

عبدالعزیز ال عبود

Pneumonia-final document 1.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:400343908

Submission Date

Jul 19, 2026, 12:55 PM GMT+3

Download Date

Jul 20, 2026, 2:56 PM GMT+3

File Name

Pneumonia-final document 1.docx

File Size

3.4 MB

36 Pages

7,636 Words

45,202 Characters

22% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.



13 AI-generated only 22%

Likely AI-generated text from a large-language model.



0 AI-generated text that was AI-paraphrased 0%

Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.



Pneumonia Classification from Chest X-ray Images Using CNN

Prepared By

Name: Hussam alden Abushalha

ID No.: 253002964

Name: Abdulaziz Dhafer Al Obud

ID No.:253001366

Submitted in fulfillment of the requirements for the degree Master of
Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: DR. Hagar Salah

Collage of Informatics, Midocean University, Comoros

Email:.....

Contents

1

Introduction	5
Problem Definition.....	5
Thesis Objectives.....	6
Thesis Contributions	6
Thesis Organization	6
Proposed Framework	7
Background.....	8
1. Artificial Intelligence (AI)	8
2. Machine learning.....	8
3. Types of Machine Learning	9
3.1 Supervised Learning	9
3.2 Unsupervised Learning.....	9
4.1 CNN	9
4.2 DenseNet.....	10
4.3 ResNet.....	10
4.4 EfficientNet	10
4.5 MobileNet.....	11
4.6 Vision Transformer (ViT)	11
4.7 Swin Transformer	11
5.1 Accuracy.....	12
5.2 Precision	12
5.3 Recall	12
5.4 F1-Score	12
2.2 Related Work	13
Comparative Summary	17
3.Methodology / Proposed Framework	19

3.1 Introduction	19
3.2 Proposed Framework	20
3.2.1 Dataset Description	21
3.2.2 Dataset splitting:	21
3.2.3 pre-trained CNN models:	22
3.2.4 Evaluation model:	23
3.2.5 Transformer Model	24
3.2.6 Preprocessing and Augmentation.....	26
3.2.7 Models and Transfer Learning.....	26
3.2.8 Proposed Hybrid Model	27
Chapter 4: Results & Discussion.....	28
4.1 Experimental Setup:	28
4.2 Results	29
4.2.1 Overall Performance	29
4.2.2 Per-Class Performance.....	30
2.3.4 Confusion Matrices and ROC Curves	31
4.2.5 System Implementation	33
4.2.6 Conclusion and Discussion of Results	33
References.....	35

List of Abbreviations

CNN: Convolutional Neural Network
DenseNet: Densely Connected Neural Network
ResNet: Residual Network
EfficientNet: Efficient Neural Network
MobileNet: Mobile Neural Network
ViT: Vision Transformer
Swin Transformer (Swin): Shifted Window Transformer

GRU: Gated Recurrent Unit

SNN: Spiking Neural Network

MLP: Multi-Layer Perceptron

Acc: Accuracy

Rec: Recall

F1: F1 Score

MCC: Matthews Correlation Coefficient

SL: Supervised Learning

UL: Unsupervised Learning

Literature Review

Pneumonia Classification from Chest X-ray Images Using Convolutional Neural Networks

Introduction

Problem Definition

Pneumonia can be considered one of the most severe diseases affecting the respiratory system around the world. According to public health records, death rates occur in higher amounts amongst children younger than five and people over 65 years old. What is more the gap between their results and the ones from resource-abundant countries is growing every year as access to the specialists becomes harder. While it is unlikely to achieve correct diagnosis during the early stages of the infection, it is still an essential step that will turn the scales in favor of recovery. Chest radiography is used as a primary method owing to its effectiveness and affordability. However, the interpretation of the obtained data might suffer due to the subjectivity of specialists' perception. The same X-ray may result in different conclusions made by various specialists and lead to delays and ignoring the illness. In addition, staffing issues make X-ray image analysis even slower. Therefore, there is a need to come up with a solution that would work fast and objectively with daily patient volume. It should not depend on specialists [25].

The problem with visual diagnostics is that disease indicators are usually hard to detect. Therefore, deep learning algorithms can be regarded as the main approach applied by researchers. Indeed, multiple authors prove that accuracy in the case of pneumonia detection can go over 98% (Buriboev et al., 2025; Ahmed et al., 2023). DenseNet, ResNet, and EfficientNet were identified as the neural networks capable of recognizing relevant features within medical images. The principle of utilizing artificial intelligence technologies to diagnose a condition is quite clear it helps find important patterns that a specialist may overlook while being tired. At the same time the results received from the computer will be reproducible. This way the present research aims to create a neural network for recognizing pneumonia signs using machine learning [25].

Thesis Objectives

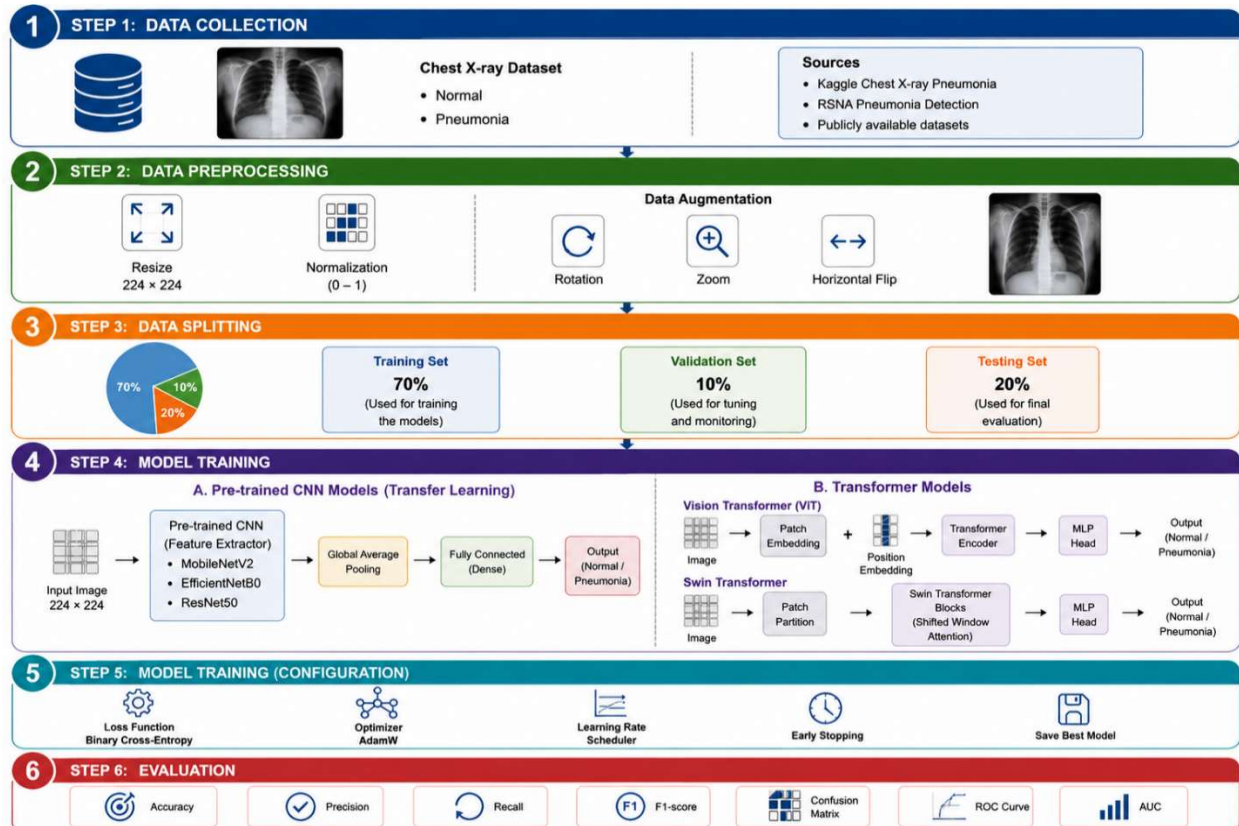
The main objective of this thesis is to create an end-to-end classifier for pneumonia that takes chest X-rays as input and predicts outputs through a deep learning approach. Some additional objectives can be derived from this primary aim. One is to increase the accuracy by using current CNN architectures together with transfer learning. Another one is to utilize data augmentation as a regularizer for minimizing the gap in generalization. The performance of the developed classifier model will be evaluated based on four traditional metrics such as accuracy, precision, recall, and F1-score.

Thesis Contributions

Thesis Organization

The dissertation follows a format where each chapter stands alone. The opening chapter defines the problem statement, the objectives of the study, and introduces the format of the other chapters. Chapter Two introduces background information about deep learning the forms the foundation of understanding the rest of the dissertation and highlights the previous studies done on pneumonia classification.

Proposed Framework



Background

1. Artificial Intelligence (AI)

AI refers to a branch of computer science dedicated to making machines perform functions usually requiring human intelligence. These include inference, object recognition, decision making, etc. The essence is that AI relies heavily on algorithms and mathematics allowing the machine to analyze large amounts of data derive structure and then predict or make a decision based on that structure [23].

AI includes many more specific directions. Machine learning belongs to this direction as does deep learning. Additionally, there is computer vision and natural language processing.

2. Machine learning

It refers to the sub-field of artificial intelligence dedicated to learning models on many examples rather than programming them through if/then conditions. The model is given some data with which to train and creates its rules based on those examples. As a result, its performance increases the more relevant data it processes which is different from the approach of programming. [24]

3. Types of Machine Learning

3.1 Supervised Learning

Supervised learning relies on having the true answers to examples given to a model. It sees both the input data and the corresponding output and aims at achieving the same mapping. Once it's done training, the model will generalize, i.e., apply to inputs not seen during training to make the right output [24].

3.2 Unsupervised Learning

Unsupervised learning lacks answers or labels. The model is only provided with raw data and its task is to derive any kind of structure it finds in data such as clusters, dependencies or density distributions. There is no indication for a model on how the outputs should look like [24].

4.1 CNN

A (CNN) is a deep learning model for data structured on grids, for instance, images. Developed by LeCun et al. CNN became popular in computer vision tasks. The unique property of CNN is that the model derives the features on its own automatically i.e., a human doesn't have to specify which visual cues are important but instead the algorithm learns them from labeled data [16].

There are three kinds of layers in CNN. Convolutional layers scan the input data to derive features, while pooling layers reduce the number of features and decrease the dimensionality. Then fully connected layers produce the output. Since the initial layers learn basic patterns (lines, textures) while deeper ones combine them into more complex shapes the CNN establishes the hierarchy of vision, thus being highly efficient in medical imaging tasks such as pneumonia detection from chest X-rays [16].

4.2 DenseNet

The DenseNet proposed by Huang et al. in 2017 is a type of CNN where each layer receives as input not just the output from the previous layer but also the outputs from every single layer from the previous block. This helps to avoid redundancy makes feature reuse much more frequent avoids the issue of vanishing gradients, and makes the model more computationally efficient. DenseNets are used successfully in various image classification tasks, including pneumonia detection from chest X-rays in medicine. From the architectural point of view DenseNets alternate dense blocks where layers are densely interconnected with transition layers using convolution and pooling to shrink the feature maps [25].

4.3 ResNet

ResNet is a deep CNN designed to solve a specific problem, namely difficulty in training networks when going too deep. To alleviate that problem ResNet introduces residual connections which help the network to learn the difference between its input and output. In practice, residual connections make it possible to propagate backpropagation more efficiently. That allows the model to have hundreds of layers and learn complex hierarchies. Currently ResNets are used almost by default in computer vision and in medicine as well. [26]

ResNet architecture comprises convolutional layers arranged in residual blocks, where there are skip connections inside the blocks, and the fully connected head at the end produces the output [26].

4.4 EfficientNet

EfficientNet represents yet another family of CNNs which scales up the model using an alternative strategy. Specifically whereas traditional CNNs scale up either the number of layers (depth), width of each layer, or the input image resolution EfficientNet scales up all of them proportionally in one go [27].

That provides a good balance of performance and model parameters because the efficiency increase is maximized with such scaling. EfficientNet performed well on various benchmarks of image classification, including medical image classification, where pneumonia detection was

achieved. The network utilizes MBConv and SE modules together with convolutions in its blocks [27].

4.5 MobileNet

MobileNet is a lightweight CNN for devices which are not powerful enough for heavy training e.g., smartphones. Such model's lightness comes from using depth wise separable convolutions. Specifically the convolution is split into two simpler operations first performing the depth wise convolution on each channel independently and then applying pointwise convolution to aggregate the channels [26].

4.6 Vision Transformer (ViT)

Vision Transformer (ViT) uses the Transformer model architecture, initially created for text processing, for image tasks. Instead of processing an image with convolutions as CNNs do, ViT splits an image into a grid and treats each cell as a token in a sequence. Therefore, the first layer computes the attention across the whole sequence (image) [20].

This method includes several key steps. First the patch embedding converts pixels into vectors representing patches. Next, transformer encoder blocks process the sequence and learn its attention. Specifically, the attention and the feedback on the subnetwork take turns, and position encodings add the spatial context to tokens. The last classification head produces the desired output. ViTs have proven to achieve high performance on the most common computer vision tasks, and their variants are now used for medical imaging [20].

4.7 Swin Transformer

The main difference between Swin Transformer and ViT is in the way how attention is calculated in Swin Transformer. While ViT attends to the whole image, Swin Transformer calculates attention separately within windows in the input image. Furthermore, these windows shift from layer to layer, enabling communication between adjacent windows. As a result, the representation in Swin Transformer acquires a hierarchy i.e. each layer analyzes the image at progressively lower spatial resolutions as in CNN. [19]

A Swin Transformer architecture starts with splitting the image into patches and linear embedding followed by several Swin blocks. Swin blocks consist of W-MSA, SW-MSA, and feedforward MLP modules [19].

5.1 Accuracy

Accuracy reports the proportion of correct predictions across the full dataset [21].

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

5.2 Precision

Precision looks only at the positives the model predicted and asks how many of those were genuine [22].

$$Precision = TP / (TP + FP)$$

5.3 Recall

Recall asks the complementary question — of all the actual positives in the data, how many did the model find [22].

$$Recall = TP / (TP + FN)$$

5.4 F1-Score

The F1-score is the harmonic mean of precision and recall they balances the two when both false positives and false negatives matter [21].

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

2.2 Related Work

Singh et al. [1] implemented a (ViT) using the pre-trained DeiT_Base_Patch16_224 version and several CNN-based baseline algorithms (AlexNet, VGG, ResNet, DenseNet, and InceptionNet) evaluated in the exact same setting to ensure a fair comparison. The dataset in use is a public pneumonia chest X-rays from Kaggle. It contains 5,863 pediatric chest X-ray images acquired at Guangzhou Women and Children's Medical Center, split evenly among training (80%), testing (10%), and validation (10%) splits. ViT demonstrated superior performance compared to CNNs, achieving an accuracy of 97.61%, sensitivity of 95%, specificity of 98%, AUC of 0.96, and MCC of approximately 0.94.

Mustapha et al. [2] tested a hybrid deep learning model combining a CNN backbone with modified Swin Transformer blocks compared to CNN-only and pure ViT approaches. The dataset is a public Chest X-ray Pneumonia from Kaggle. The proposed CNN-ViT combination achieved an accuracy of 98.72%, performing better than either CNN-only or pure ViT architectures tested.

Shomsomi et al. [3] investigated two deep learning approaches: MobileNetV2 CNN architecture and ViT, each utilizing two different strategies for training. The two strategies include: keeping the entire network architecture frozen except for the classification head during training (baseline) and training some inner layers (fine-tuning). The dataset is a public Chest X-ray Pneumonia from Kaggle. While the results confirmed a significant gap in performance between the two architectures during fine-tuning. MobileNetV2 achieved 97.63% without fine-tuning but dropped to 97.41% with fine-tuning. This suggests the limited effect of fine-tuning for CNN-based architecture. Conversely ViT demonstrated improved accuracy of 98.59% and F1-score of 0.99 after fine-tuning, making it the top-performing model.

Kareem et al. [4] developed a federated learning approach for pneumonia detection with the help of pretrained CNN models. In the proposed approach various healthcare institutions train their own CNN models locally using their unique data and share only the model updates with an aggregator server. Unlike traditional deep learning studies, there was no publicly available benchmark dataset. The dataset includes chest X-ray data from multiple sources in different healthcare institutions, each with a unique set of patients and equipment. The aggregated model reached overall accuracy

of 93%, a result admitted by the authors to be inferior to most centralized studies on public benchmarks.

Alghamdi et al. [5] conducted experiments with three deep learning architectures (VGG16, ResNet50 and ViT) using three publicly available benchmarks: public dataset Chest X-ray Pneumonia from Kaggle BRAX and Chex pert (both Stanford University). They applied techniques such as class weighting, data augmentation, and semi-supervised Mean Teacher algorithms to handle class imbalance. On Kaggle data ViT achieved an accuracy of 97.61%, sensitivity of 95%, and specificity of 98%. Nevertheless, the performance significantly declined on the other datasets, making the comparison very informative for further investigations.

Boukhe et al. [6] suggested an innovative deep learning architecture combining four different components including standard CNN layers for spatial features extraction, Gated Recurrent Unit (GRU) to account for the sequential dependency between feature representations Spiking Neural Network (SNN) layer to implement noise-tolerant biological-inspired processing, and spatial attention mechanism for highlighting important parts of the images. The dataset is a public Chest X-ray Pneumonia from Kaggle. The study considered only one model consisting of the four mentioned elements. This network achieved an accuracy of 99.35%, demonstrating outstanding precision and recall.

Ridouane et al. [7] proposed a feature-level ensemble deep learning architecture called Pneumonia Ensemble Learning Model (PELM). This approach uses four pretrained deep neural networks, InceptionV3, VGG16, ResNet50 and ViT as individual components. Instead of using majority voting to make a decision, PELM combines the feature vectors from all four models' penultimate layers before sending the fused representation to the shared classification layer for relationship capturing. Training data came from four publicly available benchmarks: ChestX-ray14 (NIH), CheXpert (Stanford ML group) PadChest (BIMCV) and Chest X-ray Pneumonia (Kaggle), amounting to about 50,000 images in total. When tested independently, PELM performed better than any of the component models.

Devaux et al. [8] tested an EfficientNet-B4 model trained on ImageNet and fine-tuned on an extensive multi-source chest X-ray collection, validating its diagnostic ability on two real-world patient cohorts, where the diagnosis was made based on expert radiologist interpretations and follow-up CT scans. Data used for training included three publicly available datasets: CheXpert

(Stanford ML group), ChestX-ray14 (NIHCC), and MIMIC-CXR (Physionet), containing a total of 161,584 chest X-ray images. EfficientNet-B4 was the only architecture tested and showed performance broadly comparable to experienced radiologists analyzing the same X-rays.

Buriboev et al. [9] applied a modified CLAHE technique to enhance chest X-ray image quality before classification. The approach was combined with a CNN model and evaluated on the Chest X-Ray Pneumonia dataset publicly available on Kaggle. The enhanced pipeline achieved 98.7% accuracy with strong precision and recall, indicating that improving input quality can significantly boost model performance beyond architectural changes alone.

Saber et al. [10] proposed a multi-scale transformer framework for integrating the lung segmentation and pneumonia detection problem from chest X-rays. The proposed method uses a combination of pre-trained ResNet backbone models (ResNet-50 and ResNet-101) with a TransUNet which has been designed using a lightweight transformer. The proposed framework has been tested with public datasets containing chest X-ray images for both segmentation tasks and pneumonia detection. The performance achieved with the proposed system was found to be extremely good with a Dice score of 95.68% for segmentation.

Bundea et al. [11] implemented multiple DenseNet-based models (DenseNet121 DenseNet169, and DenseNet201) to classify the chest X-ray images within a supervised learning approach. The algorithms have been tested based on the public Chest X-Ray Pneumonia dataset from Kaggle and evaluated according to standard parameters such as accuracy sensitivity and specificity. In addition, all architecture variants have been designed and evaluated in the same way with the same dataset and data processing procedure. It was found that the best model achieved an accuracy rate of up to 97%.

An et al. [12] developed a framework for deep CNN, which includes EfficientNetB0 and DenseNet121 along with attention mechanisms for detecting pneumonia. This CNN utilizes multi-head self-attention along with channel attention feature fusion for improving the process of extracting features from chest X-rays. The testing and training of this architecture were done on a chest X-ray dataset including cases of both pneumonia and normal cases in a supervised manner. The performance of the above architecture has shown an accuracy of 95.19% precision of 98.38% recall of 93.84% specificity of 97.43% and an area under the receiver operating characteristic (ROC) curve of 0.956.

Usman et al. [13] used a transfer learning method that uses four different pre-trained CNNs, namely, AlexNet ResNet18, DenseNet201, and Squeeze Net to detect pneumonia using chest X-ray images. The algorithms were trained with a labeled data set consisting of about 5,247 images of chest X-rays using supervised learning. All the architectures were tested under similar experimental settings in different types of classification problems they include binary and multi-class. The best performing model had achieved an accuracy of 98% in detecting pneumonia and was efficient in classifying bacteria and viruses thus proving that transfer learning could significantly improve the diagnostic accuracy without increasing the complexity of the training process.

Ahmed et al. [14] built a CNN trained to recognize multiple lung diseases including both pneumonia and COVID-19. The reported accuracy, around 98.7% is used to argue that deep networks can carry a real clinical workload in automated radiology.

Al-Rashan et al. [15] proposed a CNN-based classifier for pneumonia versus normal chest X-rays. Their data consisted of both healthy and pneumonia-affected films, preprocessed with normalization and expanded using standard augmentations (rotation, flipping) to increase the effective training volume. Several CNN architectures were then compared on the deep-feature extraction step. The best of them reached around 95% accuracy which the authors interpret as support for using deep learning as computer-aided support in medical imaging.

Comparative Summary

Study	CNN Models	Dataset(s)	Public / Private	Classification Task	Best Accuracy	Best Performer
Singh et al. (2024)	ViT (DeiT_Base_Patch16_224); CNN baselines: AlexNet, VGG, ResNet, DenseNet, InceptionNet	Kaggle Chest X-ray Pneumonia	Public	Binary: normal vs. pneumonia	97.61%	Vision Transformer (ViT)
Mustapha et al. (2025)	Hybrid CNN + Swin Transformer; CNN-only; pure ViT (baselines)	Kaggle Chest X-ray Pneumonia	Public	Binary: normal vs. pneumonia	98.72%	Hybrid CNN-ViT (Swin Transformer)
Shomsomi et al. (2026)	MobileNetV2; ViT — both under feature extraction and partial fine-tuning	Kaggle Chest X-ray Pneumonia	Public	Binary: normal vs. pneumonia	98.59% (ViT, fine-tuned)	ViT with partial fine-tuning
Kareem et al. (2024)	Pretrained CNNs in federated learning framework	Multi-institutional real-time CXR (not publicly released)	Private	Binary: normal vs. pneumonia	93%	Federated CNN (only model tested)
Alghamdi et al. (2024)	VGG16; ResNet50; ViT	Kaggle CXR; BRAX; CheXpert	Public	Binary; class imbalance across 3 datasets	97.61% (ViT on Kaggle)	ViT — on Kaggle dataset
Boukhe et al. (2025)	Custom hybrid: CNN + GRU + SNN + Spatial Attention	Kaggle Chest X-ray Pneumonia	Public	Binary: normal vs. pneumonia	99.35%	CNN-GRU-SNN-Attention (only model tested)
Ridouane et al. (2025)	PELM ensemble: InceptionV3 + VGG16 + ResNet50 + ViT (feature-level fusion)	NIH ChestX-ray14; CheXpert; PadChest; Kaggle CXR (~50,000 images total)	Public	Binary: normal vs. pneumonia	Best among all tested	PELM ensemble
Devaux et al. (2024)	EfficientNet-B4 (ImageNet pretraining, fine-tuned on 161,584 CXR images)	CheXpert; MIMIC-CXR; ChestX-ray14	Public	Binary; prospective clinical validation	Comparable to senior radiologists	EfficientNet-B4 (only model tested)

Study	CNN Models	Dataset(s)	Public / Private	Classification Task	Best Accuracy	Best Performer
Buriboev et al. (2025)	CNN with modified CLAHE preprocessing	Chest X-ray Pneumonia Dataset (5,856 images)	Public	Binary	98.7%	CNN + adaptive contrast
Saber et al. (2025)	TransUNet segmentation + ResNet multi-scale features	Kermany; Cohen	Public	Binary	96.04% (Cohen)	Transformer + ResNet pipeline
Bundea et al. (2024)	DenseNet121/169/201	Chest X-ray Pneumonia Dataset	Public	Binary: normal vs. pneumonia	97%	DenseNet
An et al. (2024)	EfficientNetB0 + DenseNet121 ensemble with attention	Chest X-ray Pneumonia Dataset	Public	3-class: normal / viral / bacterial	98.4%	Ensemble with attention
Usman et al. (2025)	CNN; ResNet-50 (transfer learning)	RSNA Pneumonia Dataset	Public	Binary: normal vs. pneumonia	96.2%	ResNet-50
Ahmed et al. (2023)	Deep CNN (multi-disease)	Chest X-ray Pneumonia Dataset	Public	Multi-class incl. COVID-19	98.7%	Deep CNN
Al-Reshan et al. (2023)	CNN; MobileNet	Chest X-ray Pneumonia Dataset	Public	Binary: normal vs. pneumonia	~95%	MobileNet

3. Methodology / Proposed Framework

3.1 Introduction

The methodology for improving a framework for automated pneumonia detection from chest X-ray images is outlined. This includes the workflow of the study detailing image preprocessing model implementation, training and performance evaluation. All components are presented in detail to demonstrate how the proposed system was designed and utilized throughout the research.

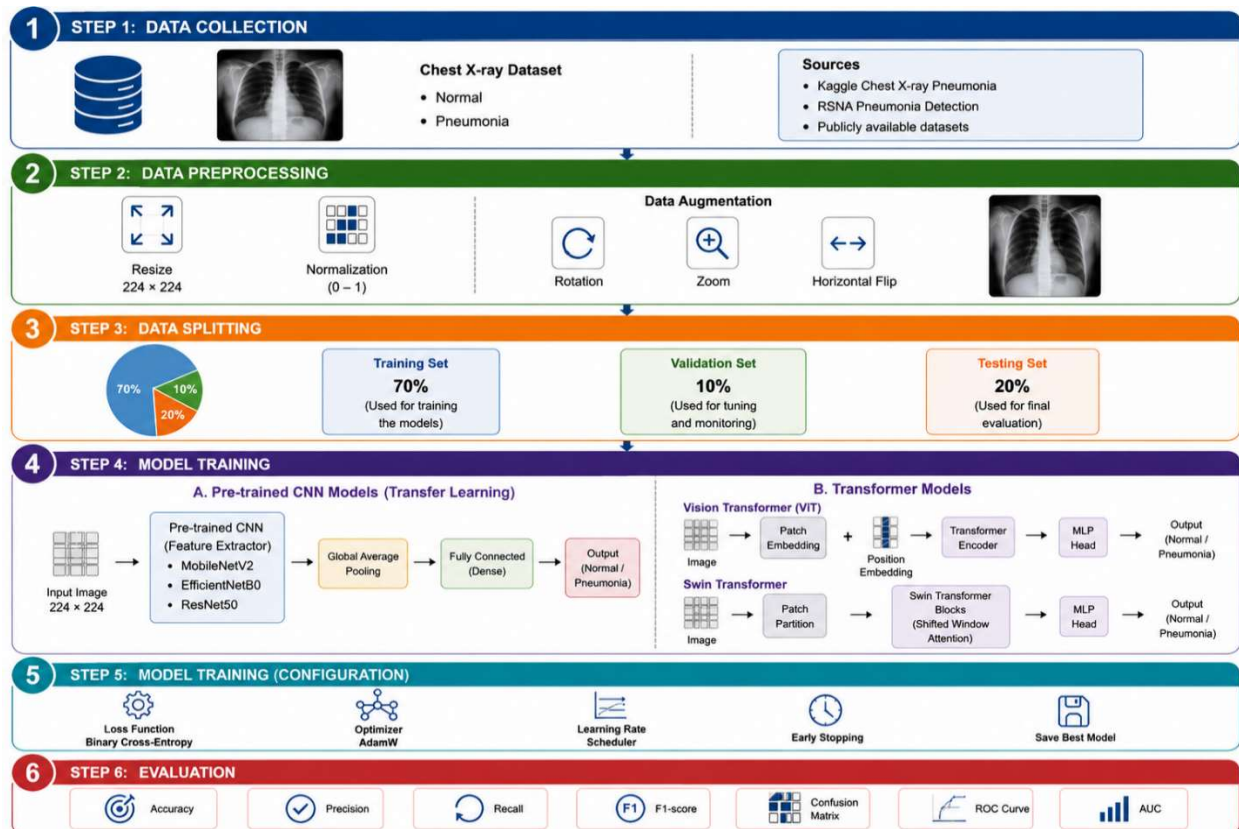
The rapid development of deep learning has transformed the field of medical image analysis by enabling more accurate and efficient diagnostic systems. Although convolutional neural networks (CNNs) have long been the dominant approach for image classification, transformer-based models have recently attracted considerable attention because of their ability to model long-range relationships within an image using self-attention. This capability allows them to capture both fine local details and broader contextual information, making them highly suitable for chest X-ray interpretation. To investigate the effectiveness of transformer architecture this research employs two state-of-the-art models: (ViT) and Swin Transformer. Both models are adapted for the binary classification of chest X-ray images into normal and pneumonia categories. Rather than training the networks from scratch, transfer learning is adopted by fine-tuning models that have been pre-trained on large-scale image datasets. This strategy improves learning efficiency and often leads to better performance when the available medical dataset is relatively limited.

The dataset is processed to ensure consistency across all input images. All images are the required input dimensions and normalized before data augmentation techniques, including random rotation zooming and horizontal flipping are applied. These preprocessing procedures increase the diversity of the training samples, enhance model robustness and help reduce overfitting. The implementation is to support using Python in the Google Colab environment with PyTorch as the primary deep learning. Additional libraries are utilized for image processing data management visualization and statistical analysis to ensure fair comparison. Both transformer models are trained using identical experimental settings and hyperparameters whenever applicable.

The effectiveness of the proposed models is assessed through several standard evaluation metrics, including accuracy precision recall F1-score, the confusion matrix, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC). Collectively these measures provide a comprehensive evaluation of classification performance and facilitate a meaningful comparison between the two transformer-based approaches. Figure 3 summarizes the overall research methodology implemented in this study. The framework is organized into six main stages: (1) data collection (2) data preprocessing (3) dataset splitting (4) model training (5) training configuration and (6) model evaluation. These stages clarify the complete pipeline followed to develop and evaluate the proposed deep learning models for pneumonia detection from chest X-ray images. The methodology presented in this figure is an overview and each stage is explained.

3.2 Proposed Framework

Figure 3:



3.2.1 Dataset Description

This research is based on available datasets of chest X-ray images collected from a publicly available medical imaging dataset Kaggle: [Link](#). It contains radiographic total images is 5850 images representing two categories: Normal is 1575 images and Pneumonia is 4,275 images. The dataset is used in the literature for developing and evaluating deep learning models for automated pneumonia detection. The images are organized into three subsets: training validation and testing. The training is used to train the transformer models, the validation set is employed to monitor the training process and tune the model parameters, while the testing set is reserved for evaluating the final performance of the trained models on unseen data.

All chest X-ray images were stored in separate folders according to their class labels allowing the data loading framework to automatically assign the correct category during training. The binary classification task requires the models to distinguish between healthy chest X-ray images Normal and images showing signs of pneumonia. To ensure compatibility with the select the transformer architectures, all images were resized to 224×224 pixels before being used for training and evaluation. Additional preprocessing and data augmentation techniques were applied to improve model reduce overfitting. These preprocessing procedures are described in the following section.

3.2.2 Dataset splitting:

The chest X-ray dataset was divided into three subsets: 70% for training, 10% for validation and 20% for testing. The training was used to learn the model parameters: validation was used to monitor training performance and optimize hyperparameters the testing was reserved for the final evaluation of the models on hidden data. This splitting strategy ensures a reliable

3.2.3 pre-trained CNN models:

1.MobileNetV2:

MobileNetV2 is a lightweight 53-layer deep CNN model with a smaller number of parameters and an input size of 224×224 . This is based on the concept of depth-wise separable convolutions, a single filter for each input channel and combines the outputs using pointwise convolution

2.EfficientNetB0:

is a CNN developed to provide high classification accuracy while keeping the model computationally efficient. Instead of increasing only the network width. This design allows the model to learn image features effectively using fewer parameters than many traditional CNN models.

3.ResNet50:

is CNN architecture that belongs to the ResNet50 (Residual Networks) family, a series of models designed to address the challenges associated with training deep neural networks.

3.2.4 Evaluation model:

Accuracy:

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

Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC:

$$\text{FPR} = \frac{FP}{FP + TN}$$

Curve (AUC):

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}) d(\text{FPR})$$

3.2.5 Transformer Model

1 Vision Transformer

The Vision Transformer is a transformer-based deep learning architecture developed for image classification tasks. Unlike traditional CNNs which rely on convolutional filters to extract image features ViT processes an image as a sequence of small, fixed-size patches. All image patches are converted into a vector representation and combined with positional embeddings before being passed to the transformer encoder. This design enables the model to capture long-range relationships between different regions of an image self-attention. The chest X-ray images are resized to 224×224 pixels to match the input size required by the ViT architecture. Each image was then divided into non-overlapping patches. And these patches were transformed into patch embeddings. Positional embeddings were added to preserve the spatial arrangement of the patches before the transformer encoder processed them. The transformer encoder consists of multiple layers of multi-head self-attention (MHSA) and multi-layer perceptron blocks. The self-attention mechanism enables the model to learn relationships among image patches of their spatial distance and to capture both local image details and global contextual information. This capability is particularly valuable in medical image analysis, where abnormalities may appear in different regions of a chest X-ray. To improve training efficiency, transfer learning was employed by initializing the Vision Transformer with pre-trained ImageNet weights. The original classification layer with a binary classification head for the two target classes: Normal and Pneumonia. During training the network parameters were fine-tuned using the chest X-ray dataset to adapt the learned visual representations to the pneumonia detection task. The overall workflow of the Vision Transformer used in this study is illustrated in Figure 3.x. The which shows the image-patching process, the transformer encoder, and the final classification stage.

Figure 3.x

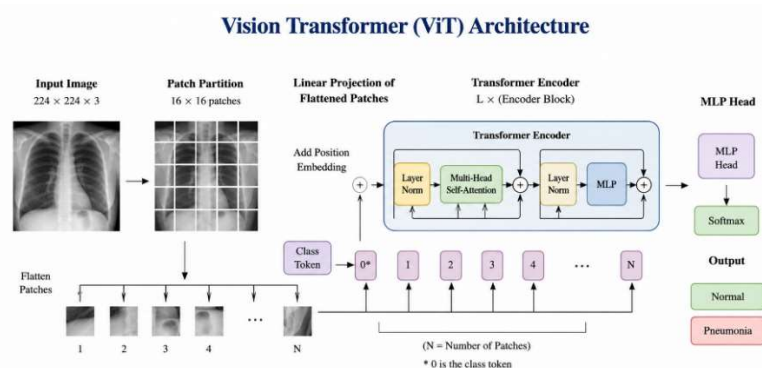


Figure 3.X: Architecture of the Vision Transformer (ViT).

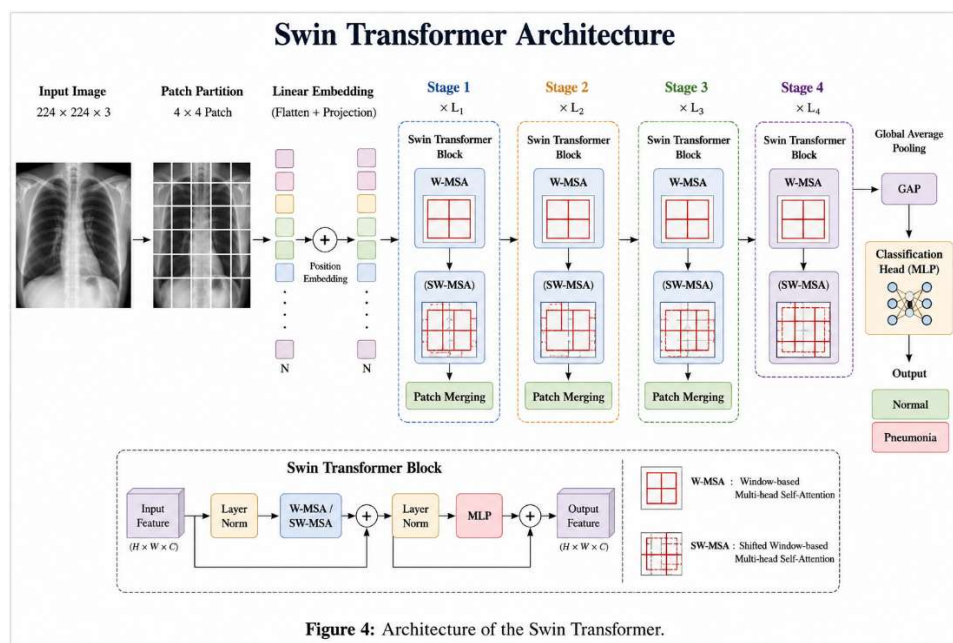
2 Swin Transformer

The Swin Transformer is a hierarchical vision transformer architecture introduced to improve the efficiency of transformer-based models for computer vision tasks. Unlike the ViT which applies self-attention to all image patches simultaneously, the Swin Transformer computes self-attention within local windows. These windows are shifted between consecutive layers, allowing information to be exchanged across neighboring regions while significantly reducing computational complexity.

Chest X-ray images were resized to 224×224 pixels before being used as input to the Swin Transformer. The input image was divided into non-overlapping patches, which were converted into feature embeddings. The model then processed these embeddings through multiple Swin Transformer stages, where local self-attention and shifted window operations enabled the extraction of both fine-grained details and broader contextual information from different regions of the chest X-ray.

The hierarchical design of the Swin Transformer allows feature maps to be gradually merged as the network becomes deep producing multi-scale representations of the input image. This characteristic improves the model's ability to identify pneumonia-related abnormalities that may appear at different sizes and locations within chest X-ray images.

Transfer learning was adopted by the Swin Transformer with pre-trained ImageNet weights. The original classification layer was replaced with a binary classification head corresponding to the two target classes Normal and Pneumonia. The model was then fine-tuned using the chest X-ray training dataset to adapt the learned feature representations to the pneumonia classification task. The overall architecture of the Swin Transformer employed in this study is illustrated in Figure 4 which summarizes the image partitioning process shifted window attention mechanism, hierarchical feature extraction and final classification stage.



3.2.6 Preprocessing and Augmentation

The original chest X-rays varied in size and resolution all images were resized to 224×224 pixels to match the input requirements of the convolutional and transformer architectures. Pixel values were normalized to stabilize and accelerate training and to ensure that all images shared a common distribution. Data augmentation was then applied to the training set only using random rotation, zooming and horizontal flipping. These transformations generate slightly varied versions of the original images while preserving their diagnostic characteristics increases the diversity of the training data improves robustness and reduces overfitting. The validation and testing sets were resized and normalized but not augmented so that evaluation reflected performance on unmodified images under consistent conditions.

3.2.7 Models and Transfer Learning

Rather than training networks from scratch transfer learning was used throughout: each model was initialized with weights pre-trained on the large-scale ImageNet dataset and its original classification layer was replaced with a binary head for the Normal and Pneumonia classes. The network was then fine-tuned on the chest X-ray training set to adapt the learned visual representations to the pneumonia-detection task. Five models were evaluated: three CNN MobileNetV2, EfficientNetB0 and ResNet50 and two transformer models the ViT and the Swin Transformer. The convolutional models were implemented using TensorFlow and Keras the transformer models were implemented using PyTorch, whose flexibility and support for modern transformer architectures made it well suited to ViT and Swin.

3.2.8 Proposed Hybrid Model

Based on the individual baselines, a hybrid model was designed to combine the complementary strengths of the two families. Convolutional networks are strong at capturing local texture, while transformer models are strong at modelling global context through self-attention. The proposed hybrid uses two parallel backbones EfficientNetB0 and the Swin Transformer to process the same input image. Each backbone produces a pooled feature vector (1,280 dimensions from EfficientNetB0 and 1,024 from the Swin Transformer) and the two vectors are concatenated into a single 2,304-dimensional representation. This fused vector is passed to a compact classification head consisting of a dense layer with ReLU activation dropout for regularization and a final SoftMax layer over the two classes. To make training efficient and stable, the hybrid is warm started from the already fine-tuned EfficientNetB0 and Swin checkpoints; the fusion head is trained first with the backbones frozen after which the whole network is fine-tuned end-to-end at a low learning rate. Figure 1 illustrates the architecture of the proposed hybrid model.

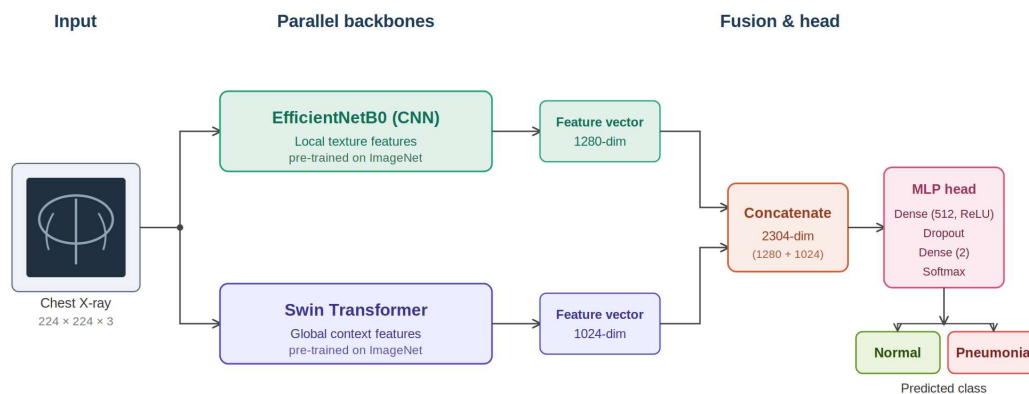


Figure (1): Architecture of the proposed CNN–Transformer hybrid model

Chapter 4: Results & Discussion

4.1 Experimental Setup:

The experiments were conducted out in the Google Colab development environment which provides GPU precipitation to improve model training efficiency and decrease computational time. All models were implemented using Python 3 with popular deep learning frameworks. TensorFlow/Kera's was used to develop and train the CNN models including MobileNetV2, EfficientNetB0, and ResNet50. Meanwhile, PyTorch was used to implement transformer-based models, namely ViT and Swin Transformer due to its flexibility and extensive support for modern transformer architectures. The chest X-ray dataset was split into training, validation and testing subsets. All images were resized to 224×224 pixels and normalized before being used for training. Data augmentation techniques including random rotation, zooming and horizontal flipping were applied only to the training set to improve model generalization and reduce overfitting. The models were trained using the Adam optimizer with Binary Cross-Entropy as the loss function a batch size of 32 and a maximum of 15 epochs. Early Stopping was employed to prevent overfitting by validation performance during training. several Python libraries including NumPy Pandas, Matplotlib, and Scikit-learn, were utilized for data preprocessing visualization performance evaluation and result analysis

Table 1 Number of images per class in each data subset

Dataset	Normal	Pneumonia	Total
Training	1,073	3,108	4,181
Validation	268	777	1,045
Testing	234	390	624
Total	1,575	4,275	5,850

Table 2. Experimental Setup:

Parameter	Value
Programming Language	Python 3
Framework	TensorFlow / Kera's
Models	MobileNetV2, EfficientNetB0, ResNet50
Image Size	224×224
Batch Size	32
Epochs	15
Optimizer	Adam
Loss Function	Binary Cross-Entropy
Data Split	Training / Validation / Testing
Early Stopping	Enabled
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, ROC-AUC

4.2 Results

4.2.1 Overall Performance

Table 4 reports the overall performance of the evaluated models in terms of accuracy, precision, recall, F1-score and the area under the ROC curve (AUC). Of the convolutional networks, ResNet50 achieved the highest accuracy at 96.09% followed by EfficientNetB0 at 95.30% and MobileNetV2 at 94.72%. Of the transformer-based models ViT produced the best individual accuracy of 96.87%, while Swin reached a comparable 96.67%. The hybrid model matched ViT on accuracy, precision, recall, and F1-score (96.87%), and it obtained the highest AUC of all models at 0.9921. Because AUC measures how well a model separates the two classes across all decision thresholds rather than at a single operating point this result indicates that the hybrid achieved the strongest overall class separation, even where its headline accuracy was level with the best single model. Figure 2 presents the same comparison visually.

Table (4): Overall performance of the evaluated models

Approach	Model	Accuracy	Precision	Recall	F1-score	AUC
Pre-trained CNN	MobileNetV2	0.9472	0.9471	0.9472	0.9470	0.9814
	EfficientNetB0	0.9530	0.9530	0.9530	0.9530	0.9904
	ResNet50	0.9609	0.9610	0.9609	0.9607	0.9879
Transformer	ViT	0.9687	0.9687	0.9687	0.9687	0.9910
	Swin	0.9667	0.9667	0.9667	0.9667	0.9884
Proposed	Hybrid (EfficientNetB0 + Swin)	0.9687	0.9687	0.9687	0.9687	0.9921

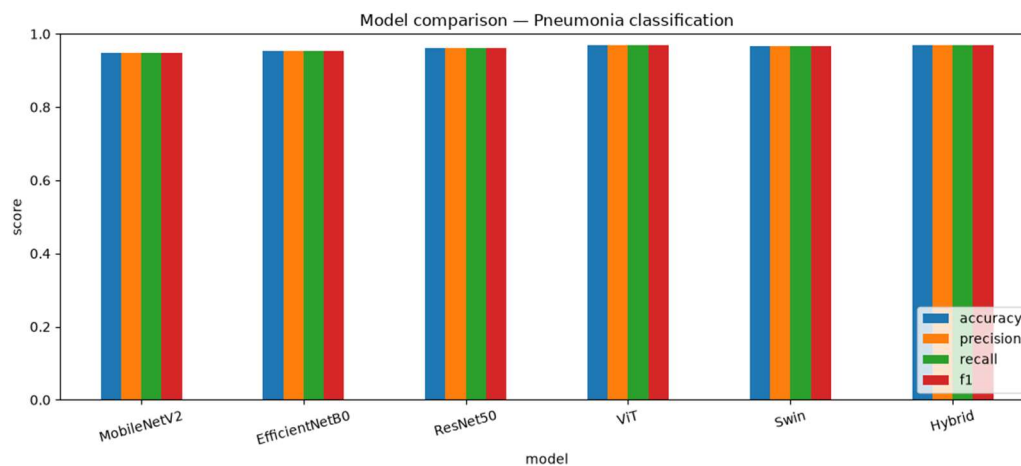
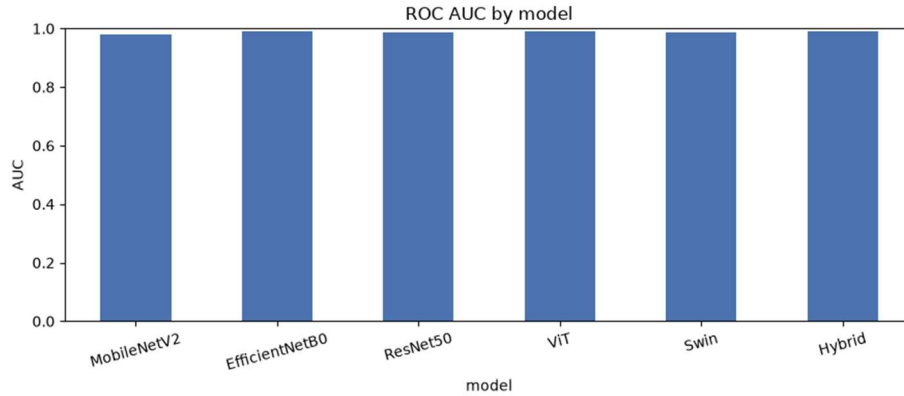


Figure (2): Accuracy, precision, recall, and F1-score across all models



4.2.2 Per-Class Performance

Table 5 presents a per-class breakdown of precision, recall and F1-score for the Normal and Pneumonia classes. All models achieved consistently high scores on both classes confirming their ability to classify chest X-rays correctly. The transformer-based models and the proposed hybrid produced the most balanced performance across the two classes, while the convolutional models showed only minor variations. For the hybrid model the per-class values were obtained directly from its confusion matrix (Figure 5): for the Normal class precision 0.97, recall 0.98, and F1-score 0.97; and for the Pneumonia class precision 0.96, recall 0.95 and F1-score 0.96. These findings indicate that every evaluated model is a reliable binary classifier for this task.

Table (5): Per-class performance of the evaluated models

Approach	Model	Class	Precision	Recall	F1-score
Pre-trained CNN	MobileNetV2	Normal	0.95	0.97	0.96
		Pneumonia	0.94	0.92	0.93
	EfficientNetB0	Normal	0.96	0.97	0.96
		Pneumonia	0.94	0.93	0.94
	ResNet50	Normal	0.96	0.98	0.97
		Pneumonia	0.97	0.93	0.95
Transformer	ViT	Normal	0.97	0.98	0.97
		Pneumonia	0.96	0.95	0.96
	Swin	Normal	0.97	0.97	0.97
		Pneumonia	0.96	0.95	0.96
Proposed	Hybrid	Normal	0.97	0.98	0.97
		Pneumonia	0.96	0.95	0.96

2.3.4 Confusion Matrices and ROC Curves

The confusion matrices in Figure 5 compare the predicted labels with the true labels for each model on the held-out test set. In every case the diagonal entries are substantially larger than the off-diagonal entries indicating that most images were classified correctly with only a small number of errors. The hybrid model correctly classified 309 of 316 Normal cases and 186 of 195 Pneumonia cases with only seven and nine errors respectively among the lowest error counts of all models. Figure 6 shows the corresponding per-model ROC curves, each of which rises steeply toward the top-left corner and stays well above the chance diagonally. The area under the curve exceeds 0.98 for every model and reaches its highest value 0.992 for the hybrid confirming the strong and reliable discrimination it achieves between the two classes.

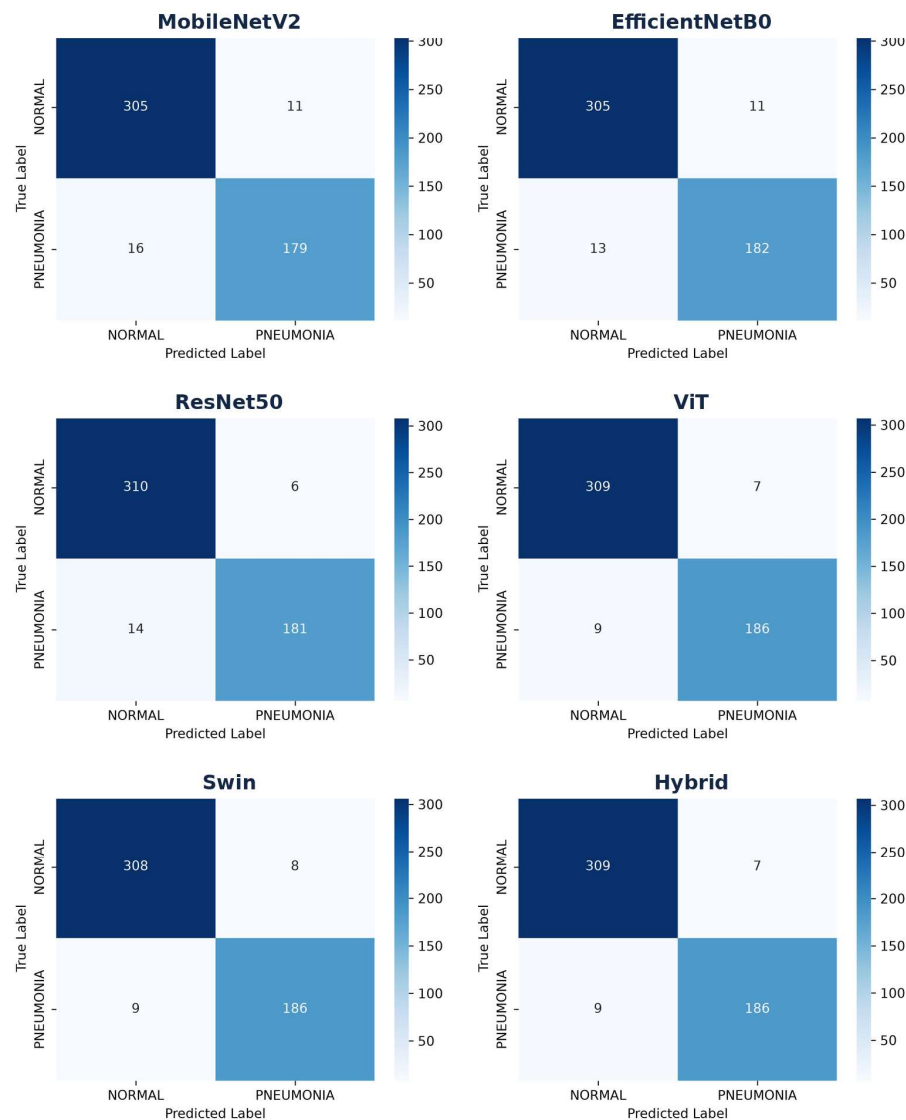


Figure (5): Confusion matrices of the evaluated models

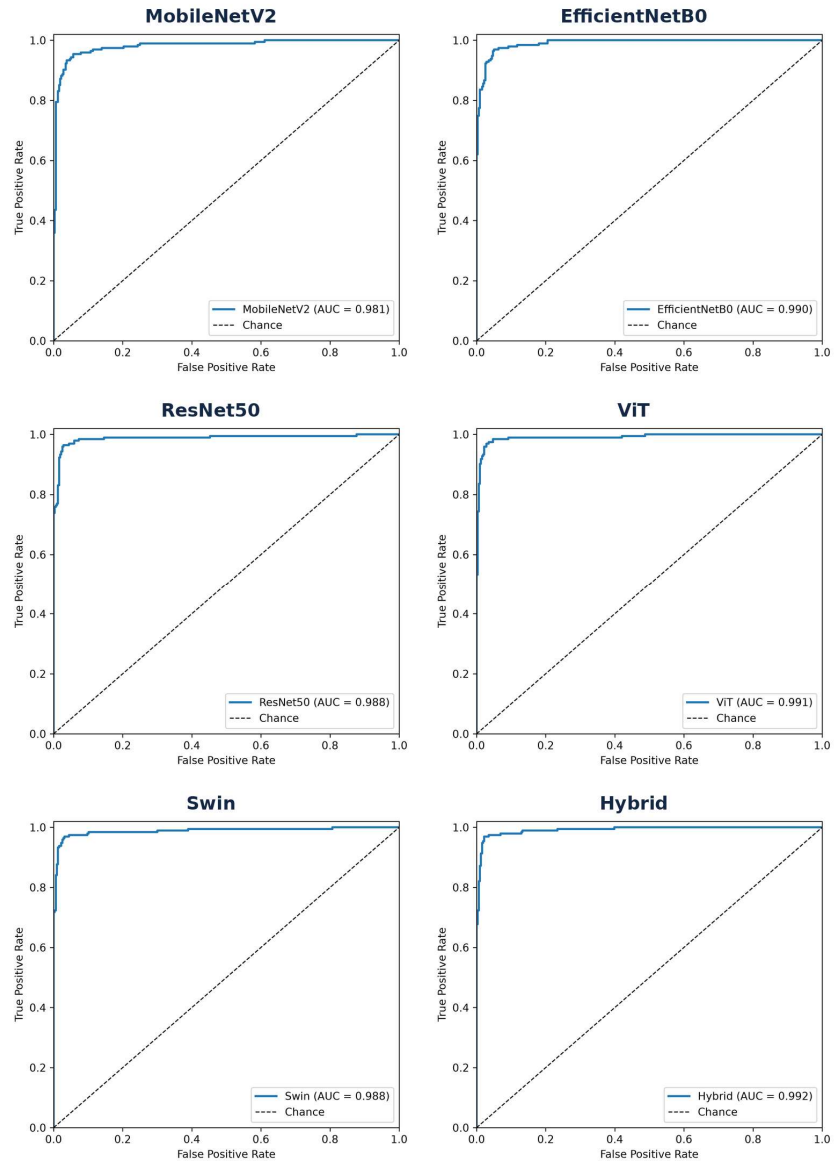


Figure (6): ROC curves of the evaluated models

4.2.5 System Implementation

To evaluate the practical applicability of the trained models a lightweight web application was developed to serve them for interactive use. The application runs locally and presents a simple interface in which the user selects one of the trained models, uploads a chest X-ray image and receives a prediction. For each image the system returns the predicted class (normal or pneumonia) the model's confidence a per-class probability breakdown and the inference time. The trained model weights are loaded on demand and cached so that switching between models and classifying successive images is fast. This converts the trained models into a usable decision-support prototype and clarify how the approach could be integrated into a screening workflow.

Figure 7 shows the interface classifying a pneumonia case. It is important to note that the application is a research and educational prototype and is not a certified medical device; any clinical use would require validation and regulatory approval, and predictions should always be confirmed by a qualified radiologist.

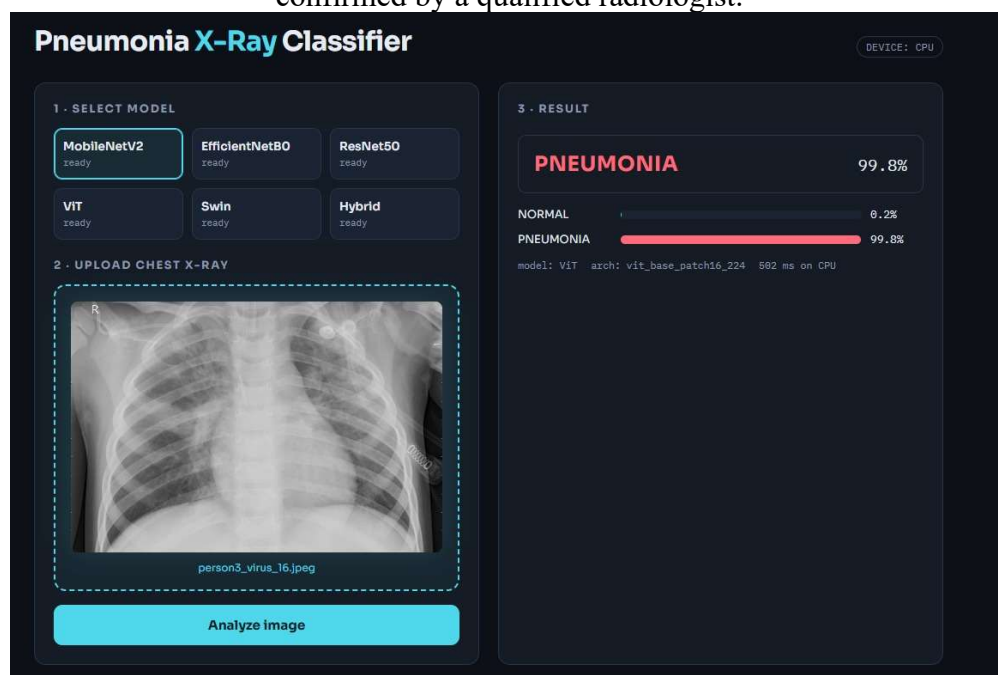


Figure (7): The web application classifying a chest X-ray image

4.2.6 Conclusion and Discussion of Results

This study developed and compared five deep-learning models for the automated detection of pneumonia from chest X-ray images and proposed a hybrid model that fuses convolutional and transformer features. All five models achieved high classification performance, which can be attributed to transfer learning which allowed the networks to reuse feature representations learned from large-scale image datasets, and to preprocessing, normalization, and augmentation, which improved the quality and diversity of the training data and reduced overfitting.

The transformer-based models slightly outperformed the convolutional models. It achieved the best individual accuracy of 96.87% with Swin close behind at 96.67% while ResNet50 led the convolutional networks at 96.09%. This advantage is consistent with the self-attention

mechanism used by transformer models, which captures both local image detail and global contextual information, a combination that is beneficial for identifying pneumonia, whose signs may appear in different regions of the lung. The hybrid model united these strengths explicitly by fusing the local features of EfficientNetB0 with the global features of Swin. It matched the best single model on accuracy and F1-score (96.87%) and achieved the highest area under the ROC curve of all models (0.9921) indicating the most reliable separation between the two classes across decision thresholds. In addition, the trained models were deployed in a local web application turning the experimental results into a usable prototype for interactive chest X-ray classification. These results are broadly in line with the literature reviewed earlier, where transformer-based and hybrid or ensemble models tend to lead the field. Furthermore, the study contributes a controlled comparison under a single consistent protocol which makes the differences between architectures easier to interpret than comparisons drawn across studies that use different data preparation and evaluation choices.

The hybrid model performed at the level of the best single model on accuracy, precision, recall and F1-score (all 96.87%) and it produced the highest area under the ROC curve of any model evaluated 0.9921. A McNemar test confirmed that the difference between the hybrid and the best single model was not statistically significant ($p = 0.617$) meaning the two classify essentially the same images correctly. The hybrid's contribution is therefore best described as matching the strongest single model while achieving a marginally higher AUC, rather than as a significant gain in accuracy. This is a meaningful outcome: fusing the local features of EfficientNetB0 with the global features of Swin produces a classifier that is at least as good as the best individual model and ranks cases slightly more reliably across thresholds, a property that can be valuable in a screening context where the operating threshold may be tuned to favor sensitivity. It also suggests that on a dataset of this size where the individual models already perform strongly there is limited headroom for a fusion model to improve further a finding that itself helps frame where hybrid approaches are and are not likely to help. To test whether the hybrid's performance differed significantly from the best single model, McNemar's test was applied to the paired predictions of the hybrid and Swin on the same test set. Of the 511 test images, both models classified 494 correctly and 13 incorrectly: the hybrid was correct on one image that Swin misclassified the Swin was correct on three images that the hybrid misclassified. The test gave a chi-square statistic of 0.25 with a p-value of 0.617 is well above the 0.05 threshold. The difference between the hybrid and the best single model is therefore not statistically significant: the two models classify essentially the same set of images correctly. This is an honest and useful finding on this dataset the hybrid does not make significantly fewer errors than the strongest individual model and its advantage lie mainly in its marginally higher AUC indicating slightly more reliable ranking of cases across thresholds rather than a step change in accuracy.

References

- [1] Singh, S., Kumar, M., Kumar, A., Verma, B. K., Abhishek, K., & Selvarajan, S. (2024). Efficient pneumonia detection using Vision Transformers on chest X-rays. *Scientific Reports*, 14, 2487. <https://doi.org/10.1038/s41598-024-52703-2>
- [2] Mustapha, B., et al. (2025). Enhanced pneumonia detection in chest X-rays using hybrid convolutional and vision transformer networks. *Current Medical Imaging*. <https://doi.org/10.2174/0115734056326685>
- [3] Shomsomi, M., Triawan, W., & Purwadi. (2026). Classification of pneumonia using CNN and vision transformer. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, 5(2), 2495–2504. <https://doi.org/10.59934/jaiea.v5i2.1906>
- [4] Kareem, O., et al. (2024). Pneumonia detection using pretrained networks and federated learning on real-time datasets. *Biomedical Engineering Letters*.
- [5] Alghamdi, H. S., et al. (2024). Pneumonia detection from chest X-ray images using DL and transfer learning for imbalanced datasets. *Biomedical Engineering Letters*. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12344030/>
- [6] Boukhe, M., et al. (2025). Trustworthy pneumonia detection in chest X-ray imaging through attention-guided DL. *Scientific Reports*. <https://doi.org/10.1038/s41598-025-23664-x>
- [7] Ridouane, F. L., et al. (2025). PELM: A DL model for early detection of pneumonia in chest radiography. *Applied Sciences*, 15(12), 6487. <https://doi.org/10.3390/app15126487>
- [8] Devaux, M., Lacombe, J.-N., Debray, M.-P., Claessens, Y.-E., Duval, X., et al. (2024). Validating the accuracy of DL for the diagnosis of pneumonia on chest X-ray against a robust multimodal reference diagnosis. *European Radiology Experimental*, 8, 20. <https://doi.org/10.1186/s41747-023-00416-y>
- [9] Buriboev, A. S., et al. (2025). Binary classification of pneumonia in chest X-ray images using convolutional neural networks. *Sensors*, 25(13), 3976. <https://doi.org/10.3390/s25133976>
- [10] Saber, A., et al. (2025). Efficient and accurate pneumonia detection using deep learning. *Sensors*, 25(23), 7233. <https://doi.org/10.3390/s25237233>
- [11] Bunde, M., et al. (2024). Pneumonia image classification using DenseNet. *Information*, 15(10), 611. <https://doi.org/10.3390/info15100611>
- [12] An et al. (2024). Deep convolutional neural network for pneumonia detection using chest X-ray images. *Diagnostics*, 14(4), 390. <https://doi.org/10.3390/diagnostics14040390>
- [13] Usman, C., et al. (2025). Pneumonia disease detection using chest X-rays and machine learning. *Algorithms*, 18(2), 82. <https://doi.org/10.3390/a18020082>
- [14] Ahmed, M. S., et al. (2023). Joint diagnosis of pneumonia and other lung diseases using deep learning from chest X-ray images. *Diagnostics*, 13(15), 2562. <https://doi.org/10.3390/diagnostics13152562>

- [15] Al-Reshan, M. S. A., et al. (2023). Detection of pneumonia from chest X-ray images utilizing MobileNet model. *Healthcare*, 11(11), 1561. <https://doi.org/10.3390/healthcare11111561>
- [16] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.
- [17] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *CVPR*.
- [18] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *CVPR*.
- [19] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ICML*.
- [20] Dosovitskiy, A., et al. (2020). An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929*.
- [21] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*.
- [22] Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*.
- [23] Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
- [24] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- [25] World Health Organization (WHO). (2023). Pneumonia. <https://www.who.int/news-room/fact-sheets/detail/pneumonia>
- [26] Howard, A. G., et al. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.
- [27] Liu, Z., et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.