

عبدالعزيز ال عبود

Pneumonia Classification from Chest X-ray Images Using CNN %288%29.docx

 Midocean University

Document Details

Submission ID

trn:oid:::2945:409204751

Submission Date

Aug 24, 2026, 4:25 PM GMT+3

Download Date

Aug 24, 2026, 9:33 PM GMT+3

File Name

Pneumonia Classification from Chest X-ray Images Using CNN (8).docx

File Size

2.1 MB

42 Pages

7,959 Words

45,558 Characters

21% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Match Groups

-  **149 Not Cited or Quoted 21%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 12%  Internet sources
- 13%  Publications
- 16%  Submitted works (Student Papers)

Integrity Flags

1 Integrity Flag for Review

-  **Hidden Text**
243 suspect characters on 3 pages
Text is altered to blend into the white background of the document.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 149 Not Cited or Quoted 21%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 12% Internet sources
- 13% Publications
- 16% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.mdpi.com	1%
2	Publication	"Fifth Congress on Intelligent Systems", Springer Science and Business Media LLC...	<1%
3	Submitted works	Ultra Courses on 2026-08-14	<1%
4	Internet	mdpi-res.com	<1%
5	Submitted works	Midocean University on 2025-09-05	<1%
6	Internet	pmc.ncbi.nlm.nih.gov	<1%
7	Publication	Daramola, Toluwani Oluwatobi. "On the Capacity of Larger-Scale MIMO Systems i...	<1%
8	Submitted works	Loyola University, Chicago on 2025-10-19	<1%
9	Internet	dspace.daffodilvarsity.edu.bd:8080	<1%
10	Submitted works	Middle East Technical University on 2026-06-08	<1%

11	Publication	K. V. Sambasivarao, Anasuya Sesha Roopa Devi Bhima. "Artificial Intelligence, Co...	<1%
12	Submitted works	The University of the West of Scotland on 2026-08-17	<1%
13	Submitted works	University of Hertfordshire on 2023-12-04	<1%
14	Internet	eurradioexp.springeropen.com	<1%
15	Internet	par.nsf.gov	<1%
16	Publication	Benzorgat Mustapha, Yatong Zhou, Chunyan Shan, Zhitao Xiao. "Enhanced Pneu...	<1%
17	Publication	Perna Agrawal, Savita Gandhi. "PneumoSense++: An Enhanced Pneumonia Detec...	<1%
18	Submitted works	University of Hull on 2026-07-15	<1%
19	Publication	"Advances in Artificial Intelligence and Smart Environments: Proceedings of ICAIS...	<1%
20	Submitted works	University of Warwick on 2026-08-15	<1%
21	Internet	norma.ncirl.ie	<1%
22	Publication	Dev P. Chakraborty. "Observer Performance Methods for Diagnostic Imaging - Fo...	<1%
23	Internet	infocomp.dcc.ufla.br	<1%
24	Publication	Faisal Alshanketi, Abdulrahman Alharbi, Mathew Kuruvilla, Vahid Mahzoon, Sham...	<1%

25	Publication	Srinivas Sethi, Bibhudatta Sahoo, Deepak Tosh, Suvendra Kumar Jayasingh, Soura...	<1%
26	Internet	public-pages-files-2025.frontiersin.org	<1%
27	Publication	Özen, Sadettin. "Snow Cover Detection Over Forested and Mountainous Regions f...	<1%
28	Submitted works	Berlin School of Business and Innovation on 2026-07-06	<1%
29	Publication	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ...	<1%
30	Internet	repository.aaup.edu	<1%
31	Submitted works	College of Public Health on 2026-03-09	<1%
32	Submitted works	Heriot-Watt University on 2022-08-15	<1%
33	Submitted works	University of Stirling on 2026-04-30	<1%
34	Submitted works	University of Sydney on 2024-07-23	<1%
35	Internet	seclab.bu.edu	<1%
36	Submitted works	Liverpool John Moores University on 2021-02-13	<1%
37	Submitted works	University of Greenwich on 2026-04-20	<1%
38	Publication	Mohanappriya K., Vennila C., Keerthivasan N.. "AgroDualNet: a dual deep learnin...	<1%

39	Publication	Sitanath Biswas, Sayan Chakraborty. "Frontiers in Engineering, Science and Tech...	<1%
40	Internet	arxiv.org	<1%
41	Publication	Ching, Serena Low Woan. "Detection of COVID-19 Pneumonia on Computed Tomo...	<1%
42	Submitted works	Coventry University on 2025-04-06	<1%
43	Submitted works	Kingston University on 2026-01-04	<1%
44	Submitted works	UNICAF on 2025-11-13	<1%
45	Submitted works	University of Queensland on 2025-05-30	<1%
46	Internet	github.com	<1%
47	Internet	ir.vistas.ac.in	<1%
48	Internet	rcastoragev2.blob.core.windows.net	<1%
49	Internet	riunet.upv.es	<1%
50	Submitted works	Al-Hussein Bin Talal University on 2026-06-14	<1%
51	Publication	Alves, Victor Manuel. "Classification of Pulmonary Nodules in 2-[18f]FDG PET/CT I...	<1%
52	Submitted works	Beirut Arab University on 2025-02-28	<1%

53	Submitted works	Brunel University on 2026-03-27	<1%
54	Submitted works	Coventry University on 2025-12-05	<1%
55	Publication	Kazi Abdul Mannan, Arifa Akter Shapla. "Explainable CNN for Automated Lung Op...	<1%
56	Publication	T.R. Mahesh, Muskan Gupta, Abhilasha Thakur, Surbhi Bhatia Khan, Mohammed ...	<1%
57	Submitted works	Technical University of Košice on 2026-06-03	<1%
58	Submitted works	Tilburg University on 2026-07-06	<1%
59	Submitted works	University of Auckland on 2025-10-26	<1%
60	Submitted works	University of Hertfordshire on 2025-09-04	<1%
61	Submitted works	University of Hertfordshire on 2026-05-17	<1%
62	Internet	trepo.tuni.fi	<1%
63	Submitted works	Afe Babalola University on 2026-07-17	<1%
64	Submitted works	Asia Pacific University College of Technology and Innovation (UCTI) on 2026-08-10	<1%
65	Submitted works	Brunel University on 2025-04-19	<1%
66	Submitted works	Higher Education Commission Pakistan on 2023-03-11	<1%

67	Publication	Poonam Nandal, Tapas Kumar, Meeta Singh, Mamta Dahiya. "Progressive Compu...	<1%
68	Submitted works	RDI Distance Learning on 2026-01-21	<1%
69	Submitted works	St Ives High School on 2026-07-13	<1%
70	Submitted works	University of South Florida on 2026-05-08	<1%
71	Publication	Vidhya N, Hariniga A S, Kowshick C, Geetha Shanmugam, Deepa Balusamy, Mythil...	<1%
72	Internet	biblio.ugent.be	<1%
73	Internet	download.bibis.ir	<1%
74	Internet	etd.aau.edu.et	<1%
75	Internet	ijsret.com	<1%
76	Internet	ioinformatic.org	<1%
77	Internet	repository.ju.edu.et	<1%
78	Internet	www.alfarabiuc.edu.iq	<1%
79	Internet	www.fim.uni-passau.de	<1%
80	Publication	"CARS 2019—Computer Assisted Radiology and Surgery Proceedings of the 33rd I...	<1%

81	Submitted works	Berlin School of Business and Innovation on 2024-10-18	<1%
82	Submitted works	Berlin School of Business and Innovation on 2024-10-21	<1%
83	Submitted works	Cavite State University on 2026-06-05	<1%
84	Submitted works	Edith Cowan University on 2025-09-28	<1%
85	Publication	Ferdib-AI Islam, Tonmay Sen, Md. Shamim Parvej Shamim, Partha Protim Gharam...	<1%
86	Publication	Ghulfam Hussain. "Prostate Cancer Detection in Bi-parametric MRI Using Deep Le...	<1%
87	Submitted works	Glasgow Caledonian University on 2026-04-13	<1%
88	Publication	Houmem Slimi, Imen Cherif, Sabeur Abid, Mounir Sayadi. "Enhanced pneumonia ...	<1%
89	Publication	John, Chimundu. "Machine Learning Approach for Predictive Maintenance in an E...	<1%
90	Submitted works	King Mongkut's University of Technology Thonburi on 2026-02-20	<1%
91	Submitted works	Liverpool John Moores University on 2026-05-11	<1%
92	Submitted works	Midocean University on 2025-09-05	<1%
93	Submitted works	Southampton Solent University on 2026-01-09	<1%
94	Submitted works	Sunway Education Group on 2025-06-01	<1%

95	Publication	Tang, Zehua. "A Comparative Study of Deep Learning Models for Bird-Drone Reco...	<1%
96	Submitted works	Technical University of Košice on 2026-06-24	<1%
97	Submitted works	The University of the West of Scotland on 2024-08-07	<1%
98	Submitted works	University of East London on 2024-05-09	<1%
99	Submitted works	University of Exeter on 2026-04-29	<1%
100	Submitted works	University of Hertfordshire on 2026-04-27	<1%
101	Submitted works	University of Keele on 2026-02-15	<1%
102	Publication	V. Sekhar, P. Ajay Kumar Reddy, K. Logesh, T. Raghunadha Reddy, Dara Raju. "Ne...	<1%
103	Submitted works	VIT University on 2026-02-16	<1%
104	Internet	indah.ump.edu.my	<1%
105	Internet	theses.lib.polyu.edu.hk	<1%
106	Internet	www.researchgate.net	<1%
107	Internet	www.researchsquare.com	<1%



88

Pneumonia Classification from Chest X-ray Images Using CNN

Prepared By

Name: Hussam alden Abushalha

ID No.: 253002964

5

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

List of Abbreviations	5
Chapter 1: Introduction	6
1.1 Problem Definition	6
1.2 Thesis Objectives	7
1.3 Thesis Contributions	7
1.4 Thesis Organization	7
Chapter 2: Background and Literature Review	8
2.1 Background	8
2.1.1 Artificial Intelligence (AI)	8
2.1.2 Machine Learning	9
2.1.3 Types of Machine Learning	9
2.1.4 Convolutional Neural Networks	9
2.1.5 DenseNet	Error! Bookmark not defined.
2.1.6 ResNet	10
2.1.7 EfficientNet	10
2.1.8 MobileNet	11
2.1.9 Vision Transformer (ViT)	11
2.1.10 Swin Transformer	11
2.2 Related Work	12
2.3 Comparative Summary	16
Chapter 3: Methodology	19
3.1 Introduction	19
3.2 Proposed Framework	20
3.3 Dataset Description	21

3.4 Dataset Splitting	21
3.5 3.5 Pre-trained CNN Models	22
3.6 Transformer Models	23
3.6.1 Vision Transformer	23
3.6.2 Swin Transformer	23
3.7 Preprocessing and Augmentation	24
3.8 Models and Transfer Learning	24
3.9 Proposed Hybrid Model	24
3.10 Evaluation Metrics	25
Chapter 4: Results and Discussion	27
4.1 Experimental Setup	27
4.2 Results	28
4.2.1 Overall Performance	28
4.2.2 Per-Class Performance	30
4.2.3 Confusion Matrices and ROC Curves	32
4.2.4 Effect of Patient-Level Data Leakage	34
4.2.5 System Implementation	36
4.3 Discussion of Results	37
Chapter 5: Conclusion and Future Work	37
References	39

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

Chapter 1: Introduction

1.1 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 among children younger than five and people over 65 years old. The gap between their outcomes and those in resource-abundant countries is growing every year as access to specialists becomes harder. Although achieving a correct diagnosis during the early stages of the infection is unlikely, it remains a crucial step that can improve the chances 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 lead to different conclusions from various specialists, resulting in delays and the potential for overlooking the illness. In addition, staffing issues make X-ray image analysis even slower. Therefore, there is a need for a solution that works 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 difficult to detect. Therefore, researchers primarily apply deep learning algorithms as their main approach. Researchers primarily apply deep learning algorithms as their main approach. Multiple authors have shown that accuracy in pneumonia detection can exceed 98% (Buriboev et al., 2025; Ahmed et al., 2023). DenseNet, ResNet, and EfficientNet were identified as the neural networks capable of recognizing medical images. The principle of utilizing artificial intelligence technologies to diagnose a condition is obvious; it helps identify important patterns that a specialist may overlook while being exhausted. 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].

1.2 Thesis Objectives

The aim of this study is to construct a deep learning framework to distinguish pneumonia from normal cases in chest X-ray images. Transfer learning is used to explore various CNN architectures to enhance their classification performance. Image augmentation techniques are used to increase variability in the dataset and to generalize to unseen images. The effectiveness of the developed models is evaluated by comparing with the standard performance measures.

1.3 Thesis Contributions

Design and evaluation of deep learning approaches for automatic classification of pneumonia from X-ray images, including transfer learning.

- * Conducted a comparative evaluation of the use of three pre-trained CNNs (MobileNetV2, EfficientNetB0, and ResNet50) and two transformers (Vision Transformer, ViT, and Swin Transformer).
- * Introduced the hybrid model of CNN (EfficientNetB0) & Transformer (Swin Transformer), which can extract the local features as well as global features and achieve better results than individual architectures.
- * The hybrid CNN–Transformer model EfficientNetB0 + Swin Transformer was assessed on a patient-disjoint test set, where it reached an accuracy of 97.01%. The results from the assessment of the hybrid CNN–Transformer model outperformed the three CNNs for pneumonia detection on chest X-ray images, while there was no statistically significant difference to the strongest transformer used individually ($p = 0.524$).
- * Implementation of a prototype web application where the models can be put into practice for the prediction of pneumonia with the help of chest X-rays.

1.4 Thesis Organization

Chapter 1 introduces the research work and outlines the major problems that it seeks to address. The chapter highlights the objectives of the work and presents the overall structure of the thesis.

Chapter 2 details the theoretical basis for pneumonia classification focusing on the leading deep learning models that can analyze images (chest X-ray images).

Chapter 3 outlines the methodology employed throughout the research. the including image preprocessing, the application of specific models and the proposed method of analysis. It details the experimental setup and examines the metrics used to gauge model performance.

Chapter 4 describes the experimental work conducted as part of the research for this thesis and an analysis of the results for the various models developed to address the problem of classifying cases of pneumonia from chest X-ray images.

Chapter 5 outlines the conclusions for the research performed within this thesis and, last but not least, provides an outlook on future work and potential optimizations.

Chapter 2: Background and Literature Review

2.1 Background

2.1.1 Artificial Intelligence

It aims to enable computers to carry out tasks that would normally require a human to use their cognitive abilities. the such as recognizing patterns and objects, making decisions based on provided information, etc. AI tasks require the use of computational and mathematical methods to identify useful connections in large amounts of data. These connections can then be used to make decisions or predictions [23].

AI is a large field of research including many related topics. The topic of ML and DL is to enable systems to learn from data; the topic of CV is to process visual information; and the topic of NLP is to address human language.

2.1.2 Machine Learning

subfield dedicated to learning models from many examples rather than programming them through if/then conditions. The model is given data to train on 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]

2.1.3 Types of Machine Learning

2.1.3.1 Supervised Learning

Effective training of a model requires access to the correct output for every example that the model is trained on and later given in a prediction situation. The model is learning a relationship between the input and output and uses this relationship to make predictions on previously unseen examples [24].

2.1.3.2 Unsupervised Learning

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

2.1.4 Convolutional Neural Networks

are deep learning models to process structured visual data, for example, images. Computer vision tasks are often performed with CNNs because they can learn relevant features for an image classification task from labeled images instead of manually designing feature extraction. The type of CNN architecture includes three categories of layers. The first are the convolutional layers, which scan through the image and identify relevant patterns. The pooling layers then

down sample the resulting feature maps from the previous layer, often reducing their spatial size to decrease the number of parameters in the following fully connected layers. The resulting feature maps are fed through the fully connected layers to perform the necessary classification in the final output. Feature learning occurs at each layer in the network, initially as the first few feature maps pick up simple features of the image such as edges and textures. The deeper layers learn increasingly complex features, making CNNs very suitable for computer vision applications, like analyzing medical images and classifying cases of pneumonia from chest X-rays [16].

2.1.5 ResNet

is a CNN architecture to improve the training of deep neural networks. By using shortcut connections. Information shortcut connections enable information to pass through selected layers in the network, rather than restricting it to a sequential path. This technique allows information to pass through selected layers in the network instead of being restricted to a sequential path. Allowing information to flow through the network in this manner makes the training process more stable as the network depth increases. The information flow through the network Allowing information to pass through the selected layers improves the flow of information in the network. This improvement is achieved by allowing information to pass through the selected layers. ResNet has several residual blocks, each of which learns the input's changes and applies them to the output. As the network goes through the successive residual blocks in the architecture, it is able to learn to represent the image at several different levels of complexity. It can be used for several tasks in the area of computer vision, including medical image analysis and classification [26].

2.1.6 EfficientNet

is a family of models for image classification in the class of CNNs, which tackles the problem of model scaling by considering the three dimensions: depth, width, and input resolution and scales all of them in a compound way. It tries to achieve the best trade-off between accuracy and

resources.

EfficientNet is based on the MBConv block and adds Squeeze-and-Excitation (SE) modules. The architecture has been applied to various image classification tasks, including several **medical imaging classification tasks, such as classifying chest X-rays with pneumonia** [27].

2.1.7 MobileNet

is a lightweight CNN for devices that are not powerful enough for heavy training, e.g., smartphones. This model's lightness comes from using convolutions. Convolution is split into two simpler operations: first performing depthwise convolution on each channel independently. Then applying pointwise convolution to aggregate the channels [26].

2.1.8 Vision Transformer (ViT)

The ViT processes an image first by splitting it up into **patches. Each patch is then turned into a vector, and this vector is combined with a positional embedding**, which is then passed on to the Transformer encoder. In each of the layers of the encoder, there is an MHA component, which captures interactions between patches in the image. This is followed by an MLP component, which further processes the embeddings. **The output of each of the encoder layers is in turn passed to a classification head to return a class that the image corresponds to.** The architecture has proven **to perform well in a variety of tasks in computer vision** and is recently getting a lot of attention for applying it to medical image analysis [20].

2.1.9 Swin Transformer

The main difference between Swin Transformer and ViT is 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. 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].

2.2 Related Work

Singh et al. [1] A pre-trained Vision Transformer, DeiT_Base_Patch16_224, was evaluated against several established CNNs, i.e., AlexNet, VGG, ResNet, DenseNet, and InceptionNet. The models were tested under the same settings. The publicly available Kaggle dataset for the task of chest X-ray pneumonia classification was used in the work. The dataset comprised 5,863 pediatric X-ray images that were collected by the Guangzhou Women and Children's Medical Center. The images were split into 80% for training, 10% for validation, and the remaining 10% for testing purposes. The ViT achieved the highest accuracy of 97.61% and was followed by a sensitivity of 95% and a specificity of 98%. The model also obtained the highest AUC of 0.96 and the highest MCC of approximately 0.94 in this study.

Mustapha et al. [2] present a hybrid deep learning architecture. It consists of a CNN backbone combined with modified Swin Transformer blocks. The architecture is compared to a CNN-only approach as well as to a standalone ViT, using the publicly available Chest X-ray Pneumonia dataset from Kaggle. The proposed hybrid CNN-ViT architecture achieves an accuracy of 98.72% and thus outperforms the CNN as well as the pure ViT approach.

Shomsomi et al. [3] we investigated two DL methods: MobileNetV2, a CNN architecture, and ViT, a CNN architecture. We also tested two different training strategies: the entire network (except for the classification head) is frozen during training (baseline), and some of the inner layers are trained (fine-tuning). The results were obtained using a public chest X-ray pneumonia dataset from Kaggle. A significant performance gap was observed between the two architectures tested using the fine-tuning strategy. achieved an accuracy of 97.63% without fine-tuning, but an accuracy of 97.41% with fine-tuning. This means that there is a negligible impact of fine-tuning on CNN-based architecture. The accuracy of 98.59% and the best F1-score of 0.99 were obtained using the initialized ViT.

Kareem et al. [4] We investigate pneumonia classification in a federated learning setting. Instead of collecting all medical images into a central database and then training a CNN there, in our setting each hospital trains a CNN on the images of its patients in local storage. Only model updates are communicated between hospitals. Using this way of privacy-aware training of a pre-trained

CNN in a federated setting, we get a test accuracy of 93% for ResNet-50, which, however, is below the often-reported numbers for the same architecture in a centralized setting but was achieved without pooling patient data from different hospitals.

Alshanketi et al. [5] The authors evaluate the performance of VGG, ResNet, and ViT models on three different datasets for chest Xray analysis, namely Kaggle Chest X-Ray Pneumonia, BRAX, and Chex pert. The methods to deal with the class imbalance problem are transfer learning, data augmentation, class weighting, and semi-SL. The authors investigate methods to address the class imbalance problem, including transfer learning, data augmentation, class weighting, and semi-supervised learning. The results for ViT reach 97.61% accuracy on the Kaggle dataset. The performance on BRAX and Chex pert is not as strong. The authors found out that great results on one dataset do not automatically translate to great results on other datasets using the same approach.

Slimi et al. [6] The framework includes four key components. In the first place, convolutional layers are employed to capture the spatial information contained in the chest X-ray images. Next, a gated recurrent mechanism can learn the relationships between the features extracted by the previous layers. The framework also includes a spike-based neural processing component that has been included in order to increase the computational efficiency of the framework and to improve its robustness to noise. An attention component is included in order to increase the attention that is focused on the most relevant regions of the images that are needed to classify the images in the best possible way. The author evaluated the framework using a publicly available dataset of chest X-ray images. He reported an accuracy of 99.35% and high values.

Yanar et al. [7] proposed a feature-level ensemble architecture called the PELM . The approach fuses the feature representations of four pretrained networks — InceptionV3, VGG16, ResNet50 and ViT — extracted before their classification layers and combined into a single representation passed to a shared classifier. The ensemble outperformed each of its individual components.

Hofmeister et al. [8] took a more skeptical approach to evaluation. They trained a convolutional model to diagnose pneumonia and assessed it on two prospective cohorts totaling 430 patients, in which the reference diagnosis was established afterwards by a multidisciplinary expert panel using multimodal clinical data rather than by radiologists reading the radiograph alone. Under this stricter reference the model reached an area the ROC curve of 0.74, far below the values typically reported when deep-learning systems are benchmarked against radiologist interpretation, although its accuracy remained comparable to that of the radiologists themselves and significantly higher than that of emergency physicians. The authors concluded that the clinical relevance of such models may have been systematically overestimated by weak reference standards — a conclusion closely related to the evaluation problem examined in Section 4.2.4 of this thesis.

Buriboev et al. [9] applied a modified version of a CLAHE to enhance the input images before classification and applied a CNN for classification, which achieved 98.7% on the testing set for the Chest X-Ray Pneumonia dataset and also showed that improving input quality can improve classification accuracy independently of the architecture used for classification.

Saber et al. [10] in addition to lung segmentation, this work incorporates pneumonia detection from chest X-rays using a multi-scale transformer framework. The framework integrates a lung segmentation using a transformer-enhanced Transnet segmentation stage and pneumonia detection using pretrained ResNet backbones to produce multi-scale features that are then fed into a transformer module. The method achieves a Dice score of 95.68% for the segmentation task and achieves classification accuracy of 93.75% on the dataset and 96.04% on the dataset.

Bundea et al. [11] implemented more DenseNet-based models (DenseNet121, DenseNet169 and DenseNet201) to classify chest X-ray images using a supervised learning approach. The algorithms were tested on The Chest X-Ray pneumonia public dataset from Kaggle and evaluated according to standard parameters such as accuracy, sensitivity, and specificity. In addition, all architecture variants were designed and evaluated in the same way, using 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] This framework is a deep learning model that incorporates EfficientNetB0 and DenseNet121 modules along with several attention components to employ multi-head self-attention and channel-attention feature fusion to extract the most relevant features from input chest X-ray images. The deep learning framework is trained in a supervised learning setting using given pneumonia and normal case images, achieving an accuracy of 95.19%, a precision of 98.38%, a recall of 93.84%, a specificity of 97.43%, and an area under the ROC curve of 0.956.

Usman et al. [13] the use of transfer learning in the classification of images of lung air-containing opacities of varying types is investigated. A CNN is trained from scratch and compared to a pre-trained ResNet-50 model, which was fine-tuned on the RSNA pneumonia dataset. The fine-tuned pre-trained ResNet-50 performed better than the model that was developed from scratch. This further highlights the advantage of using TL, particularly for medical imaging where there are limited numbers of training samples for large deep learning architectures.

Ahmed et al. [14] A deep CNN is presented to classify multiple categories related to pneumonia using chest X-ray images. Unlike approaches that are limited to binary classification between pneumonia and normal (or other diseases), the proposed approach can classify into multiple categories (i.e., different diagnostic classes) within a single CNN framework. The approach has achieved an accuracy of around 98.7% in rating pneumonia-related categories from chest X-ray images.

Al Reshan et al. [15] Pneumonia classification using the MobileNet for detecting pneumonia from given chest x-ray images. This is a binary classification approach in which the given images are classified as either normal or indicative of pneumonia, following image normalization and enhancement during the preprocessing step. Achieved an accuracy of around 95% using the given MobileNet approach for the given binary classification task.

2.3 Comparative Summary

Table 2.1: Summary of previous studies on pneumonia detection

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. (2023)	Pretrained CNNs in federated learning (ResNet-50)	Multi-institutional distributed CXR data	Private	Binary: normal vs. pneumonia	93%	Federated ResNet-50

Study	CNN Models	Dataset(s)	Public / Private	Classification Task	Best Accuracy	Best Performer
Alshanketi et al. (2025)	VGG; ResNet; ViT	Kaggle CXR; BRAX; CheXpert	Public	Binary; class imbalance across 3 datasets	97.61% (ViT on Kaggle)	ViT — on Kaggle dataset
Slimi et al. (2025)	CNN + GRU + spiking layer + attention	Public chest X-ray dataset	Public	Binary: normal vs. pneumonia	99.35%	Attention-guided hybrid
Yanar et al. (2024)	PELM ensemble: InceptionV3 + VGG16 + ResNet50 + ViT	Public chest X-ray benchmarks	Public	Binary: normal vs. pneumonia	Best among all tested	PELM ensemble
Hofmeister et al. (2024)	CNN vs. multimodal expert reference	Two prospective cohorts (n = 430)	Clinical	Binary; prospective clinical validation	AUC 0.74	CNN $\approx$ radiologists
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 transformer	Kermany; Cohen	Public	Binary	96.04% (Cohen); 93.75% (Kermany)	Transformer + ResNet pipeline

Study	CNN Models	Dataset(s)	Public / Private	Classification Task	Best Accuracy	Best Performer
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. (2024)	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. (2024)	CNN; MobileNet	Chest X-ray Pneumonia Dataset	Public	Binary: normal vs. pneumonia	~95%	MobileNet

Chapter 3: Methodology

3.1 Introduction

We describe the methodology used to develop a deep learning framework for detecting pneumonia cases from patients' chest X-ray images. Deep learning architectures for medical image analysis are becoming increasingly specialized to handle specific types of data. CNNs are currently the most popular approach for many image classification tasks; new architectures, especially those based on transformers, have recently received a lot of attention. These architectures can extract several features from an image and use the self-attention mechanism to identify relevant relationships between different parts of the image. ViT and Swin Transformer chest x-ray images are classified as either normal or pneumonia cases using ViT and the Swin Transformer. The ViT and Swin Transformer architectures are used for the binary classification of chest x-ray images, distinguishing between normal and pneumonia cases. The two architectures are able to capture local details as well as the global context of an image. Transfer learning is used, where pre-trained models that have learned to represent images from large datasets are fine-tuned on the smaller medical dataset used in this study. The previously learned representations are fine-tuned for the task of pneumonia classification in this study.

3.2 Proposed Framework

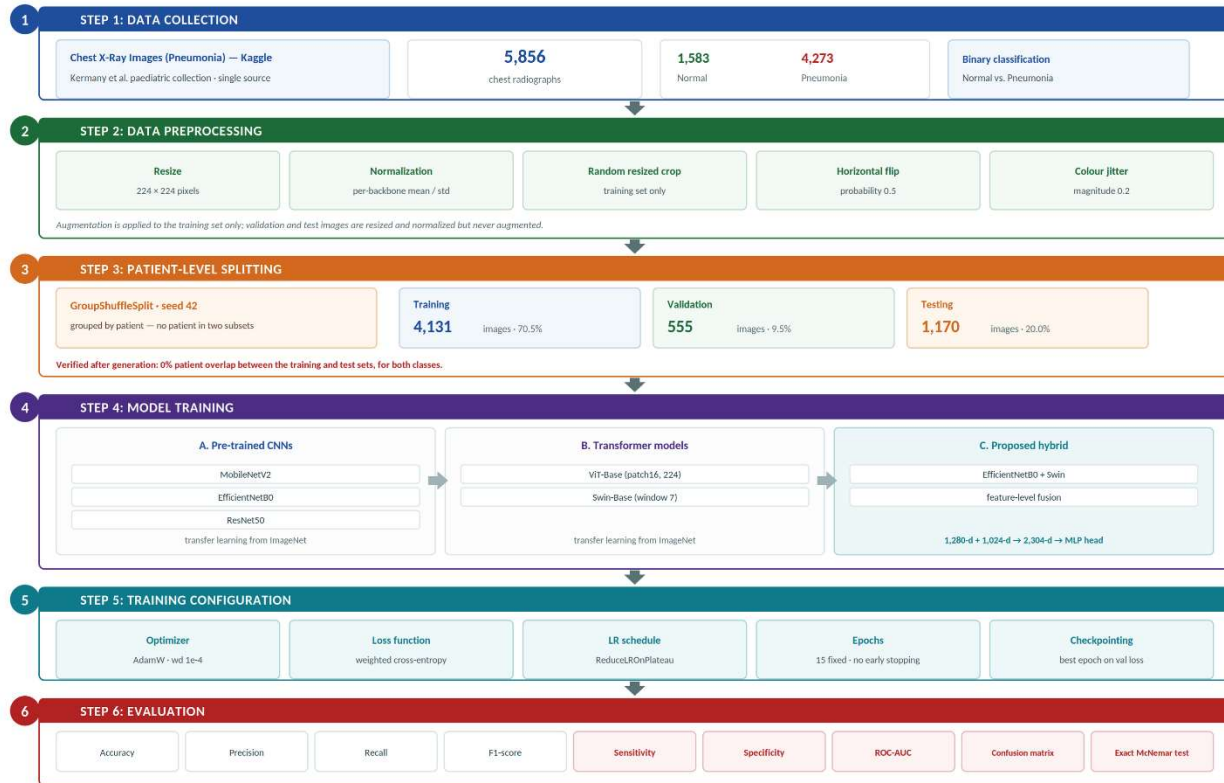


Figure 3.1: Overview of the proposed research framework

3.3 Dataset Description

The Public Chest X-Ray Images (Pneumonia) dataset from Kaggle has been used for this research. The dataset originates from the pediatric chest radiograph collection reported by Kermayn et al. The complete dataset contains 5,856 chest X-ray images, which were divided into two classes: 1,583 normal cases and 4,273 pneumonia cases. Many previous studies for automated pneumonia classification used this large dataset so that different approaches for deep learning could be compared on the same basis. Each image has been resized to 224 * 224 pixels, which is the input size for the used models. The classification task is a binary classification problem, where every image is classified as either a normal or pneumonia case.

3.4 Dataset Splitting

the chest X-ray dataset, we split the data into a training set, validation set, and test set of approximately 70%, 10%, and 20% of the size, respectively. We fixed a random seed (42) so that all 6 models get trained and tested on the same data partitions.

Patient-level data leakage in training and validation sets: images are split by patient instead of individually. The resulting dataset is comprised of 4,273 images of pneumonia from the records of 1,844 unique patients (some of whom have been X-rayed on multiple occasions). We used 'GroupShuffleSplit' to perform the patient-based split and determined patient identifiers from the filenames of the dataset's images. Consequently, all images of a given patient were included in the same subset (i.e., training set, validation set, or test set). We verified the split to confirm that no patients from the test set had appeared in the training set or in the validation set. The implications of this approach for the results of this paper are outlined in Section 4.2.4.

1 Table 3.1 reports the number of images per class in each subset.

70 **Table 3.1: Number of images per class in each data subset**

Dataset	Normal	Pneumonia	Total
Training	1,114	3,017	4,131
Validation	155	400	555
Testing	314	856	1,170
Total	1,583	4,273	5,856

3.5 3.5 Pre-trained CNN Models

3.5.1 MobileNetV2:

A lightweight CNN for efficient image classification.

3.5.2 EfficientNetB0:

27 An efficient CNN architecture that strikes a balance between network depth, width and input size.

3.5.3 ResNet50:

A residual CNN architecture for deep feature learning by using skip connections.

3.6 Transformer Models

3.6.1 Vision Transformer

Processes images as sequences of patches using self-attention.

Processes images as sequences of patches using self-attention.

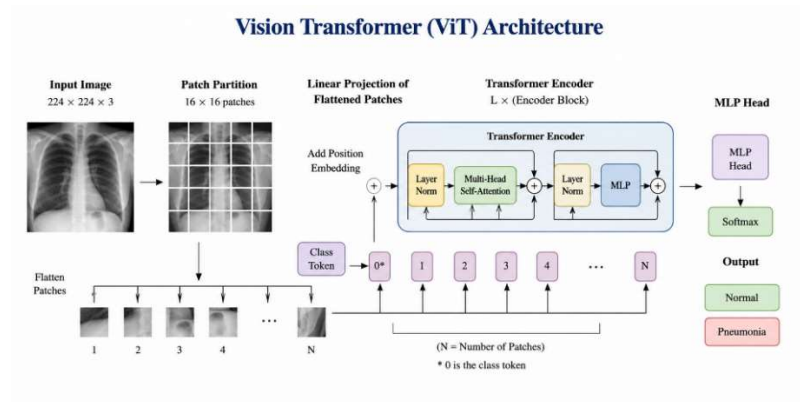


Figure 3.2: Architecture of the Vision Transformer

3.6.2 Swin Transformer

Swin uses attention within shifted local windows to capture hierarchical image features.

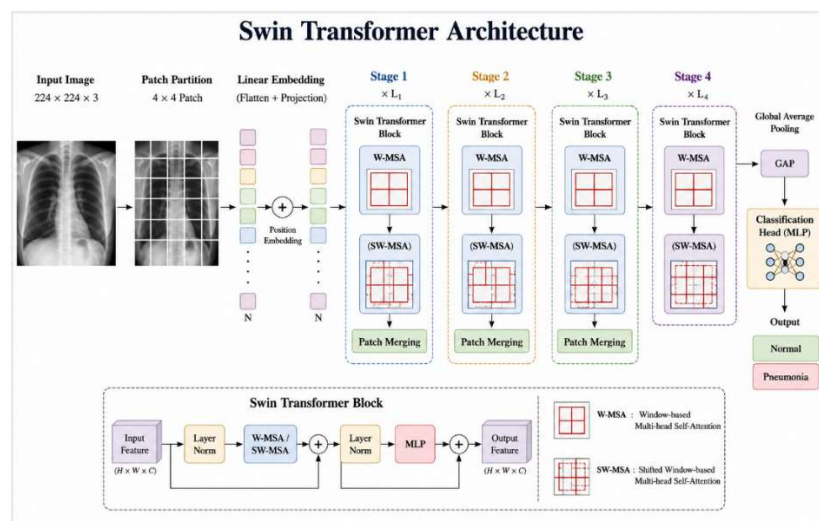


Figure 3.3: Architecture of the Swin Transformer

3.7 Preprocessing and Augmentation

All images of chest X-rays were resized to 224x224 pixels and normalized for use in learning of pre-trained models. Randomized cropped images, horizontally flipped images with a probability of 0.5, and color-jittered images with a probability of 0.2 were used as data augmentation only for the training set. The validation set and test set were not augmented for consistent evaluation. Although a horizontal flip is not useful for identifying the location of certain findings in chest X-ray images, it is used in this study because the goal is to develop a binary pneumonia classification model.

3.8 Models and Transfer Learning

All models were implemented using transfer learning. They were all initialized using the ImageNet pre-trained weights, and the classification layer of each model was replaced with a binary classifier for the normal / pneumonia classes. After that, all models were fine-tuned on the chest X-ray training data for 15 epochs in an end-to-end fashion.

3.9 Proposed Hybrid Model

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 modeling 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.

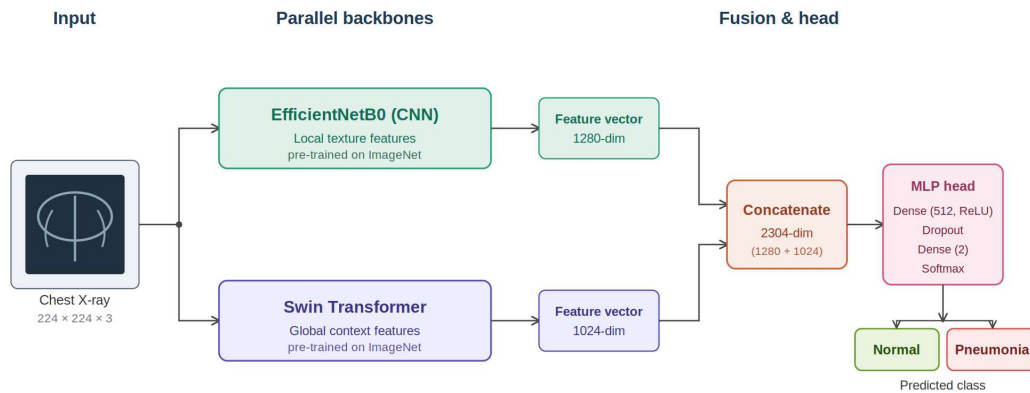


Figure 3.4: Architecture of the proposed CNN-Transformer hybrid model

3.10 Evaluation Metrics

Accuracy

Accuracy is an evaluation metric that measures the proportion of correctly classified cases out of the total number of instances in a dataset. [13]

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

Precision

Precision is a metric that measures the proportion of correctly predicted positive cases out of all cases the model predicts as positive. [14]

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

Recall

Recall is a metric that measures a model's ability to identify all actual positive cases in a dataset correctly. [14]

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

F1 Score

The F1-score is a metric that combines precision and recall into a single

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

Chapter 4: Results and Discussion

4.1 Experimental Setup

The experiment is on a local Windows 11 machine equipped with an Intel Core i9, 32 GB RAM, and a laptop GPU with 8 GB VRAM. A PyTorch implementation using the timm library for models was used with all models initialized with pre-trained weights for ImageNet. We ran all experiments for 15 epochs at a batch size of 32 using AdamW with a weight decay of 1e-4 and a weighted cross-entropy loss to deal with class imbalances. We used a learning rate of 1e-4 for all CNNs and 3e-5 for all transformer models. We used ReduceLROnPlateau with a reduction factor of 0.3 after two epochs of no improvement and mixed precision for GPU training. We fixed a random seed of 42 for all Python, NumPy, and PyTorch code. For evaluation, the best checkpoint as measured by the validation loss for each model was chosen. Since all models went through the full number of scheduled iterations, early stopping was not used, and the procedure described here is best-checkpoint selection. NumPy, Pandas, Matplotlib, scikit-learn, Seaborn, and Stats models were used for data processing, model evaluation, creation of plots, and statistical analysis.

Table 4.1: Experimental configuration

Parameter	Value
Programming Language	Python 3.13
Framework	PyTorch 2.11 (CUDA 12.8) with timm
Models	MobileNetV2, EfficientNetB0, ResNet50, ViT-Base, Swin-Base, Hybrid
Image Size	224 × 224
Batch Size	32
Epochs	15 (fixed); 3 + 8 for the hybrid
Optimizer	AdamW, weight decay 1e-4
Loss Function	Cross-entropy weighted by class frequency (1.854 / 0.685)
Data Split	70/10/20, patient-disjoint (GroupShuffleSplit), seed 42

Early Stopping	Not used; best-epoch weights restored on validation loss
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, Sensitivity, Specificity, ROC-AUC, McNemar test

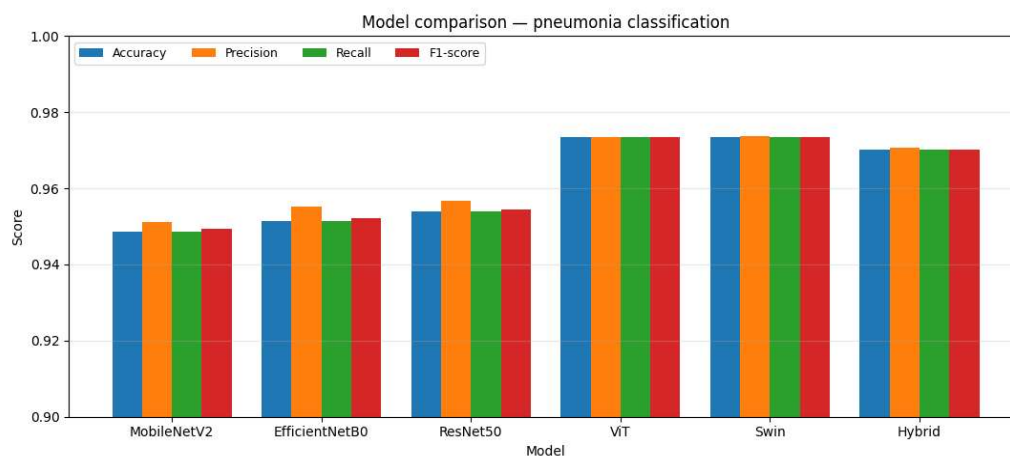
4.2 Results

4.2.1 Overall Performance

Table 4.2 The evaluated models in terms of accuracy, precision, recall, F1-score, and the area under the ROC were measured on the patient-disjoint test set of 1,170 images. The two transformer models tied for the highest accuracy at 97.35%, each classifying 1,139 of the 1,170 test images correctly. The three convolutional networks formed a distinctly lower group: ResNet50 at 95.38%, EfficientNetB0 at 95.13%, and MobileNetV2 at 94.87%. The proposed hybrid reached 97.01%, above every convolutional model but below both transformers. The Swin Transformer obtained the highest AUC of any model at 0.9962, followed by ViT at 0.9948 and the hybrid at 0.9946. These differences are large enough to separate the two architecture families. The Wilson 95% confidence interval for either transformer is 96.26–98.13%, while EfficientNetB0's is 93.74–96.22% and MobileNetV2's is 93.45–96.00%; the transformer interval does not overlap the interval of any convolutional model. Within each family the models cannot be separated: ViT and Swin are tied outright, and the three convolutional networks sit within half a percentage point of one another. The correct reading is therefore that the architecture family matters on this task, while the specific choice within a family does not. To test whether the difference among the proposed hybrid and the strongest single model was statistically meaningful, McNemar's test was applied to the paired predictions of the hybrid and ViT on the same test set, using the exact binomial form because the number of discordant pairs is small. Of the 1,170 test images, the two models agreed on 1,148 and disagreed on 22: the hybrid was uniquely correct on nine and ViT on thirteen. The test returns $p = 0.524$, well above the 0.05 threshold, so the four-image advantage held by ViT is not statistically significant. The hybrid is therefore as effective as the best single model, with no significant advantage.

Table 4.2: Overall performance of the evaluated models

Approach	Model	Accuracy	Precision	Recall	F1-score	AUC
Pre-trained CNN	MobileNetV2	0.9487	0.9511	0.9487	0.9493	0.9866
	EfficientNetB0	0.9513	0.9551	0.9513	0.9521	0.9925
	ResNet50	0.9538	0.9566	0.9538	0.9545	0.9896
Transformer	ViT	0.9735	0.9734	0.9735	0.9735	0.9948
	Swin	0.9735	0.9736	0.9735	0.9735	0.9962
Proposed	Hybrid (EfficientNetB0 + Swin)	0.9701	0.9706	0.9701	0.9702	0.9946

**Figure 4.1: Accuracy, precision, recall and F1-score across all models**

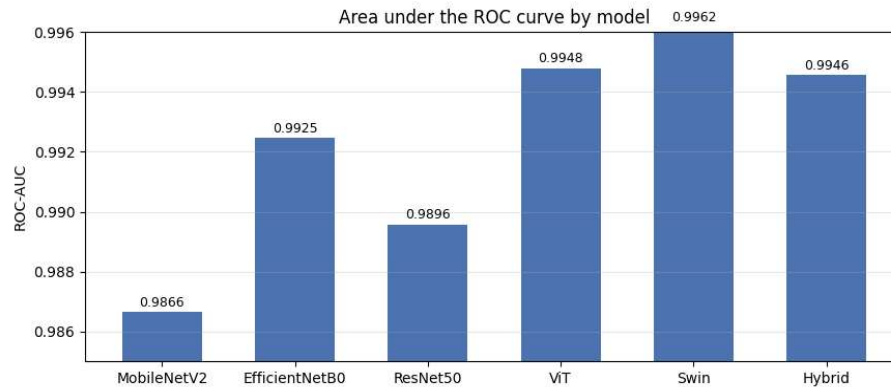


Figure 4.2: Area under the ROC curve by model

4.2.2 Per-Class Performance

Table 4.3 presents a per-class breakdown of precision, recall and F1-score for the Normal and Pneumonia classes, and the asymmetry of the dataset shows clearly. Precision on the Pneumonia class is high for every model, between 0.98 and 0.99, because a prediction of pneumonia is usually correct when pneumonia accounts for three quarters of the images; precision on the Normal class is markedly lower, between 0.87 and 0.95, because the smaller class absorbs proportionally more of the false positives. The two transformers are the most balanced of the six, each reaching 0.95 on all three Normal-class measures against 0.87 to 0.88 for the convolutional networks.

Restating these results as sensitivity and specificity, with the Pneumonia class treated as positive, brings out the difference that matters most clinically. The Vision Transformer missed 14 of the 856 pneumonia cases and the Swin Transformer 17, giving sensitivities of 0.984 and 0.980. The three convolutional networks missed between 42 and 47, giving sensitivities between 0.945 and 0.951. In absolute terms a convolutional model overlooks roughly three times as many pneumonia cases as a transformer on the same images. Comparing MobileNetV2 with ViT, accuracy differs by 2.5 percentage points, but the number of pneumonia cases overlooked falls from 43 to 14 — a difference that would matter far more in a screening setting than the accuracy figures alone suggest. Specificity runs the other way but by a much smaller margin, with EfficientNetB0 highest at 0.968 and ViT and MobileNetV2 lowest at 0.946.

Table 4.3: Per-class performance of the evaluated models

Approach	Model	Class	Precision	Recall	F1-score
Pre-trained CNN	MobileNetV2	Normal	0.87	0.95	0.91
		Pneumonia	0.98	0.95	0.96
	EfficientNetB0	Normal	0.87	0.97	0.91
		Pneumonia	0.99	0.95	0.97
	ResNet50	Normal	0.88	0.96	0.92
		Pneumonia	0.99	0.95	0.97
Transformer	ViT	Normal	0.95	0.95	0.95
		Pneumonia	0.98	0.98	0.98
	Swin	Normal	0.95	0.96	0.95
		Pneumonia	0.98	0.98	0.98
Proposed	Hybrid	Normal	0.93	0.96	0.95
		Pneumonia	0.99	0.97	0.98

4.2.3 Confusion Matrices and ROC Curves

Figure 4.3 compares the predicted labels with the true labels for each model on the held-out test set of 1,170 images. In every case the diagonal entries are substantially larger than the off-diagonal entries. The Vision Transformer correctly classified 297 of 314 normal cases and 842 of 856 pneumonia cases, giving 17 false positives and 14 false negatives. Swin recorded 14 and 17, and the proposed hybrid 12 and 23. The convolutional networks show the same shape with markedly more missed pneumonia: MobileNetV2, 17 and 43; ResNet50, 12 and 42; and EfficientNetB0, 10 and 47—the fewest false positives of any model and the most false negatives. Figure 4.4 shows the corresponding ROC curves, each of which rises steeply towards the top-left corner and remains well above the chance diagonal; the area under the curve exceeds 0.986 for every model and reaches its highest value, 0.996, for Swin.

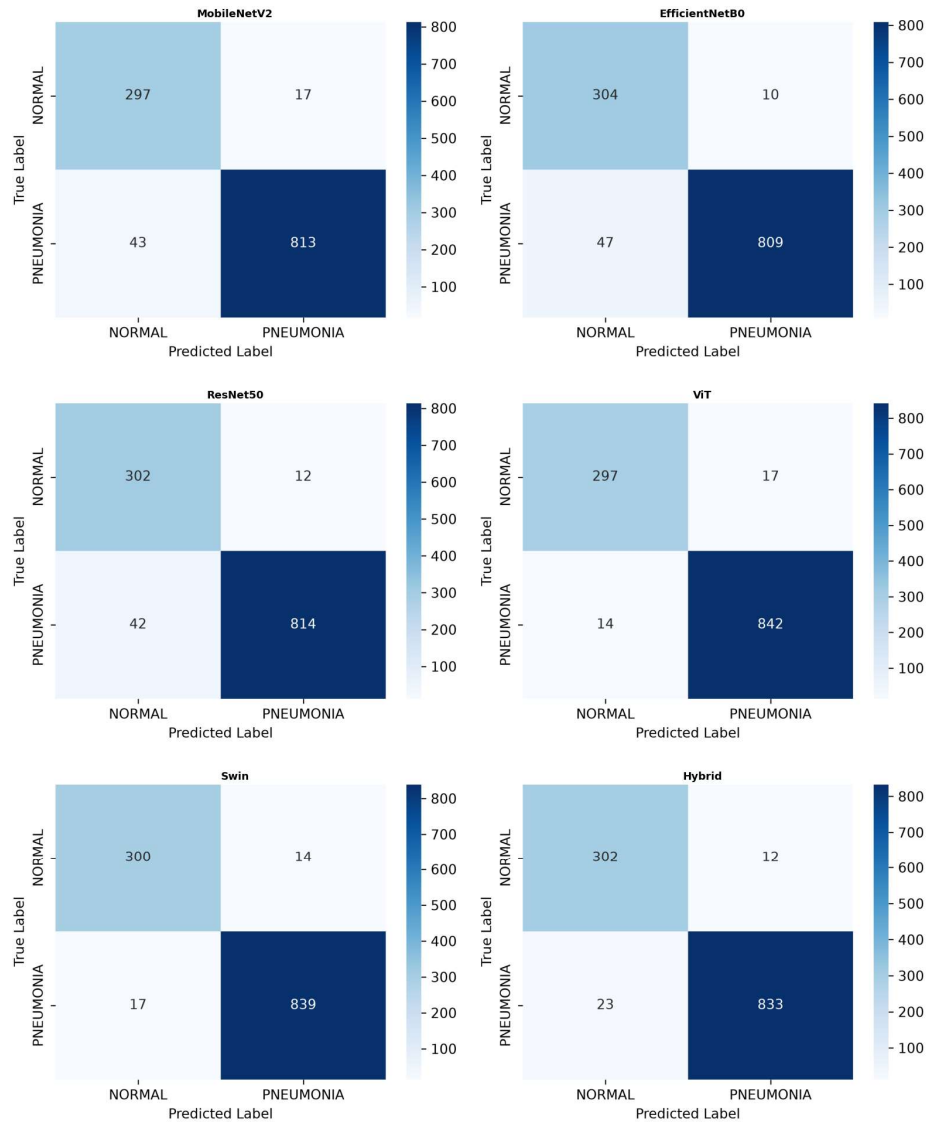


Figure 4.3: Confusion matrices of the evaluated models

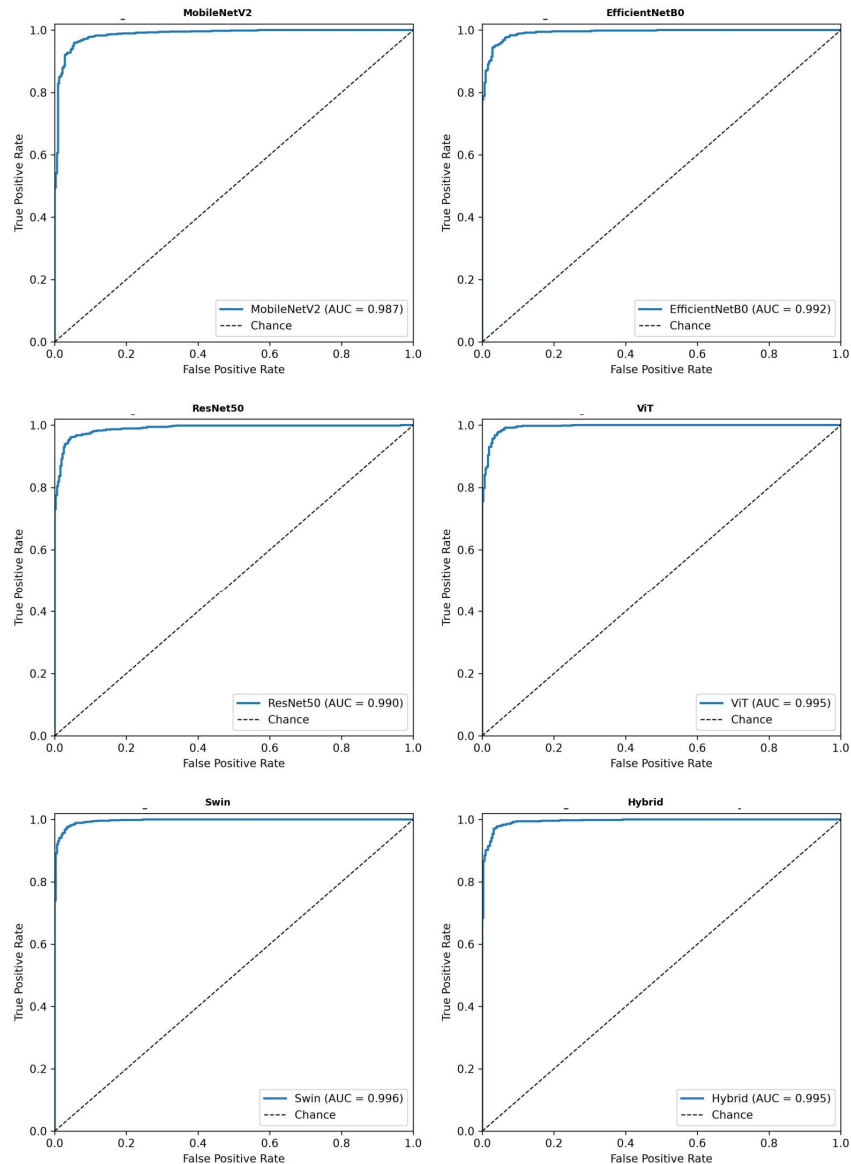


Figure 4.4: ROC curves of the evaluated models

4.2.4 Effect of Patient-Level Data Leakage

The results reported above were obtained on a patient-disjoint split. To quantify what patient overlap costs in measured performance, the entire experiment was also run on a file-level split of the same 5,856 images, in which images were assigned to subsets individually without regard to which patient they came from — the strategy used by much of the published work on this benchmark. Under that split, 74.2% of pneumonia test images came from a patient who also

appeared in the training set; the Normal class, being close to one image per patient, showed only 5.0% overlap.

The effect is systematic and it is not uniform across architectures. All three convolutional networks scored higher under the file-level split — MobileNetV2 by 0.95 points, ResNet50 by 1.04 and EfficientNetB0 by 1.54 — while the two transformers and the hybrid scored slightly lower. The false-negative counts diverge more sharply still: removing the overlap roughly doubled the number of pneumonia cases the convolutional networks missed, from 32–38 to 42–47, while the transformers improved from 24–35 to 14–17. A plausible reading is that convolutional networks, whose receptive fields emphasize local texture, benefit disproportionately from having seen another radiograph of the same chest, whereas attention-based models rely more on global structure that transfers between patients. This is offered as a hypothesis rather than a demonstration: the two runs use different test images as well as different splitting strategies, and each figure comes from a single training run, so the comparison confounds the leakage effect with sampling and run-to-run variation.

Table 4.4: Measured performance under a file-level split and a patient-disjoint split of the same dataset

Model	Accuracy (file-level)	Accuracy (patient-disjoint)	Difference	FN (file-level)	FN (disjoint)
MobileNetV2	95.82%	94.87%	−0.95	38	43
EfficientNetB0	96.67%	95.13%	−1.54	32	47
ResNet50	96.42%	95.38%	−1.04	35	42
ViT	97.18%	97.35%	+0.17	24	14
Swin	96.42%	97.35%	+0.93	35	17
Hybrid	96.59%	97.01%	+0.42	33	23

The practical conclusion is firmer than the mechanistic one. Under the file-level split the six models appeared nearly indistinguishable, with every confidence interval overlapping every other; under the patient-disjoint split a clear gap opens between the transformer and convolutional families. A comparison drawn on the leaky split would therefore have supported the wrong conclusion about which architectures suit this task. This finding echoes that of Hofmeister et al. [8], who showed that benchmarking against a weak reference standard inflates apparent diagnostic performance; here the weak element is the splitting protocol rather than the reference diagnosis, but the lesson is the same. Patient-grouped splitting should be treated as a minimum requirement for work on this benchmark rather than as a refinement.

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 clarifies how the approach could be integrated into a screening workflow. Figure 4.5 shows the interface classifying a pneumonia case. 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 a qualified radiologist should always confirm predictions.

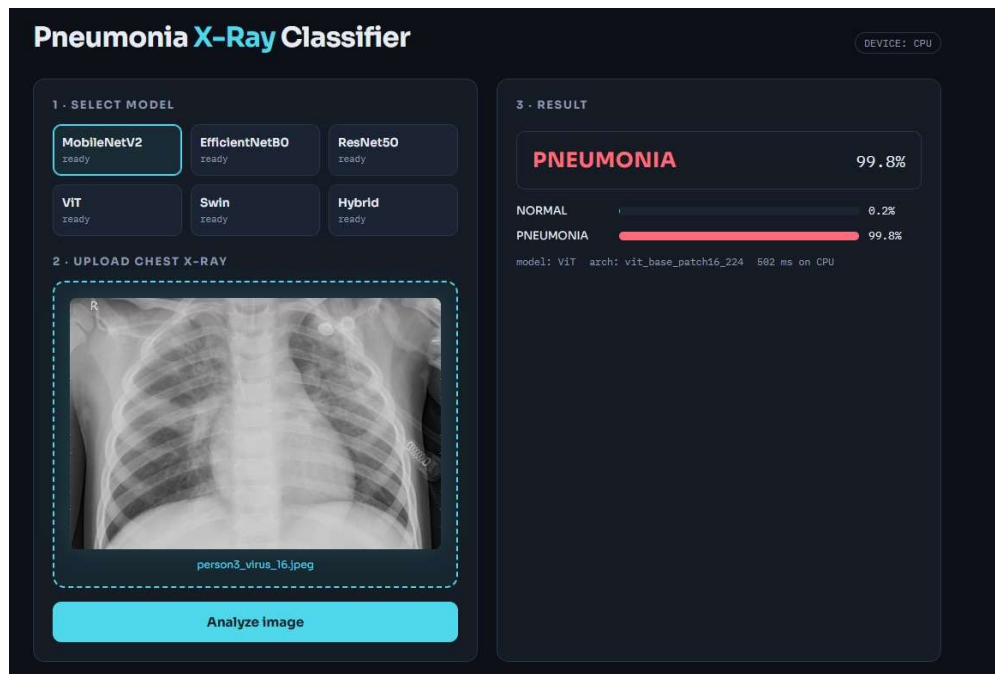


Figure 4.5: The web application classifying a chest X-ray image

4.3 Discussion of Results

Figure 4.3 compares the predicted labels with the true labels for each model on the held-out test set of 1,170 images. In every case the diagonal entries are substantially larger than the off-diagonal entries. The Vision Transformer correctly classified 297 of 314 normal cases and 842 of 856 pneumonia cases, giving 17 false positives and 14 false negatives. Swin recorded 14 and 17, and the proposed hybrid 12 and 23. The convolutional networks show the same shape with markedly more missed pneumonia: MobileNetV2, 17 and 43; ResNet50, 12 and 42; and EfficientNetB0, 10 and 47—the fewest false positives of any model and the most false negatives. Figure 4.4 shows the corresponding ROC curves, each of which rises steeply towards the top-left corner and remains well above the chance diagonal; the area under the curve exceeds 0.986 for every model and reaches its highest value, 0.996, for Swin.

Chapter 5: Conclusion and Future Work

This thesis found that transformer-based models outperform convolutional networks for pneumonia classification on chest radiographs, and that fusing the two families into a hybrid did not improve on the transformer alone. Several directions follow from these findings and from the limitations set out above. The most pressing is external validation: the models should be re-evaluated on larger multi-centre datasets spanning different equipment, patient populations and image quality, and on adult as well as paediatric cohorts, so that generalization can be assessed rather than assumed. The evaluation itself should also be strengthened, with repeated runs across several random seeds to quantify variance, DeLong's test for comparing AUC values, and threshold tuning that reflects the higher cost of a false negative in screening.

Explainability should be treated as a requirement rather than an enhancement. Methods such as Grad-CAM or SHAP would show which regions of an image drive each prediction, allowing clinicians to judge whether the model attends to plausible findings. Prospective comparison against radiologist readings, on cases where ground truth is established independently — as in the protocol used by Hofmeister et al. [8] — would be the natural step after that.

On the modelling side, the hybrid design space deserves fuller exploration. Only one backbone pairing and one fusion strategy were tested here; alternatives such as pairing the convolutional branch with ViT rather than Swin, attention-based rather than concatenation-based fusion, or joint training from scratch would establish whether the ceiling observed in this work is a property of the dataset or of the particular fusion adopted. Extending the task beyond binary classification, to distinguish bacterial from viral pneumonia or to screen for multiple lung pathologies, would broaden the clinical value of the approach. Finally, the web prototype should be developed into a properly engineered decision-support system with authentication, audit logging, de-identification and PACS integration before it is placed in front of any clinical user.

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., Zhou, Y., Shan, C., & Xiao, Z. (2025). Enhanced pneumonia detection in chest X-rays using hybrid convolutional and vision transformer networks. *Current Medical Imaging*, 21, e15734056326685. <https://doi.org/10.2174/0115734056326685250101113959>
- [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, A., Liu, H., & Velisavljevic, V. (2023). A federated learning framework for pneumonia image detection using distributed data. *Healthcare Analytics*, 4, 100204. <https://doi.org/10.1016/j.health.2023.100204>
- [5] Alshanketi, F., Alharbi, A., Kuruvilla, M., Mahzoon, V., Siddiqui, S. T., Rana, N., & Tahir, A. (2025). Pneumonia detection from chest X-ray images using deep learning and transfer learning for imbalanced datasets. *Journal of Imaging Informatics in Medicine*, 38(4), 2021–2040. <https://doi.org/10.1007/s10278-024-01334-0>
- [6] Slimi, H., Balti, A., Abid, S., & Sayadi, M. (2025). Trustworthy pneumonia detection in chest X-ray imaging through attention-guided deep learning. *Scientific Reports*, 15, 40029. <https://doi.org/10.1038/s41598-025-23664-x>
- [7] Yanar, E., Hardalaç, F., & Ayturan, K. (2025). PELM: A deep learning model for early detection of pneumonia in chest radiography. *Applied Sciences*, 15(12), 6487. <https://doi.org/10.3390/app15126487>
- [8] Hofmeister, J., Garin, N., Montet, X., Scheffler, M., Platon, A., Poletti, P.-A., Stirnemann, J., Debray, M.-P., Claessens, Y.-E., Duval, X., & Prendki, V. (2024). Validating the accuracy of deep learning for the diagnosis of pneumonia on chest x-ray against a robust multimodal

reference diagnosis: A post hoc analysis of two prospective studies. *European Radiology Experimental*, 8, 20. <https://doi.org/10.1186/s41747-023-00416-y>

- [9] Buriboev, A. S., Abduvaitov, A., & Jeon, H. S. (2025). Binary classification of pneumonia in chest X-ray images using modified contrast-limited adaptive histogram equalization algorithm. *Sensors*, 25(13), 3976. <https://doi.org/10.3390/s25133976>
- [10] Saber, A., Fateh, A., Parhami, P., Siahkarzadeh, A., Fateh, M., & Ferdowsi, S. (2025). Efficient and accurate pneumonia detection using a novel multi-scale transformer approach. *Sensors*, 25(23), 7233. <https://doi.org/10.3390/s25237233>
- [11] Bunde, M., & Danciu, G. M. (2024). Pneumonia image classification using DenseNet architecture. *Information*, 15(10), 611. <https://doi.org/10.3390/info15100611>
- [12] An, Q., Chen, W., & Shao, W. (2024). A deep convolutional neural network for pneumonia detection in X-ray images with attention ensemble. *Diagnostics*, 14(4), 390. <https://doi.org/10.3390/diagnostics14040390>
- [13] Usman, C., Rehman, S. U., Ali, A., Khan, A. M., & Ahmad, B. (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., Rahman, A., AlGhamdi, F., AlDakheel, S., Hakami, H., AlJumah, A., AlIbrahim, Z., Youldash, M., Alam Khan, M. A., & Basheer Ahmed, M. I. (2023). Joint diagnosis of pneumonia, COVID-19, and tuberculosis from chest X-ray images: A deep learning approach. *Diagnostics*, 13(15), 2562. <https://doi.org/10.3390/diagnostics13152562>
- [15] Al Reshan, M. S., Gill, K. S., Anand, V., Gupta, S., Alshahrani, H. M., AlSulaiman, A. A., & Shaikh, A. (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).
- [28] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., ... Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 172(5), 1122–1131.e9. <https://doi.org/10.1016/j.cell.2018.02.010>

- [29] Mooney, P. (2018). Chest X-ray images (pneumonia) [Data set]. Kaggle.
<https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia>
- [30] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.
<https://doi.org/10.1007/BF02295996>
- [31] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212.
<https://doi.org/10.1080/01621459.1927.10502953>
- [32] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3), 837–845. <https://doi.org/10.2307/2531595>
- [33] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.
<https://doi.org/10.1109/ICCV.2017.74>
- [34] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 4765–4774.
- [35] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510–4520.