Environment	2.	Google Colab.
Configuration	3.	Import all necessary libraries and packages.
	4.	Load the images.
Directories Configuration	5.	Load directories for training, testing, and validation.
Training and Testing	6.	Build the proposed models. For transfer learning, use models trained on the ImageNet dataset.
	7.	Fine-tune the models by adding the fully connected layer and the SoftMax activation function.
Model Compilation	8.	The model compiles with an Adam optimizer and a learning rate of 0.000001.
	9.	Set 100 epochs for model training.
	10.	As a model checkpoint, use the validation accuracy to monitor.
	11.	Save model.
Performance Evaluation	12.	Generate classification report.
	13.	Generate models' accuracy and loss reports.
	14.	Generate ROC-AUC Curve.
	15.	Generate Grad Cam and Grad Cam++
	16.	Generate SHAP and Lime
Prediction	17.	Load best model.
	18.	Load random images.
	19.	Predict disease classes.

4.1. Accuracy

Accuracy represents the percentage of correctly predicted images among all predictions, as outlined in Eq. (9). This score evaluates the model's ability to make correct predictions (TP + TN) relative to the total predictions (TP + TN + FP + FN), where TP is true positive, TN is true negative, FP is false positive, and FN is false negative [49].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

4.2. Precision

Precision, [50] denoting the percentage of correctly identified positive outcomes, is calculated using Eq. (10).

$$\text{Precision (P)}: \quad P = \frac{TP}{TP+FP} : \tag{10}$$

4.3. Recall

The recall value assesses the accuracy of positive predictions by comparing true positive findings to the total number of actual positive samples (TP + FN), expressed by Eq. (11).

$$\text{Recall} \oplus : \quad R = \frac{TP}{TP+FN} : \tag{11}$$

4.4. F1-score

In evaluating a model's overall performance, researchers often turn to metrics like the F1-score. This score, defined in Eq. (12), captures the harmonic mean of a model's precision and recall.

$$\text{F1} - \text{score (F1)}: \quad F1 = 2 \times \frac{P \times R}{P+R} \tag{12}$$

4.5. Explainable AI (XAI)

Explainable AI (XAI) employs various techniques, including SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), Grad-CAM (Gradient-weighted Class Activation Mapping), and Grad-CAM++. SHAP assigns values to each feature, quantifying its impact on model predictions. LIME creates locally faithful models to explain predictions for specific instances. Grad-CAM visualizes which regions of an input image are influential for a deep neural network's decision, focusing on the gradient of the target class. Grad-CAM++ enhances Grad-CAM by incorporating positive and negative contributions to better highlight critical regions. These methods contribute to XAI by offering feature importance, local explanations, and visualizations. They help bridge the gap between complex AI models and human understanding, fostering transparency and trust in decision-making processes. Integrating these techniques ensures a comprehensive approach to explainability in AI across diverse applications.

4.5.1. SHAP

SHAP (SHapley Additive exPlanations) is a method in explainable AI that provides a consistent and mathematically sound way to fairly allocate contributions of each feature to the prediction made by a machine learning model. The Shapley value from cooperative game theory serves as the foundation for SHAP values.

$$\phi_i(f) = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N|-|S|-1)!}{|N|!}[f(S \cup i) - f(S)] \tag{13}$$

the Shapley value quantifies the average contribution of a feature to all possible combinations of features, ensuring a fair distribution of credit across the features. The SHAP values provide a unified framework for interpreting complex machine learning models and understanding the impact of individual features on predictions.

4.5.2. LIME

LIME (Local Interpretable Model-agnostic Explanations) is an explainability technique that aims to provide interpretable explanations for individual predictions made by machine learning models. LIME approximates the local behavior of a complex model around a specific instance by training a simpler, interpretable model. The basic idea behind LIME is to perturb the input data around a particular instance, observe the corresponding predictions from the black-box model, and then use this perturbed data to train a local interpretable model.

4.6. Grad-Cam & Grad-Cam++

Grad-CAM (Gradient-weighted Class Activation Mapping) and Grad-CAM++ are visualization techniques used to understand and interpret the decision-making processes of convolutional neural networks

Table 6
Model evaluation metrics performance.

Models	Recall	Precision	F1-score	Overall accuracy	Training accuracy	Validation Accuracy	Training Loss	Validation Loss
EfficientNetB0	0.95	0.94	0.93	0.94	0.96	0.94	0.10	0.15
InceptionV3	0.92	0.92	0.92	0.92	0.93	0.91	0.17	0.22
Lenet	0.89	0.87	0.93	0.91	0.92	0.91	0.18	0.24
Proposed Model	**0.98**	**0.99**	**0.99**	**0.99**	**1.00**	**0.99**	**0.000000249**	**0.015**

(CNNs). These methods highlight important regions of an input image that contribute the most to a particular class prediction.

$$L^c_{\text{CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \tag{14}$$

Grad-CAM++ improves upon Grad-CAM by incorporating positive and negative contributions. The Class Activation Map for Grad-CAM++ is given by:

$$L^c_{\text{CAM++}} = \text{ReLU}\left(\sum_k \frac{\alpha^c_k}{2}\left(2A^k - \sum_l A^l\right)\right) \tag{15}$$

Both Grad-CAM and Grad-CAM++ highlight regions in the input image that significantly influence the CNN's prediction for a specific class. These visualizations aid in understanding which parts of an image the model focuses on when making decisions.

4.7. Result analysis

This section delves into the performance assessment of various models, employing accuracy and loss graphs to illuminate their behavior. The ROC-AUC curves for all applied models are presented, along with demonstrations of interpretability tools such as SHAP, LIME, Grad-CAM, and Grad-CAM++. In evaluating the proposed model, a comparative analysis with other studies is conducted to ascertain its relevance and effectiveness. Learning curves, essential in the realm of Deep CNNs, depict time or epoch against learning progress. In this study, dual learning curves are generated for the Deep Learning (DL) model, encompassing both the training and validation datasets. These curves, specifically accuracy and loss graphs, serve as crucial tools for understanding model behavior and tuning parameters. Fig. 6 portrays accuracy and loss graphs for the proposed model, EfficientNetB0, InceptionV3, and Lenet with augmentation. The curves reveal nuanced fluctuations, notably subdued in models with data augmentation techniques. In Fig. 6(b), the minimal gap between training and validation accuracy signifies a well-fitted model. Despite some fluctuations in the loss graph, the actual losses remain notably lower. Consequently, the proposed model with data augmentation emerges as the optimal fit, as reflected in its accuracy graph, surpassing other models in overall performance

The confusion matrix for proposed model, EfficientNetB0, LeNet and InceptionV3 the classification reports offer a comprehensive assessment of the performance of the classification models that deals with four distinct classes: Corona Virus Disease, Normal, Pneumonia, and Tuberculosis. Starting with Fig. 7(a) the proposed model, the model showcases outstanding precision, recall, and F1-score across all classes, indicating a high degree of accuracy in predicting instances of Corona Virus Disease, Normal, Pneumonia, and Tuberculosis. The precision, recall, and F1-score values consistently outperform the other models, positioning proposed model as the top-performing model in this evaluation. Fig. 7(b) demonstrate EfficientNetB0 follows closely behind, demonstrating commendable precision, recall, and F1-score metrics. While slightly trailing proposed model, EfficientNetB0 exhibits robust performance across the evaluated classes, showcasing its effectiveness in accurately classifying instances of various respiratory conditions. The Fig. 7(c) which is InceptionV3, the third-ranked model, maintains good precision and recall values, though not reaching the levels observed in the proposed model and EfficientNetB0. Nevertheless, InceptionV3 proves to be a solid performer, effectively identifying cases of different respiratory conditions. Fig. 7(d) demonstrates LeNet, the fourth-ranked model, delivers respectable results but falls behind the top three models in terms of precision, recall, and F1-score.

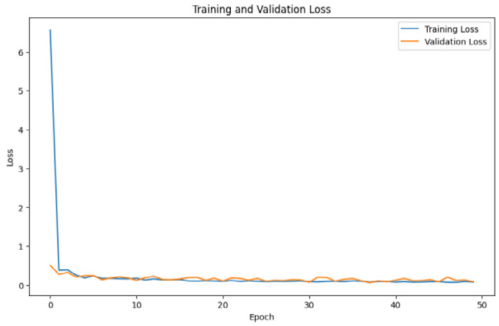
The evaluation of CNN-based Transfer Learning (TL) models and the proposed model involves key metrics such as accuracy, recall, precision, and F1-score. The training and validation accuracy and loss compared to the final (test) accuracy of the applied models are also presented in Table 6. Which illuminates that models integrated with augmentation techniques exhibit notably higher recall values, indicating a substantial enhancement in performance with the application of augmentation. EfficientNetB0 showcases an impressive recall of 94%, while InceptionV3 follows closely with a recall of 92%. Lenet achieves a recall of 89%. Notably, the proposed model consistently outperforms transfer-learning-based models, including EfficientNetB0, InceptionV3, and Lenet, irrespective of the use of augmentation techniques. Among all models, the proposed CNN model, particularly when coupled with data augmentation techniques, attains the highest recall value, reaching an impressive 98%. This underscores the superior performance of the proposed model in capturing relevant information, even amidst varying conditions and complexities in the dataset. the proposed model has achieved the highest accuracy of 99.20% with the smallest loss and smallest accuracy difference between the training and validation datasets.

The proposed model demonstrates superior performance with a recall of 0.98, precision of 0.99, and F1-score of 0.99, surpassing other models. Achieving an overall accuracy of 0.99, it exhibits consistent and reliable classification across all instances. The perfect training accuracy (1.00) and high validation accuracy (0.99) reflect its robust learning and generalization capabilities. Notably, the model maintains low training (0.000000249) and validation (0.015) loss, emphasizing its effectiveness in minimizing errors during both phases. Compared to alternative models, such as EfficientNetB0, InceptionV3, and Lenet, the proposed model consistently outperforms in key metrics, affirming its excellence in practical applications (see Table 7).

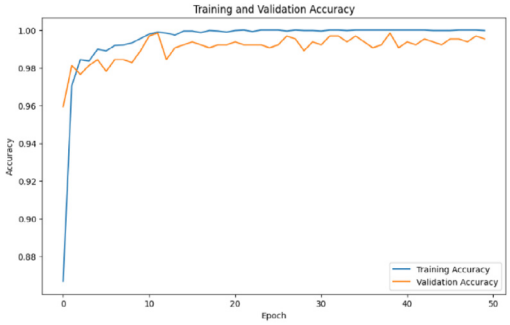
The proposed model consistently outperforms other models in both raw and preprocessed images, with and without augmentation, showcasing its versatility and effectiveness. Without augmentation, the model achieves competitive metrics, including a recall of 0.80 and precision of 0.77 in raw images, and a recall of 0.95 and precision of 0.97 in preprocessed images. However, with augmentation, the proposed model significantly improves, reaching a recall of 0.94 and precision of 0.96 in raw images, and an impressive recall of 0.98 and precision of 0.99 in preprocessed images. The augmentation strategy enhances its ability to generalize, resulting in higher recall, precision, and F1-score across different image types. The consistent excellence of the proposed model, whether with or without augmentation, establishes its reliability and suitability for both raw and preprocessed images in practical applications.

4.8. Ablation study

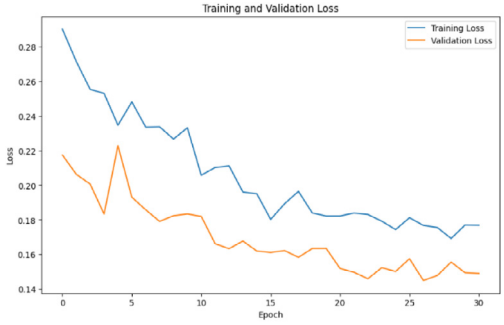
An ablation study is a type of scientific experiment or analysis commonly conducted in various fields, including medicine, physics, computer science, and engineering. The primary goal of an ablation study is to understand the importance or contribution of specific components or factors within a system, model, or process by systematically removing or "ablating" them and observing the resulting effects. This proposed



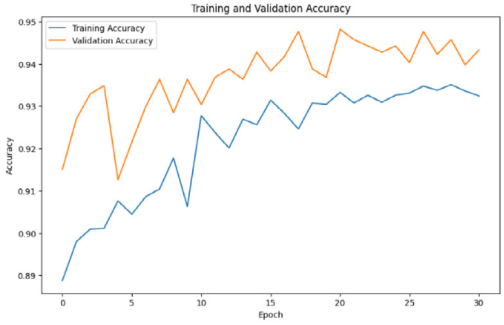
(a) Proposed Model Loss Curve



(b) Proposed Model Accuracy Curve



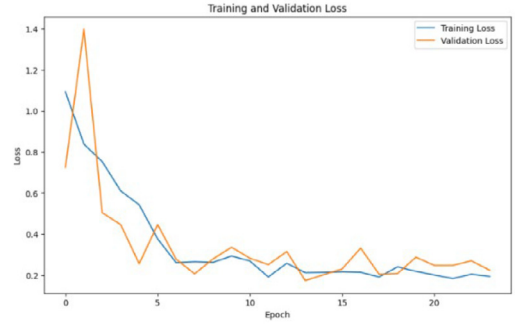
(c) EfficientNetB0 Loss Curve



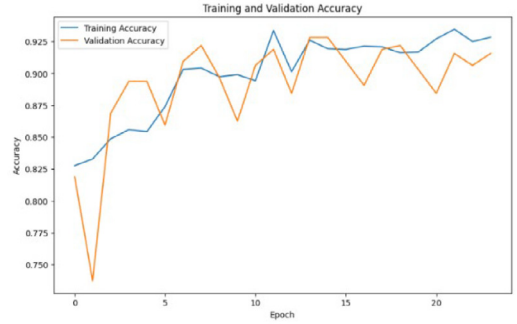
(d) EfficientNetB0 Accuracy Curve

(e) LeNet Loss Curve

(f) LeNet Accuracy Curve



(g) InceptionV3 Loss Curve



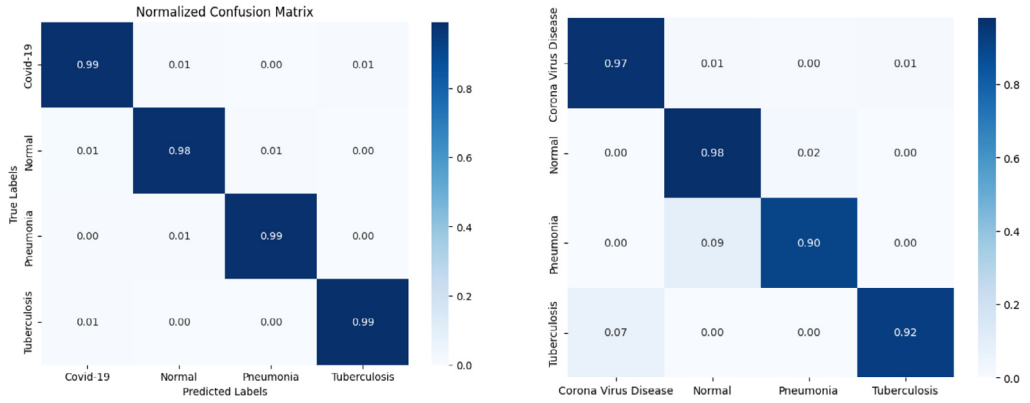
(h) InceptionV3 Accuracy Curve

Fig. 6. Loss and accuracy Graph of the proposed model, EfficientNetB0, LeNet, InceptionV3.

model is a deep convolutional neural network architecture that has demonstrated outstanding performance in a variety of computer vision applications, including image classification. It was pretrained using a sizable dataset, often ImageNet.

4.8.1. Changing flatten layer

The base model is the proposed model that serves as the backbone of the neural network. Base model output refers to the output tensor of this pretrained model. This part creates a Flatten layer. The Flatten layer is responsible for taking the multi-dimensional output tensor from the previous and reshaping it into a one-dimensional tensor showed in Table 8. This is often necessary when transitioning from convolutional layers to fully connected layers in a neural network. The output tensor of the proposed model serves as the input to the Flatten layer. This means that the feature maps or activations produced by the convolutional layers in proposed model are flattened into a one-dimensional



(a) Proposed Model Confusion Matrix

(b) EfficientNetB0 Confusion Matrix

(c) LeNet Confusion Matrix

(d) InceptionV3 Confusion Matrix

Fig. 7. Confusion Matrix of Proposed Model, EfficientNetB0, LeNet and InceptionV3.

Table 7
Results evaluation with and without Augmentation.

	Models	Augmentation	Recall	Precision	F1-score	Accuracy
(a) Raw Image	Proposed Model	Without Augmentation	0.80	0.77	0.77	0.78
		With Augmentation	0.94	0.96	0.94	0.94
	EfficientNetB0	Without Augmentation	0.76	0.81	0.73	0.76
		With Augmentation	0.87	0.90	0.84	0.87
(b) Preprocessed Images	Proposed Model	Without Augmentation	0.95	0.97	0.98	0.96
		With Augmentation	0.98	0.99	0.99	0.99
	EfficientNetB0	Without Augmentation	0.92	0.91	0.89	0.90
		With Augmentation	0.95	0.94	0.93	0.94
	InceptionV3	Without Augmentation	0.87	0.89	0.83	0.86
		With Augmentation	0.92	0.92	0.92	0.92
	LeNet	Without Augmentation	0.79	0.83	0.81	0.81
		With Augmentation	0.89	0.87	0.93	0.91

Table 8
Changing flatten layer.

Layer name	Val_Acc (%)	Val_Loss (%)	T_Acc (%)	T_Loss (%)	Finding
AveragePooling2D	99.41	0.015	99.20	0.0000000249	Identical Performance
GlobalMaxPooling2D	95.89	0.29	96.12	0.21	Accuracy Dropped
GlobalAveragePooling2D	92.31	0.31	92.49	0.29	Accuracy Dropped
MaxPooling2D	93.84	0.28	94.19	0.27	Accuracy Dropped

vector [51] here, Val stands for Validation and *T* stands for Testing. In Table 8 we can see that the AveragePooling2D layer is the best flatten layer among others which gave the highest performance and less loss.

4.8.2. Changing batch size

The batch size is a critical hyperparameter in training deep neural networks. It impacts both the training dynamics and memory usage during training. A larger batch size may lead to faster convergence during training since it computes gradients based on more samples at once. However, it can require more memory, and there might be a risk of overfitting if the batch size is excessively large. A smaller batch size reduces memory requirements but may lead to slower convergence since it computes gradients based on fewer samples. Smaller batch sizes are sometimes preferred when working with limited computational

Table 9
Changing batch size table.

Batch size	Overall loss	Overall accuracy
32	0.40	0.94
64	0.023	0.99
128	0.31	0.94

Table 10
Changing optimizer and learning rates table.

Optimizer	Learning rates	Best optimizer	Best learning rates	Best accuracy
adam	0.001 0.0001 0.00001	adam	0.00001	99.20%
nadam	0.000001 0.0000001			

resources. Increasing the batch size can potentially speed up training, but it may also require more GPU memory. It is essential to balance batch size with available resources. The choice of batch size can affect the model's generalization and convergence behavior. Smaller batch sizes may introduce more randomness into weight updates, which can sometimes help the model escape local minima. The Table 9. shows the tried experiments with different batch size like 32,64 and 128 and the best accuracy was obtained by AveragePooling2D which is 99.20%.

4.8.3. Changing optimizer and learning rates

In the original code, he Adam optimizer have been with default learning rate, which is typically set to a small value like 0.001. Then experimentations have been made with other values like 0.0001 and 0.0000001 as well. Which did not help much but did improve some training loss. The training loss improved from 0.31 to 0.000000429 which is shown in Table 10.

4.8.4. Importance of optimizer and learning rate

The choice of optimizer and learning rate is crucial in training deep neural networks. Different optimizers and learning rates can lead to variations in training speed, convergence, and model performance. The Adam optimizer is a popular choice due to its adaptive learning rate mechanism, which adjusts the learning rate during training. Sometimes, learning rate schedules are used, where the learning rate is reduced over time (e.g., through learning rate annealing or learning rate decay). This can help fine-tune the training process. Experimenting with different optimizers (e.g., Adam, SGD, RMSprop) can also affect training outcomes. Some optimizers perform better than others for specific tasks or architectures. Here it can be seen that it improved the training loss. The validation loss improved from 0.24 to 0.015.

4.8.5. Final configuration after ablation studies

The "final configuration after ablation studies" refers to the last set of hyperparameters, model architecture, and training settings that were determined to be the best based on the results of ablation studies. This research was conducted by several experiments by changing various aspects of the model, such as the flatten layer, batch size, optimizer, and learning rates. After conducting ablation studies and fine-tuning the model, here Table 11. the final configuration and performance summary of the proposed model is pretrained on ImageNet. Only the fully connected layers are trainable, while the convolutional layers are frozen. Applied data augmentation techniques during training, including rotation, width shift, height shift, shear, zoom, and horizontal flip. Batch size set to 64 for both training and validation data generators. Loss Function Categorical Cross-Entropy. Optimizer is Adam optimizer. Performance matrix on training data loss 0.000000429 accuracy: 100%. Performance matrix on test data Loss: 0.0185 & accuracy 99.20%. Performance matrix on validation data loss 0.015 & accuracy 99.41%.

Table 11
Final configuration table.

Configuration	Value
Image sizes	224 × 224
Optimization function	Adam
Layer Name	Flatten
Learning Rate	0.000001
Batch Size	64
Loss Function	Categorical Cross-Entropy
Pooling layer	AveragePooling2D

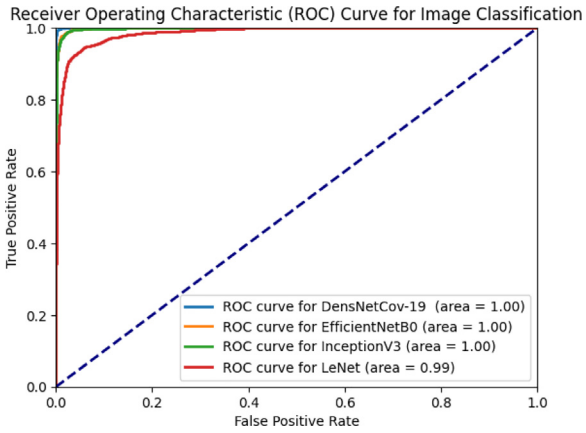


Fig. 8. ROC-AUC Curve of applied Algorithms.

4.8.6. ROC-AUC curve analysis

The ROC-AUC curves [52] of the utilized models are shown in Fig. 8. The LeNet model demonstrated the lowest AUC score among all the models, indicating relatively poorer performance. However, upon applying augmentation techniques, the LeNet model exhibited an improvement in performance, although it still lagged behind the other models. In contrast, the EfficientNetB0 and InceptionV3 models showcased superior performance compared to the LeNet model, yet fell short of the performance achieved by the proposed model. Notably, the proposed model achieved the highest AUC score, outperforming all other models in the study. Moreover, the ROC curves of the proposed model with augmentation displayed a more optimal performance, surpassing the other models, including the EfficientNetB0 and InceptionV3 models. These findings emphasize the substantial impact of augmentation techniques on the overall performance of the models, with the proposed model with augmentation emerging as the most effective model among those tested".

Numerous strides in deep learning research focus on refining the practicality and interpretability of models, particularly in medical imaging. The Gradient Weighted [53] Class Activation Mapping (Grad-CAM) method emerges as a powerful tool for visualizing densely connected neural networks, providing crucial insights into decision-making processes during classification tasks. Applied specifically to X-ray images, this approach utilizes a proposed model as an adept detection mechanism. Post-label prediction, Grad-CAM unfolds at the final convolutional layer, generating heatmaps that pinpoint critical regions within X-ray images. The accompanying code seamlessly aligns with this methodology, employing Grad-CAM to visualize activation maps in the context of X-ray classification. Illustrated in Fig. 9 of relevant research, the application of Grad-CAM vividly showcases heatmaps on X-ray images, contributing significantly to the transparency and interpretability of deep neural models, particularly in crucial medical imaging scenarios [54]. This research also leverages Grad-CAM++, an enhanced version of the visualization technique. Grad-CAM++ incorporates additional information from both the forward and backward passes of the neural network, leading to more accurate and fine-grained localization of important features.

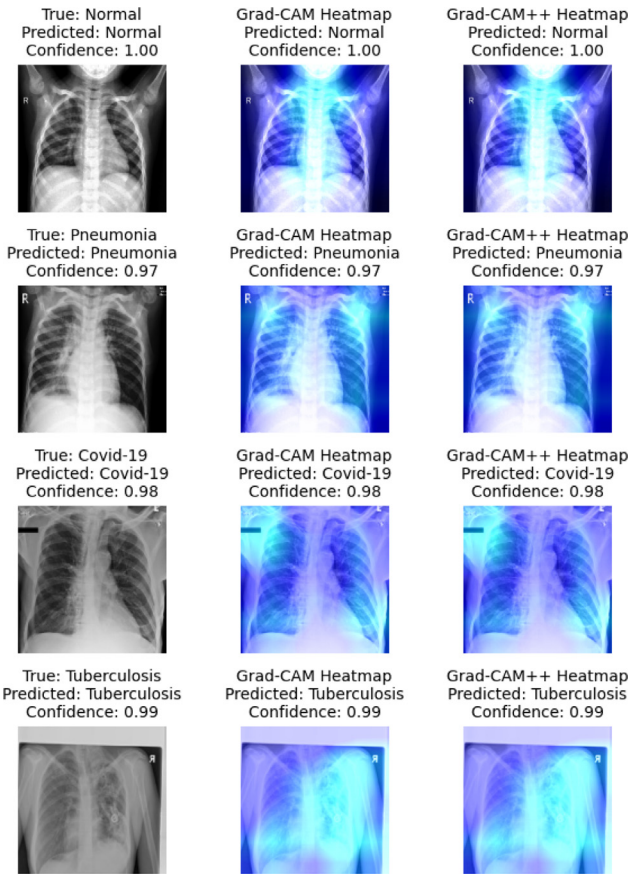


Fig. 9. CXR images utilizing Grad-CAM on the Proposed model.

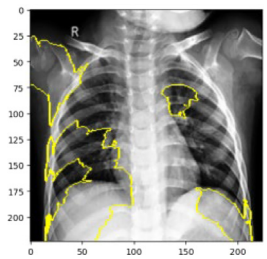


Fig. 10. Normal CXR Image Using LIME.

In the examination of our deep learning model's predictions concerning chest X-ray images depicting lung diseases, the informative methodologies of LIME and SHAP were employed [55]. LIME played the role of the local guide, bringing attention to specific regions deemed critical for disease classification within both the validation and test datasets. Features indicative of COVID-19, tuberculosis, normal lung structures, and pneumonia were highlighted. While LIME demonstrated competence in accentuating these features, occasional inaccuracies, particularly in outlining outer lung regions, were observed. Nevertheless, the essential regions for classification were identified in the majority of images. A visual representation of the map generated for two chest X-ray images depicting lung diseases using LIME can be observed in Figs. 10 and 11.

Conversely, SHAP, our more globally attuned companion, not only provided explanations but also assigned importance scores to individual pixels—akin to granting each pixel a voice in the decision-making orchestration. Impressively, SHAP proficiently identified crucial regions, utilizing pink hues to signify significance in images depicting COVID-19, tuberculosis, normal lung conditions, and pneumonia. The

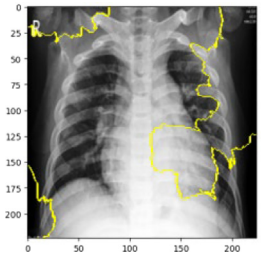


Fig. 11. Pneumonia CXR Image Using LIME.

visualizations of SHAP values served as a narrative, unveiling the nuanced story behind the model's predictions. Vibrant saliency maps of two chest X-ray images representing different lung diseases using SHAP are illustrated in Fig. 12

The SHAP [56] Gradient Explainer explains machine learning predictions by calculating Shapley values through gradients, offering model-agnostic insights with a local focus on specific instances. Conversely, the [57] SHAP Kernel Explainer approximates Shapley values through a weighted average of model predictions from feature subsets, providing both local and global interpretability. This is especially useful for black-box models where obtaining gradients is challenging. Fig. 12 visually explains how the model makes predictions using SHAP values. Notably, the red hues in the saliency maps signify positive contributions, while blue hues indicate negative contributions. In the presented pneumonia images, the prominence of red colors suggests accurate detection, showcasing the effectiveness of the SHAP Gradient Explainer in highlighting relevant features for precise classification. in Fig. 12, two pneumonia images were randomly selected, and the SHAP Gradient Explainer accurately detected them. This is evident in the red parts indicating positive signs, and the pneumonia images have more red colors compared to the COVID-19 images. This interpretability tool not only enhances the understanding of model decisions but also provides valuable insights for clinicians, contributing to the transparency and trustworthiness of the deep learning model in the context of lung disease diagnosis. While the Gradient Explainer prioritizes efficiency with gradients, the Kernel Explainer offers flexibility in scenarios with demanding gradient computations, enhancing the overall interpretability of machine learning models (see Fig. 13).

Upon scrutiny of these figures, it becomes apparent that SHAP tends to attribute importance to the presence of specific disease-related features, such as anomalies characteristic of COVID-19 or tuberculosis, distinctly highlighting these areas in pink. Conversely, images representing normal lung conditions may lack noteworthy features influencing the model, resulting in a serene blue wash across the entire image in SHAP values.

Conducting a comprehensive comparative analysis, Table 12 scrutinizes twelve existing works, highlighting the performance of their architectures, the number of classes considered, and the specific deep learning methodologies employed. This analysis serves as a critical lens through which the effectiveness of the proposed model can be further understood and evaluated.

4.9. Deployment

The deployment of the deep learning model involves introducing a finalized model into a production environment, where it can serve its intended purpose. In the context of lungs disease prediction using chest X-rays, the ultimate aim is to benefit society and cater to the needs of a broad audience. Given the prevalence of lung diseases and the complexity of their diagnosis, providing a tool for early detection is crucial for effective healthcare. The primary objective and motivation of this study revolve around deploying an End-to-End (E2E) system that can assist in predicting lung diseases from chest X-ray images. In this

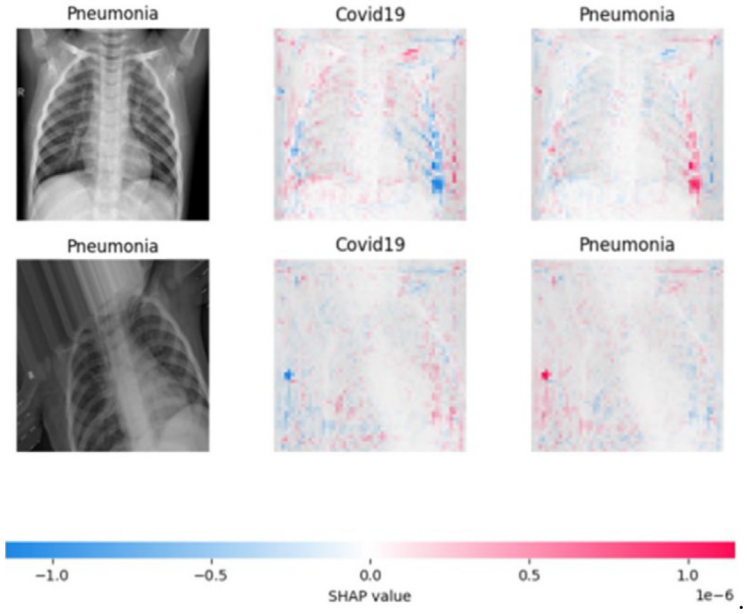


Fig. 12. Random CXR images using SHAP Gradient Explainer.

Table 12
The performance comparison between the proposed model and existing related work.

Ref	Number of classes	Method	Medical image	Acc.	Prec.	XAI Used
Hamal et al. [14]	3	4S-min-FP	CXR, CT	0.98	0.90	No
Bondugula et al. [15]	3	4S-min-FP	CXR, CT	0.91	0.93	No
Khanna et al. [16]	3	Xgboost	Numerical Data	0.87	0.81	SHAP
Tian et al. [17]	2	DenseNet-161	MRI T2-weighted images, Numeric Data	92.44	0.95	No
Nalluri et al. [18]	8	CBRACDC (Customized Hybrid Model)	CXR, Text Data	90	–	No
Chutia et al. [19]	2	DenseNet201	CXR, CT	95.34	97	Grad-CAM
Malik et al. [20]	9	Proposed CNN	CXR, CT, CSI	99.01	93	Grad-CAM
Mishra et al. [21]	3	CNN with pretrained weights on ImageNet	CXR	91	92	No
Bennour et al. [22]	4	MDCXR3-Net	CXR	97.74	97.86	No
Bennour et al. [22]	4	MDCXR4-Net	CXR	90.37	90.67	No
Niño et al. [23]	2	ResNet50	CXR	0.91	–	No
Proposed	4	**DensNetCov-19 (Customized CNN)**	**CXR**	**99.20**	**99**	**Visualization techniques used, Grad-Cam, Grad-Cam++, SHAP and LIME**

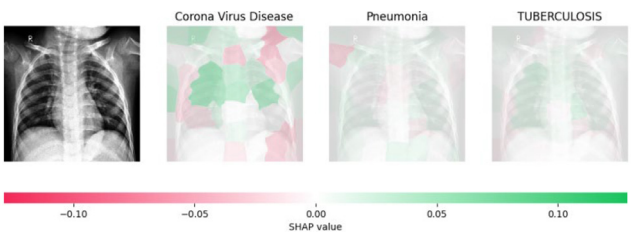
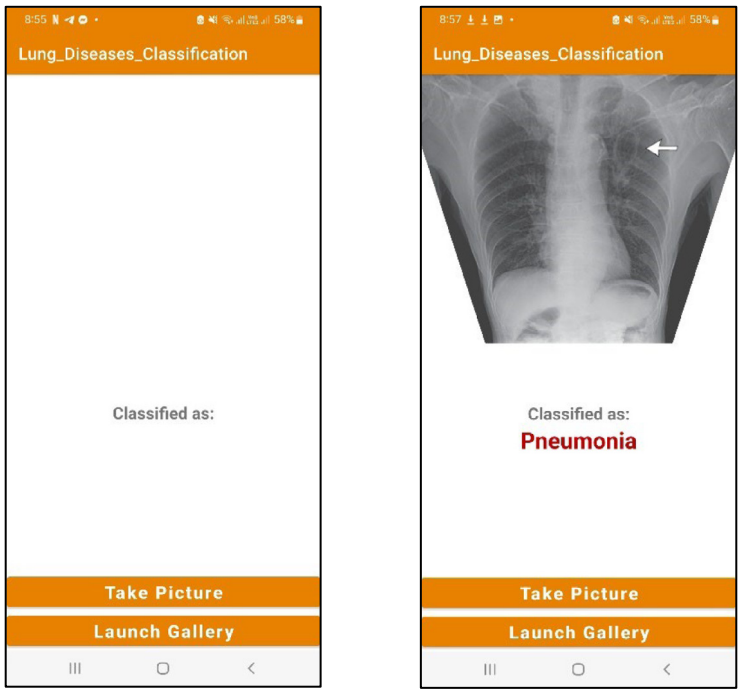


Fig. 13. CXR images using SHAP Kernel Explainer.

context, the study focused on deploying the model through an Android application. Fig. 14 showcase the user interface of the Android application along with real-time classification results. The implementation of the most effective proposed model enables the prediction of lung diseases based on chest X-ray images. In developing the system, the study employed various Python libraries and supporting technologies. The Android application's frontend design and functionality were tailored to facilitate the user's upload of a chest X-ray image, which is then classified by the proposed model, with the results displayed on the screen. It is important to note that, in this study, the deployment was specifically focused on the Android application.

The Android application's user interface is crafted using Extensible Markup Language (XML), while the backend is powered by Java. For image classification within the app, the optimal proposed model has



(a) User Interface (b) Classification result

Fig. 14. User Interface and Classification results on Android Application.

been converted to a lightweight TensorFlow version in the "tflite" format. Java is employed to seamlessly integrate this lightweight TensorFlow model into the Android application. The Android application is designed to be versatile, supporting Android versions 5.0 to 12. Users can leverage the application to classify lung diseases by either capturing a live image or selecting an existing image from their device. Our open-access Android application is available for download on our GitHub repository [58]. The lightweight nature of the proposed model ensures swift and accurate predictions, providing users with rapid and reliable results. This smartphone-based Android application offers a user-friendly experience, making lung disease classification accessible and efficient.

5. Conclusion and discussion

This research establishes a robust foundation for lung disease prediction using deep learning. We conducted rigorous experiments with various architectures, including our proposed model, InceptionV3, EfficientNetB0, and LeNet, to classify lung diseases from chest X-ray images. Our proposed model demonstrated superior performance, achieving a remarkable accuracy of 99.20% and high precision, recall, and F1-scores across all classes, which is crucial for accurate medical diagnoses. The meticulous training setup, including the Adam optimizer, categorical cross-entropy loss function, early stopping, and essential callbacks, contributed significantly to these results [59]. The broader implications of our findings suggest that integrating deep learning models into medical diagnostics can greatly enhance the accuracy and efficiency of lung disease detection, leading to better patient outcomes. The ability of the proposed model to provide high accuracy with minimal difference between training and validation datasets indicates its strong generalization capabilities. Moreover, the integration of explainable AI tools like SHAP, LIME, Grad-CAM, and Grad-CAM++ enhances the trustworthiness of the model by providing valuable insights into its decision-making processes. This transparency is essential for clinical adoption, as it helps healthcare professionals understand and trust the AI's recommendations. The high ROC-AUC scores further underscore the model's efficacy in distinguishing between different respiratory conditions, making it a reliable diagnostic tool in clinical settings.

5.1. Limitations and future work

Despite the promising results, this study has limitations. The dataset used, while comprehensive, may not fully represent the diversity of the global population. Thus, the model's performance needs validation across diverse populations to ensure generalizability. Future research should focus on integrating multi-modal data, such as combining X-ray images with other clinical data (e.g., patient history, lab results), to provide a more comprehensive analysis of patient health [60]. Developing real-time prediction systems for chest X-ray images can facilitate timely interventions, crucial in emergency and routine healthcare settings. Exploring transfer learning techniques and model generalization efforts will enhance adaptability across different populations and settings. Additionally, incorporating more advanced explainable AI methods will improve the model's interpretability, making it more user-friendly for clinicians. Continuous learning and updates based on emerging data and research will ensure the model remains relevant in the dynamic healthcare landscape. Collaboration with healthcare professionals is crucial for the seamless integration of these AI tools into routine clinical practice, ensuring that the technology complements and enhances human expertise rather than replacing it. By addressing these areas, we can contribute to the development of more accurate, accessible, and reliable diagnostic tools, ultimately improving patient care and outcomes in the field of respiratory health.

Declaration of competing interest

The authors declare that they have no conflicts of interest to disclose. This research was conducted impartially, and there are no financial, personal, or professional relationships that could be perceived as influencing the content of this manuscript.

Data availability

Data will be made available on request.

Acknowledgment

This research is supported by the Multimedia University (MMU) IR Fund (Grant no. MMUI/220041).

References

[1] Global burden of chronic respiratory diseases and risk factors, 1990–2019: an update from the Global Burden of Disease Study 2019 | The Institute for Health Metrics and Evaluation, www.healthdata.org. https://www.healthdata.org/research-analysis/library/global-burden-chronic-respiratory-diseases-and-risk-factors-1990-2019.

[2] R. Kumar Kushagra, S. Jain, Recent Advances in Machine Learning for Diagnosis of Lung Disease: A Broad View, IEEE Xplore, 2022, https://ieeexplore.ieee.org/abstract/document/9753275. (6 September 2023).

[3] M.M. Shukla, B.K. Tripathi, M. Nagle, B.K. Chaurasia, COVID-19 & Lung Disease Detection using Deep Learning, IEEE Xplore, 2022, https://ieeexplore.ieee.org/abstract/document/10008269. (6 September 2023).

[4] S. Bharati, P. Podder, M.R.H. Mondal, Hybrid deep learning for detecting lung diseases from X-ray images, Inform. Med. Unlocked 20 (2020) 100391, http://dx.doi.org/10.1016/j.imu.2020.100391.

[5] P. Afshar, et al., 3D-MCN: A 3D multi-scale capsule network for lung nodule malignancy prediction, Sci. Rep. 10 (1) (2020) http://dx.doi.org/10.1038/s41598-020-64824-5.

[6] L. Wang, H. Shen, K. Enfield, K. Rheuban, COVID-19 Infection Detection Using Machine Learning, IEEE Xplore, 2021, https://ieeexplore.ieee.org/document/9671700. (Accessed 29 July 2023).

[7] M. Hasan, M. Al Mamun, M.C. Das, M.M. Hasan, A.M. Islam, The Application and Comparison of Deep Learning Models for the Prediction of Chest Cancer Prognosis, IEEE Xplore, 2023, https://ieeexplore.ieee.org/abstract/document/10216201.

[8] P. Kalaivani, A. Dhivya, G. Dharani, S. Bharathi, C. Rajan, Deep Learning with Multi-Class Classification for Detection of Covid-19 and Pneumonia, IEEE Xplore, 2023, https://ieeexplore.ieee.org/abstract/document/10134162. (Accessed 7 September 2023).

[9] T.Q.H. Vo, C.-M. Lin, An improved densenet deep neural network model for tuberculosis detection using chest X-ray images, 11, 2023, pp. 42839–42849, http://dx.doi.org/10.1109/access.2023.3270774.

[10] T. Soewu, S.V. Uday Kalyan, M. Rakhra, D. Singh, Lung cancer detection using image processing, in: 2022 5th International Conference on Contemporary Computing and Informatics, (IC3I), Uttar Pradesh, India, 2022, pp. 1206–1211, http://dx.doi.org/10.1109/IC3I56241.2022.10072589.

[11] J.M.S. Lavanya, P. Subbulakshmi, Machine learning techniques for the prediction of non-communicable diseases, in: 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering, (ICECONF), Chennai, India, 2023, pp. 1–8, http://dx.doi.org/10.1109/ICECONF57129.2023.10084301.

[12] F. Aslam, Advancing intelligence: Unveiling the power of advanced machine learning algorithms for real-world applications, J. Eng. Res. Rep. 25 (7) (2023) 159–165, http://dx.doi.org/10.9734/jerr/2023/v25i7949.

[13] N. Kosarkar, P. Basuri, P. Karamore, P. Gawali, P. Badole, P. Jumle, Disease prediction using machine learning, in: 2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing, (ICETET-SIP-22), Nagpur, India, 2022, pp. 1–4, http://dx.doi.org/10.1109/ICETET-SIP-2254415.2022.9791739.

[14] Susmita Hamal, Bhupesh Kumar Mishra, R. Baldock, W. Sayers, Tek Narayan Adhikari, R.M. Gibson, A comparative analysis of machine learning algorithms for detecting COVID-19 using lung X-ray images, Decis. Anal. J. 11 (2024) http://dx.doi.org/10.1016/j.dajour.2024.100460, 100460–100460.

[15] Rohit Kumar Bondugula, Nitin Sai Bommi, S.K. Udgata, An efficient multi-stage ensemble deep learning framework for diagnosing infectious diseases, Decis. Anal. J. 11 (2024) http://dx.doi.org/10.1016/j.dajour.2024.100458, 100458–100458.

[16] V.V. Khanna, K. Chadaga, N. Sampathila, S. Prabhu, R.C. P., A machine learning and explainable artificial intelligence triage-prediction system for COVID-19, Decis. Anal. J. 7 (2023) 100246, http://dx.doi.org/10.1016/j.dajour.2023.100246.

[17] H. Tian, et al., DenseNet model incorporating hybrid attention mechanisms and clinical features for pancreatic cystic tumor classification, J. Appl. Clin. Med. Phys. (2024) e14380, http://dx.doi.org/10.1002/acm2.14380.

[18] S. Nalluri, R. Sasikala, Detection and difference of pneumonia from other chest/lung disease using multi-model data: A hybrid classification model, Int. J. Intell. Syst. Appl. Eng. 12 (1) (2024) 328–344, [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/3420. (Accessed 27 May 2024).

[19] Upasana Chutia, Anand Shanker Tewari, Jyoti Prakash Singh, Vikash Kumar Raj, Classification of lung diseases using an attention-based modified DenseNet model, J. Imaging Inform. Med. (2024) http://dx.doi.org/10.1007/s10278-024-01005-0.

[20] H. Malik, Tayyaba Anees, Multi-modal deep learning methods for classification of chest diseases using different medical imaging and cough sounds, PLoS One 19 (3) (2024) http://dx.doi.org/10.1371/journal.pone.0296352, e0296352–e0296352.

[21] O. Mishra, et al., Imbalanced Class Problem Analysis for Lung Cancer Detection Using Convolutional Neural Networks, IOP Publishing eBooks, 2024, pp. 6–14, http://dx.doi.org/10.1088/978-0-7503-5417-2ch6.

[22] Akram Bennour, Najib Ben Aoun, Osamah Ibrahim Khalaf, Fahad Ghabban, W.-K. Wong, Sameer Algburi, Contribution to pulmonary diseases diagnostic from X-ray images using innovative deep learning models, Heliyon 10 (9) (2024) http://dx.doi.org/10.1016/j.heliyon.2024.e30308, e30308–e30308.

[23] L. Elena, Jhojan Genaro Nuñez-Fernandez, Fany Yesica Taquila-Calderon, I. Adrianzén-Olano, P. De-La-Cruz-VdV, G. Carrión-Barco, Classification model using transfer learning for the detection of pneumonia in chest x-ray images, Int. J. Online Biomed. Eng. 20 (05) (2024) 150–161, http://dx.doi.org/10.3991/ijoe.v20i05.45277.

[24] Lavina Jean Crasta, Rupal Neema, Alwyn Roshan Pais, A novel deep learning architecture for lung cancer detection and diagnosis from computed tomography image analysis, Healthc. Anal. (2024) http://dx.doi.org/10.1016/j.health.2024.100316, 100316–100316.

[25] S. Wankhade, V. S., A novel hybrid deep learning method for early detection of lung cancer using neural networks, Healthc. Anal. (2023) 100195, http://dx.doi.org/10.1016/j.health.2023.100195.

[26] S. prasad Koyyada, T.P. Singh, An explainable artificial intelligence model for identifying local indicators and detecting lung disease from chest X-ray images, Healthc. Anal. 4 (2023) 100206, http://dx.doi.org/10.1016/j.health.2023.100206.

[27] Negar Maleki, Seyed Taghi Akhavan Niaki, An intelligent algorithm for lung cancer diagnosis using extracted features from computerized tomography images, Healthc. Anal. 3 (2023) 100150, http://dx.doi.org/10.1016/j.health.2023.100150.

[28] R. Mumtaz, M.H. Maqsood, I. ul Haq, U. Shafi, Z. Mahmood, M. Mumtaz, Integrated digital image processing techniques and deep learning approaches for wheat stripe rust disease detection and grading, Decis. Anal. J. 8 (2023) 100305, http://dx.doi.org/10.1016/j.dajour.2023.100305.

[29] O.M. Dalvi, Lungs disease dataset (4 types), www.kaggle.com, https://www.kaggle.com/datasets/omkarmanohardalvi/lungs-disease-dataset-4-types.

[30] A. Mikołajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: 2018 International Interdisciplinary PhD Workshop, IIPhDW), Świnoujście, Poland, 2018, pp. 117–122, http://dx.doi.org/10.1109/IIPHDW.2018.8388338.

[31] L. Perez, J. Wang, The effectiveness of data augmentation in image classification using deep learning, 2017, arXiv.org, https://arxiv.org/abs/1712.04621v1.

[32] Saleh Albahli, Efficient GAN-based chest radiographs (CXR) augmentation to diagnose coronavirus disease pneumonia, Int. J. Med. Sci. 17 (10) (2020) 1439–1448, http://dx.doi.org/10.7150/ijms.46684, www.ncbi.nlm.nih.gov/pmc/articles/PMC7330663/.

[33] A. Mikołajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: 2018 International Interdisciplinary PhD Workshop, (IIPhDW), Świnoujście, Poland, 2018, pp. 117–122, http://dx.doi.org/10.1109/IIPHDW.2018.8388338.

[34] Image Denoising Using Autoencoders in Deep Learning, Omdena, 2024, https://omdena.com/blog/denoising-autoencoders. (Accessed 19 June 2024).

[35] Denoising autoencoders, 2022, Available online: https://pyimagesearch.com/2020/02/24/denoising-autoencoders-with-keras-tensorflow-and-deep-learning/ (Accessed on 16 2022).

[36] T. Mahmood, J. Li, Y. Pei, F. Akhtar, An automated in-depth feature learning algorithm for breast abnormality prognosis and robust characterization from mammography images using deep transfer learning, Biology 10 (9) (2021) 859, http://dx.doi.org/10.3390/biology10090859.

[37] P. Wang, J. Wang, Y. Li, P. Li, L. Li, M. Jiang, Automatic classification of breast cancer histopathological images based on deep feature fusion and enhanced routing, Biomed. Signal Process. Control 65 (2021) 102341, http://dx.doi.org/10.1016/j.bspc.2020.102341.

[38] A.M. Reza, Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement, J. VLSI Signal Process.-Syst. Signal, Image, Video Technol. 38 (1) (2004) 35–44, http://dx.doi.org/10.1023/b:vlsi.0000028532.53893.82.

[39] K. Koonsanit, S. Thongvigitmanee, N. Pongnapang, P. Thajchayapong, Image enhancement on digital x-ray images using N-CLAHE, in: 2017 10th Biomedical Engineering International Conference, (BMEiCON), Hokkaido, Japan, 2017, pp. 1–4, http://dx.doi.org/10.1109/BMEiCON.2017.8229130.

[40] Ganesh R. Vidhya, H. Ramesh, Effectiveness of contrast limited adaptive histogram equalization technique on multispectral satellite imagery, in: Proceedings of the International Conference on Video and Image Processing, (ICVIP '17), Association for Computing Machinery, New York, NY, USA, 2017, pp. 234–239, http://dx.doi.org/10.1145/3177404.3177409.

[41] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X.-P. Zhang, X. Ding, A retinex-based enhancing approach for single underwater image, in: 2014 IEEE International Conference on Image Processing, (ICIP), Paris, France, 2014, pp. 4572–4576, http://dx.doi.org/10.1109/ICIP.2014.7025927.

[42] M. Bocher, et al., Gamma camera-mounted anatomical X-ray tomography: technology, system characteristics and first images, Eur. J. Nucl. Med. Mol. Imaging 27 (6) (2000) 619–627, http://dx.doi.org/10.1007/s002590050555.
[43] D. de Martino, et al., X-ray follow-ups of XSS J12270-4859: A low-mass X-ray binary with gamma-ray Fermi-LAT association, Astron. Astrophys. 550 (2013) A89, http://dx.doi.org/10.1051/0004-6361/201220393.
[44] V. Perumal, V. Narayanan, S.J.S. Rajasekar, Detection of COVID-19 using CXR and CT images using transfer learning and haralick features, Appl. Intell. 51 (2020) http://dx.doi.org/10.1007/s10489-020-01831-z.
[45] R.A. Charisma, F. Dharma Adhinata, Transfer learning with densenet201 architecture model for potato leaf disease classification, in: 2023 International Conference on Computer Science, Information Technology and Engineering, (ICCoSITE), Jakarta, Indonesia, 2023, pp. 738–743, http://dx.doi.org/10.1109/ICCoSITE57641.2023.10127772.
[46] Y.-H. Chiu, et al., Advancements in ct lung cancer diagnosis: the role of deep learning for effective classification, J. Med. Imaging Radiat. Sci. 54 (3 Supplement) (2023) S24, http://dx.doi.org/10.1016/j.jmir.2023.06.086.
[47] A. Ur Rehman, A. Naseer, S. Karima, M. Tamoor, S. Naz, Deep learning classifiers for computer-aided diagnosis of multiple lungs disease, J. X-ray Sci. Technol. (2023) http://dx.doi.org/10.3233/XST-230113.
[48] H. Xu, Comparison of CNN Models in Non-small Lung Cancer Diagnosis, IEEE Xplore, 2023, https://ieeexplore.ieee.org/document/10075772 (Accessed 28 April 2023).
[49] Confusion matrix, 2022, Wikipedia, https://en.wikipedia.org/w/index.php?title=Confusion_matrix&oldid=1107701525. (Accessed 15 December 2023).
[50] M.A. Khan, F. Algarni, A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS, IEEE Access 8 (2020) 122259–122269, http://dx.doi.org/10.1109/access.2020.3006424.
[51] F.M.J.M. Shamrat, et al., LungNet22: A fine-tuned model for multiclass classification and prediction of lung disease using X-ray images, J. Personalized Med. 12 (5) (2022) 680, http://dx.doi.org/10.3390/jpm12050680.
[52] M.H. Ferris, et al., Using ROC curves and AUC to evaluate performance of no-reference image fusion metrics, in: 2015 National Aerospace and Electronics Conference, (NAECON), Dayton, OH, USA, 2015, pp. 27–34, http://dx.doi.org/10.1109/NAECON.2015.7443034.
[53] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision, (ICCV), 2017, http://dx.doi.org/10.1109/iccv.2017.74.
[54] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision, (WACV), Lake Tahoe, NV, USA, 2018, pp. 839–847, http://dx.doi.org/10.1109/WACV.2018.00097.
[55] X. Guo, Y. Li, H. Ling, LIME: Low-light image enhancement via illumination map estimation, IEEE Trans. Image Process. 26 (2) (2017) 982–993, http://dx.doi.org/10.1109/tip.2016.2639450.
[56] Shap. gradientexplainer — shap latest documentation, 2023, shap-lrjball.readthedocs.io. https://shap-lrjball.readthedocs.io/en/latest/generated/shap GradientExplainer.html (Accessed 15 December 2023).
[57] shap. KernelExplainer — SHAP latest documentation, shap-lrjball.readthedocs.io. https://shap-lrjball.readthedocs.io/en/latest/generated/shap KernelExplainer.html.
[58] DrDisease69, Mobile-app, 2023, GitHub, 2023. https://github.com/DrDisease69/Mobile-App (Accessed December 15 2023).
[59] M. Azam, et al., A systematic review of hyperparameter optimization techniques in convolutional neural networks, Decis. Anal. J. 11 (2024) http://dx.doi.org/10.1016/j.dajour.2024.100470, 100470–100470.
[60] Deepika Koundal, Y. Guo, S. Chakraborty, R.F. Mansour, Emerging trends in healthcare decision support systems, Decis. Anal. J. 11 (2024) http://dx.doi.org/10.1016/j.dajour.2024.100484, 100484–100484.