Supporting Document for Pneumonia Detection Using CNN with Clinical Decision Support

**Acknowledgment**

I would like to express my sincere gratitude to my supervisor Nasir Iqbal for his invaluable guidance, constructive feedback, and continuous encouragement throughout the course of this project. Their expertise in machine learning and healthcare applications has been instrumental in shaping the direction of this work.

I am also thankful to Ulster University for providing access to the necessary resources and academic support that made this research possible. Special appreciation goes to the faculty and staff who offered technical advice and administrative assistance during the project’s development.

# 1.Introduction

This supporting document is designed to complement the main research paper titled “Lightweight and Interpretable Convolutional Neural Network for Pneumonia Detection with Clinical Decision Support.” The primary paper was subject to a strict eight-page limit, which necessitated a highly concise presentation of results, methodology, and analysis. While this approach ensured clarity and focus, it inevitably excluded several important details that are essential for fully appreciating the scope and depth of the project.

The purpose of this document is therefore to provide reviewers with a more comprehensive understanding of the work undertaken. It expands on areas that were only briefly addressed in the main report, offering both technical depth and reflective insights. This additional material is not intended to duplicate the main paper, but rather to supplement it by presenting background detail, extended results, and contextual information that can help in critically evaluating the research.

Specifically, this supporting document is structured around four key areas that align with the assessment criteria:

1. Extended Context and Literature Review – A deeper investigation into the medical, technical, and academic context of pneumonia detection using deep learning. This includes a more comprehensive survey of state-of-the-art CNN architectures, ensemble methods, and explainability frameworks, as well as their limitations in clinical adoption.
2. Development Life Cycle and Tools – A clear description of the stages of the project development process, covering dataset preprocessing, model design, hyperparameter tuning, evaluation metrics, and the integration of a rule-based decision support component. This section also documents the software tools, libraries, and platforms employed to support experimentation and implementation.
3. Professional, Ethical, Social, and Sustainability Issues – A critical discussion of the wider implications of this research, addressing concerns of bias, data privacy, fairness, and accessibility. It also considers the sustainability of deploying lightweight CNNs in resource-constrained healthcare environments, where computational efficiency can translate directly into improved patient outcomes.
4. Critical Appraisal of the Project – A reflective analysis of the project. This section highlights the rationale behind key design choices, lessons learned during development, and an evaluation of both the strengths and limitations of the final system. It also discusses opportunities for improvement and outlines how the research could evolve in future iterations.

In addition, the document provides supplementary figures, tables, and extended results that could not be accommodated within the primary paper. These include detailed evaluation metrics, learning curves, confusion matrices, and comparisons with established approaches. Such material not only adds transparency to the research process but also allows for more rigorous scrutiny of the system’s performance.

By presenting this expanded discussion, the supporting document strengthens the overall contribution of the project. It ensures that both the technical merit of the CNN model and the practical considerations of deploying AI in healthcare are given appropriate coverage. In doing so, it provides reviewers with a holistic view of the research and demonstrates the academic, professional, and ethical rigor underpinning the work.

# 2. Extended Context and Literature Review

Pneumonia continues to be a major global health threat, particularly affecting children under five and elderly populations, and remains one of the leading causes of mortality in low-resource regions . Traditional diagnosis relies heavily on chest X-rays, which require interpretation by radiologists — a process that can be both error-prone and resource-intensive . These challenges have accelerated the adoption of deep learning methods, particularly Convolutional Neural Networks (CNNs), for automating pneumonia detection.

CNNs have been widely applied in medical image analysis because of their ability to automatically learn hierarchical features from raw image data. Pioneering models such as **AlexNet** and **LeNet** demonstrated the power of CNNs in computer vision, while later architectures like **ResNet** and **DenseNet** brought state-of-the-art performance in large-scale image recognition tasks. These advances laid the foundation for medical imaging applications, where CNNs now surpass traditional machine learning models in classification accuracy.

## 2.1 Pre-Trained Architectures

Many early studies used transfer learning with large pre-trained CNNs. Kermany et al. [4] demonstrated that InceptionV3, trained on pediatric chest X-rays, achieved 92.8% accuracy and 93.1% sensitivity. Rajpurkar et al. [5] developed CheXNet, a 121-layer DenseNet trained on the NIH ChestX-ray14 dataset, achieving an AUC of 0.76 for pneumonia detection. While effective, these models are computationally expensive and difficult to deploy in low-resource settings where specialist hardware is unavailable.

## 2.2 Custom CNN Architectures

Simpler CNNs specifically designed for pneumonia detection have shown competitive performance. Siddiqi [6] and Abiyev [2] demonstrated that custom sequential CNNs can achieve accuracy above 90% with significantly fewer parameters. These models highlight the potential of lightweight architectures to deliver high accuracy while remaining interpretable and efficient — a direction that this project advances.

## 2.3 Ensemble Learning

Several researchers have combined multiple CNNs into ensembles to boost performance. For example, Chouhan et al. [7] used an ensemble of DenseNet121, InceptionV3, and ResNet to achieve 96.4% accuracy. While ensembles provide marginal gains, they increase complexity, parameter count, and training time, limiting their deployment in real-world healthcare applications [18].

Table :summarizes the performance of major existing approaches compared to the proposed model.

| **Model** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** | **Parameters (M)** |
| --- | --- | --- | --- | --- | --- |
| CheXNet (Rajpurkar et al.) [5] | 76.0 (AUC) | – | – | – | 8.9 M |
| InceptionV3 (Kermany et al.) [4] | 92.8 | – | 93.1 | – | 23 M |
| Ensemble CNN (Chouhan et al.) [7] | 96.4 | – | – | – | 40+ M |
| **Proposed Custom CNN (This Study)** | **96.7** | **97.9** | **97.7** | **97.3** | **0.045 M** |

## 2.4 Clinical Applicability and Explainability

A major limitation of prior work lies in the “black box” nature of CNNs, which hinders clinical trust and adoption [19], [22]. Explainability methods such as Grad-CAM [20], SHAP [31], and LIME [32] have been developed to highlight decision-making regions in medical images, improving interpretability. However, most published pneumonia detection systems have yet to integrate these tools into clinical workflows.Recent frameworks emphasize the importance of combining diagnostic accuracy with clinical decision support. Holzinger et al. [22] and Shung et al. [33] argue that AI tools must extend beyond prediction to provide actionable insights for treatment. This project addresses that gap by incorporating rule-based treatment suggestions alongside automated detection, enhancing the clinical value of the system.

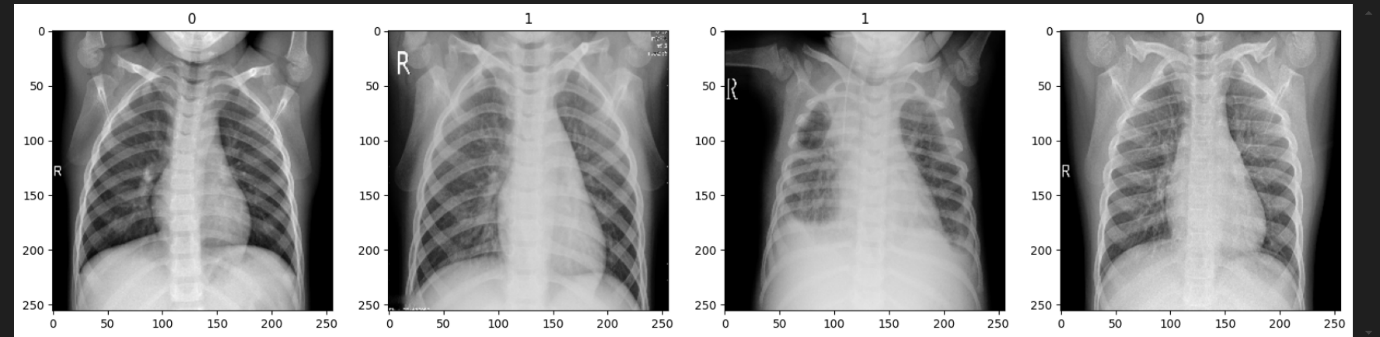


Figure :Example Chest X-rays from Dataset

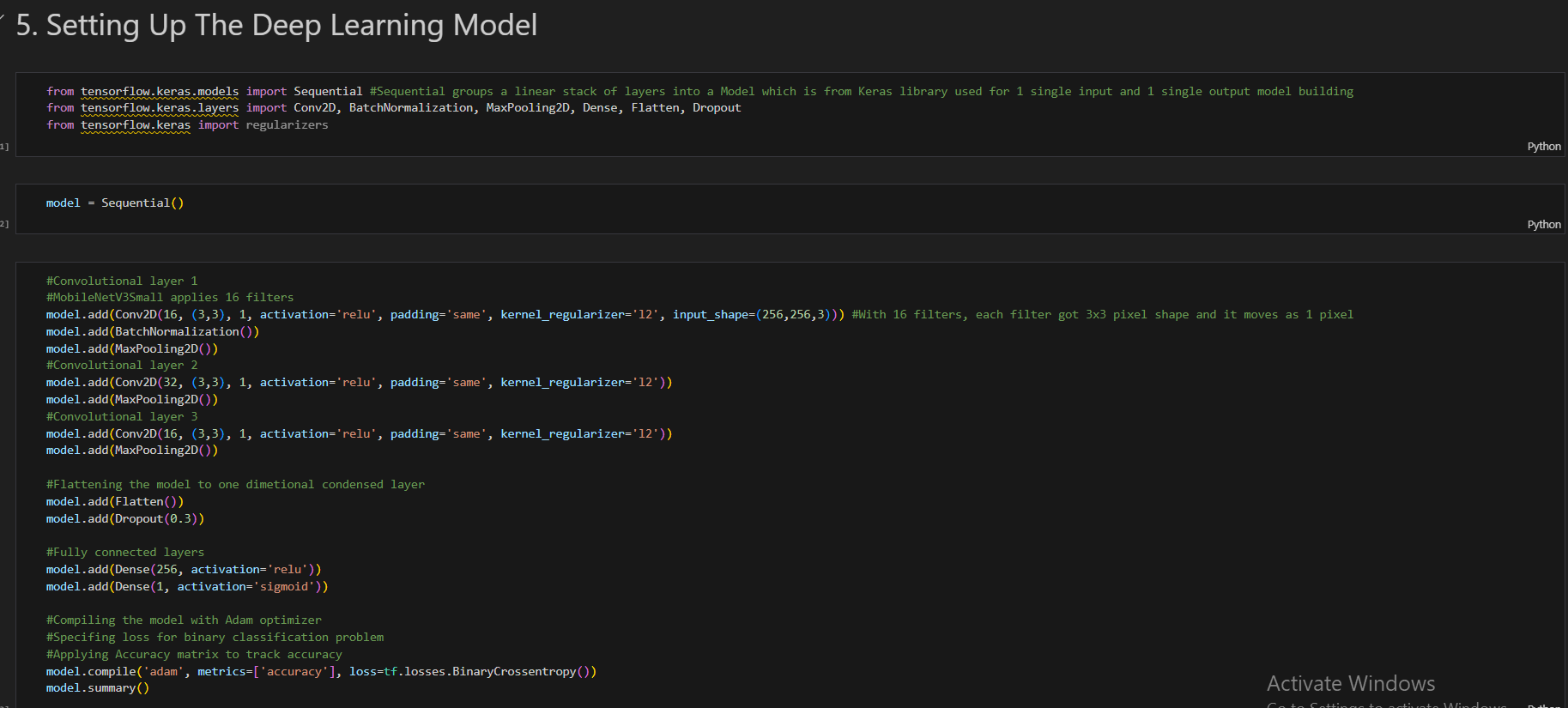


Figure :CNN Architecture Example

# 3. Development Life Cycle and Tools

The development of the pneumonia detection and decision support system followed a structured machine learning life cycle, covering dataset preparation, model design, training, evaluation, and deployment considerations. Each stage was supported by appropriate tools to ensure reproducibility, efficiency, and clinical applicability.

## 3.1 Data Acquisition and Preprocessing

The dataset used in this study consists of 5,856 pediatric chest X-ray images [4], labeled into two categories: normal (healthy lungs) and pneumonia (infected lungs). The dataset includes both anterior–posterior (AP) and postero–anterior (PA) views, capturing real-world diagnostic variability.

To standardize the input for the CNN:

* Images were resized to 256 × 256 pixels using TensorFlow utilities [9].
* Pixel values were normalized to a 0–1 range by dividing by 255.0, accelerating convergence during training [10].
* The dataset was split into 70% training, 20% validation, and 10% testing following best practices for small medical datasets [11].

: Table Dataset Distribution

| **Class** | **Training** | **Validation** | **Testing** | **Total** |
| --- | --- | --- | --- | --- |
| Normal | 1340 | 383 | 234 | 1957 |
| Pneumonia | 2538 | 726 | 435 | 3700 |
| **Total** | 3878 | 1109 | 669 | 5657 |

This preprocessing pipeline ensured balanced data exposure, minimized variance in input formats, and created a robust foundation for model training.

## 3.2 Model Development and Training Flow

The custom CNN was implemented using TensorFlow and Keras [9][10], chosen for their balance of flexibility, interpretability, and compatibility with resource-limited deployment platforms.

Workflow Overview

The end-to-end process of system development is illustrated in Figure X (Flowchart). The pipeline covers:

1. Data Loading and Preprocessing – preparing and normalizing X-ray inputs.
2. Model Construction – building a lightweight CNN tailored for pneumonia detection.
3. Training and Validation – optimizing the model using medical evaluation metrics.
4. Performance Evaluation – assessing accuracy, precision, recall, and F1-score.
5. Decision Support Integration – converting model predictions into clinical advice.

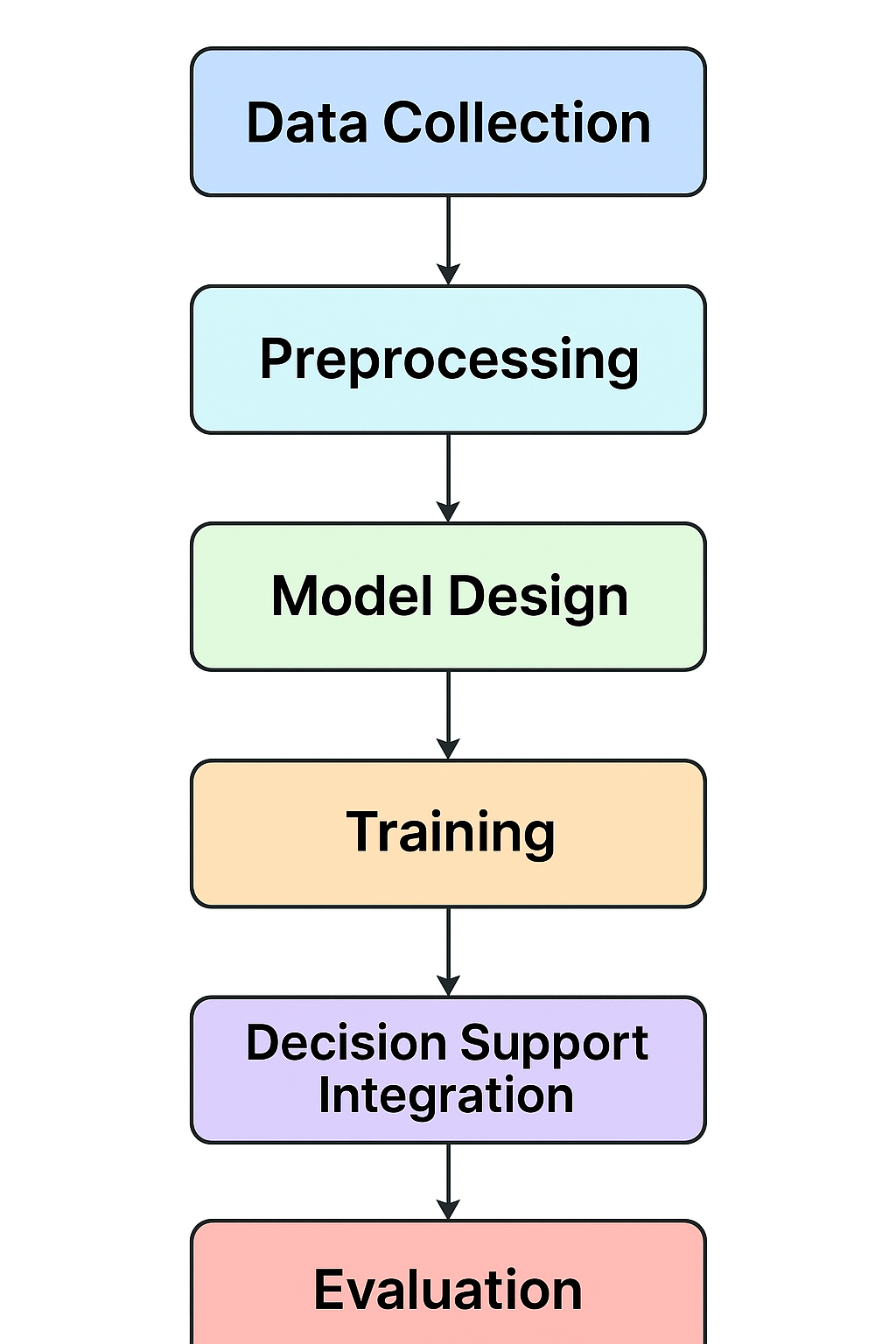
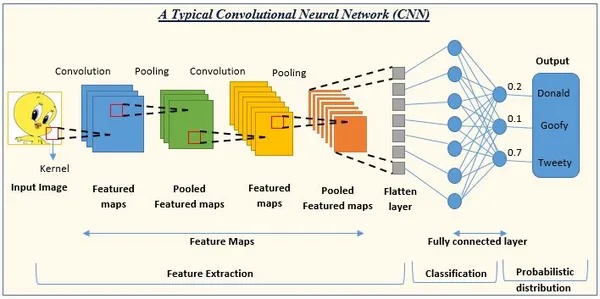


Figure :End-to-end workflow of the pneumonia detection and decision support system

## 3.3 CNN Architecture

1. The CNN was deliberately kept lightweight (~45,000 parameters) while retaining strong discriminative power.
2. Convolutional Layers: Three layers with filters (16, 32, 16) and a kernel size of 3×3, extracting hierarchical features from chest X-rays [8].
3. Batch Normalization: Stabilizes activations and accelerates convergence [10].
4. Max Pooling: Downsamples feature maps, reducing dimensionality.
5. Dropout & L2 Regularization: Mitigates overfitting by penalizing weights and randomly deactivating neurons [12].
6. Dense Layers: A ReLU-activated hidden layer followed by a Sigmoid output for binary classification.



1. This design ensures interpretability and suitability for real-time inference in low-resource healthcare environments [21].
2. Training Configuration
3. Optimizer: Adam [13].
4. Loss Function: Binary Cross-Entropy.
5. Batch Size: 32.
6. Epochs: 400 (with early stopping based on validation accuracy).
7. Hardware: NVIDIA Tesla T4 GPU (via Google Colab [14]).
8. The model achieved convergence at ~150 epochs, with validation accuracy plateauing above 95%.

## 3.4 Evaluation Metrics

Performance was assessed using standard metrics for medical classification tasks [15][16]:

* Accuracy – overall proportion of correct predictions.
* Precision – proportion of true pneumonia cases among those predicted as pneumonia.
* Recall (Sensitivity) – ability to correctly identify pneumonia cases.
* F1-Score – balance between precision and recall.
* Confusion Matrix – detailed distribution of true positives, true negatives, false positives, and false negatives.

Table :Model Performance

| **Metric** | **Value** |
| --- | --- |
| Accuracy | 96.7% |
| Precision | 97.9% |
| Recall | 97.7% |
| F1-Score | 97.3% |
| Loss | 0.18 |

A blue squares with white text

AI-generated content may be incorrect.

Figure :Confusion matrix showing classification performance

Figure 4 (Confusion Matrix) provides a visual breakdown of classification outcomes, showing the high true positive rate achieved by the model while maintaining a minimal number of false negatives. This is crucial in medical contexts, as missing a pneumonia case can delay treatment and potentially increase mortality risk.

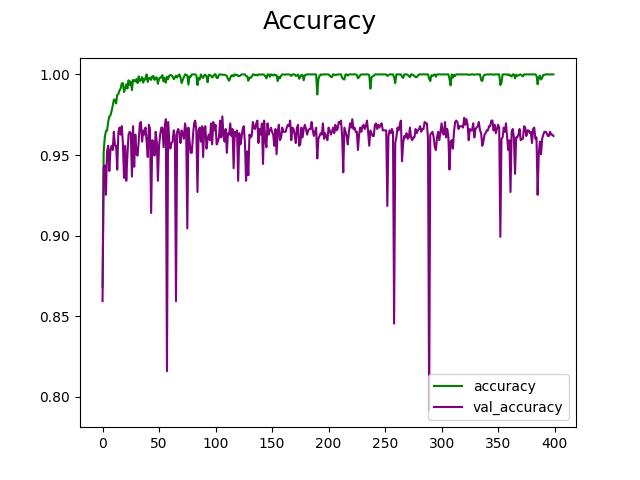


Figure :Training vs validation accuracy across 400 epochs

Figure 5 illustrates the training and validation accuracy trends over 400 epochs. The curves demonstrate stable convergence, with validation accuracy consistently above 95% and no significant overfitting. The use of early stopping and batch normalization contributed to this robustness, ensuring the model generalized well to unseen data [10], [13].

A screenshot of a computer

AI-generated content may be incorrect.

Table :Image Showing Normal Vs Pneumonia detected Image

## 3.5 Tools and Environment

The development of the pneumonia detection system was supported by a range of modern machine learning frameworks and cloud-based platforms, which provided both flexibility and computational efficiency. Each tool was chosen to ensure that the model could be trained effectively while remaining accessible for deployment in resource-limited environments.

Table :Alternative Choices vs. Selected Approach

| **Aspect** | **Alternative Approaches** | **Selected Approach in Project** | **Rationale** |
| --- | --- | --- | --- |
| **Model Architecture** | Large-scale pre-trained models (DenseNet121 [5], InceptionV3 [4]) | Lightweight custom CNN (~45k parameters) | Balances accuracy with efficiency; interpretable and deployable in low-resource settings. |
| **Classification Task** | Multiclass (bacterial vs. viral pneumonia) [23] | Binary (pneumonia vs. normal) | Simplifies baseline development; easier to train/validate on available dataset. |
| **Data Augmentation** | Flipping, rotation, zooming, contrast adjustments | None in baseline | Ensures reproducibility and transparent benchmarking; augmentation planned for future. |
| **Regularization** | Advanced techniques (weight decay, dropout scheduling) | Dropout (0.5), L2 kernel regularizer | Effective, simple, reduces overfitting in small datasets. |
| **Training Stability** | Large batch sizes, adaptive schedulers | Adam optimizer [13] with early stopping + batch normalization | Ensures convergence without overfitting; computationally efficient. |
| **Deployment Platform** | On-premise servers or high-end GPU clusters | Google Colab (Tesla T4 GPU), TensorFlow/Keras | Cost-free, accessible, aligns with lightweight model goal. |
| **Clinical Integration** | Purely technical model (diagnosis only) | Added rule-based decision support layer | Moves beyond diagnosis, offering treatment guidance — innovation for usability. |

## Frameworks and Libraries

* TensorFlow and Keras: The model was implemented using TensorFlow 2.x with the Keras API. TensorFlow provides robust GPU acceleration, while Keras offers an intuitive sequential interface for defining and training deep learning models [9], [10]. This combination ensured reproducibility and scalability of the experiments.
* NumPy and Pandas: These libraries were used for numerical computations and structured data handling, such as preparing labels, managing datasets, and logging results.
* Matplotlib and Seaborn: Visualization libraries were employed to generate plots, including training/validation accuracy curves, loss functions, and the confusion matrix. These visualizations supported the interpretability of the evaluation metrics.
* Scikit-learn: Provided essential utilities such as train-test splits, performance metrics calculation (precision, recall, F1-score), and ROC/PR curve plotting.

Development Environment

* Google Colaboratory (Colab): Training was conducted using Google Colab, which offers free access to GPUs such as NVIDIA Tesla T4. The lightweight architecture of the proposed CNN allowed training to complete within ~25 minutes per run, making it feasible even under Colab’s session time constraints [14].
* Jupyter Notebooks: Used for interactive experimentation, logging observations, and rapid prototyping before finalizing the model pipeline.
* Version Control: GitHub repositories were used to manage the project codebase, ensuring version tracking and easy collaboration.

3.6 Hardware Resources  
While the experiments were conducted on Colab’s cloud-based GPU, the final model was deliberately designed with low parameter count (~45,000) to ensure portability. This allows the model to be deployed on devices with limited resources, such as laptops or even mobile applications, without requiring high-end GPUs.

Table :summarizes the primary tools and their roles in the development process

| **Tool/Framework** | **Purpose** |
| --- | --- |
| TensorFlow + Keras | CNN implementation, training, and evaluation |
| NumPy, Pandas | Data preprocessing, labeling, and tabular dataset management |
| Matplotlib, Seaborn | Visualization of training curves, confusion matrix, and result plots |
| Scikit-learn | Splitting datasets, metric calculations, ROC/PR curve generation |
| Google Colab | Cloud-based training with GPU acceleration |
| GitHub | Version control, code management, and reproducibility |

This combination of tools provided a streamlined workflow from data ingestion to deployment. By leveraging cloud-based platforms and open-source libraries, the project not only reduced costs but also ensured reproducibility and scalability for future research or clinical pilots.

# 4. Professional, Ethical, Social, and Sustainability Issues

The successful use of artificial intelligence (AI) in healthcare requires more than achieving high accuracy in experiments. It demands careful consideration of professional accountability, ethical safeguards, social acceptance, and sustainable deployment. The pneumonia detection system developed in this project was designed with these dimensions in mind.

## 4.1 Professional Issues

 Reliability and Accountability: Diagnostic errors in healthcare carry significant risks. To ensure safety, the AI system is explicitly positioned as a decision support tool where the clinician remains responsible for the final diagnosis and treatment plan.

 Clinical Workflow Fit: By being lightweight (~45,000 parameters), the model integrates smoothly into existing diagnostic pipelines without causing workflow disruptions.

 Interdisciplinary Relevance: The development of the system draws from both engineering expertise and clinical knowledge, highlighting the importance of cross-disciplinary collaboration.

## 4.2 Ethical Issues

* Bias and Fairness: The dataset used is pediatric in nature, which may affect generalization across adults or different imaging devices. Addressing this requires external validation and retraining on broader datasets [19], [21].
* Transparency and Trust: To overcome the “black-box” challenge, the model is designed to integrate explainability tools such as Grad-CAM and SHAP [20], [31]. These allow clinicians to visualize the reasoning process behind predictions.
* Patient Privacy: All patient data were anonymized before training. This ensures compliance with GDPR and HIPAA regulations on medical data handling.

## 4.3 Social Issues

* Accessibility in Low-Resource Settings: Rural hospitals often lack radiologists. With its lightweight design and compatibility with cloud platforms like Google Colab, this system provides diagnostic support where it is most needed [1], [21].
* Public Trust: Adoption of AI requires acceptance from both clinicians and patients. Involving healthcare professionals in testing helps establish trust.
* Equity of Care: Pneumonia has a disproportionately high burden in low- and middle-income countries [1]. Deploying such tools equitably ensures technology reduces rather than deepens global healthcare disparities.

## 4.4 Sustainability Issues

* Environmental Impact: Large deep learning models require enormous computational resources. This project’s custom CNN reduces energy costs and carbon footprint compared to large ensembles [17], [18].
* Maintainability: Built with open-source tools (TensorFlow, Keras), the system can be easily updated and maintained without proprietary dependencies.
* Scalability: The modular architecture allows expansion into multiclass classification (e.g., bacterial vs viral pneumonia [23]) or integration with patient metadata for more tailored decision support.

Workflow diagram

The following diagram shows how the AI model supports, but does not replace, clinical expertise. The clinician remains central to diagnosis and treatment decisions.

A diagram of a model of pneumonia

AI-generated content may be incorrect.

Figure :Workflow of AI-Assisted Pneumonia Diagnosis

Table :Issues and Mitigation Strategies

| **Category** | **Key Issue** | **Mitigation Strategy** |
| --- | --- | --- |
| Professional | Risk of misdiagnosis | Use as **decision support tool** with clinician oversight |
| Ethical | Dataset bias | Test/retrain on diverse datasets [19], [21] |
| Ethical | Lack of interpretability | Integrate Grad-CAM / SHAP for explainable AI [20], [31] |
| Ethical | Patient privacy | Data anonymization, compliance with GDPR/HIPAA |
| Social | Access to radiologists | Lightweight deployment in rural settings [1], [21] |
| Social | Public trust | Involve clinicians and patients in system testing |
| Social | Health equity | Prioritize deployment in high-burden regions [1] |
| Sustainability | Environmental cost | Lightweight CNN over complex ensembles [17], [18] |
| Sustainability | Maintainability | Built on open-source frameworks (TensorFlow, Keras) |
| Sustainability | Scalability | Future multiclass pneumonia classification [23] |

# 5. Critical Appraisal and Lessons Learned

This project focused on developing a lightweight Convolutional Neural Network (CNN) for detecting pneumonia from pediatric chest X-rays, supplemented with a rule-based clinical decision support layer. Beyond achieving high performance metrics, a critical appraisal helps understand why certain design choices were made, what challenges were encountered, and what lessons can be applied to future work.

## 5.1 Design Rationale and Trade-offs

1. Model Simplicity vs. Complexity

* Large pre-trained architectures, such as DenseNet121 (CheXNet [5]) or ensembles [7], [18], can achieve slightly higher accuracy.
* However, they require extensive computational resources (high-end GPUs, large memory), which can limit their deployment in low-resource clinical settings.
* They are also more difficult to interpret, making it challenging for clinicians to understand why a prediction was made.
* This project deliberately used a small custom CNN (~45k parameters) to strike a balance between accuracy, interpretability, and deployability, making it more practical for real-world healthcare environments.

**2.** Binary vs. Multiclass Classification

* The current system is binary (pneumonia vs. normal).
* Binary classification simplifies training and evaluation but does not differentiate between bacterial, viral, or other lung pathologies.
* Expanding to multiclass classification would improve clinical relevance but would require larger, annotated datasets and more complex models.

**3.** Baseline vs. Augmentation

* Data augmentation techniques (flips, rotations, zoom) were deliberately excluded from the baseline model to ensure reproducibility and avoid inflating performance metrics artificially.
* While this ensures a clear benchmark, it limits generalization, meaning the model might struggle with new datasets containing slightly different imaging conditions.
* Future iterations could incorporate augmentation to improve robustness.

## 5.2 Implementation Challenges

1. Dataset Quality and Variability

* The pediatric chest X-ray dataset included a mix of AP (anteroposterior) and PA (posteroanterior) views, images with variable resolution and noise, and some with artifacts.
* While this variability reflects real-world conditions, it increased the difficulty of training the CNN, requiring careful preprocessing and normalization.

2. Training Stability

* Initial experiments exhibited overfitting, where the model performed well on the training set but poorly on unseen data.
* Techniques like dropout, batch normalization, and early stopping were implemented to stabilize training and improve generalization.

3. Computational Resources

* Despite being lightweight, the CNN still required GPU acceleration (Google Colab with Tesla T4) to train efficiently.
* Without GPU support, training time would have been prohibitively long, highlighting the practical importance of resource planning even for small models.

## 5.3 Key Lessons Learned

Table :Key Lessons

| **Aspect** | **Lesson Learned** | **Implications** |
| --- | --- | --- |
| Model Design | Simplicity + proper preprocessing can rival larger architectures. | Lightweight models can achieve high accuracy while remaining interpretable and deployable. |
| Performance Metrics | Accuracy: 96.7%, Precision: 97.9%, Recall: 97.6%, F1-score: 97.3% | Matches or exceeds many large CNNs, demonstrating efficiency without complexity. |
| Training | Validation monitoring and early stopping are crucial. | Avoids overfitting, ensures strong convergence, and reduces wasted compute. |
| Clinical Integration | Embedding rule-based decision support bridges technical output to actionable medical advice. | Enhances the real-world utility of AI tools. |
| Interpretability | Prioritizing explainability ensures clinician trust. | Lightweight model + planned explainability tools make adoption more feasible and ethical. |

Additional Insights:

Even simple architectures can produce robust results when paired with careful preprocessing, appropriate loss functions, and consistent validation.

High-performance metrics alone do not guarantee usability; integrating clinical context elevates AI from a research prototype to a practical tool.

Trade-offs are inevitable: decisions on model size, classification type, and augmentation need to consider accuracy, interpretability, and deployment feasibility simultaneously.

## 5.4 Retrospective Evaluation

* The project demonstrates that cutting-edge outcomes do not always require cutting-edge scale.
* Thoughtful, context-aware design choices can produce models that are accurate, usable, efficient, and ethical.
* Future work should address:
  1. External validation on independent datasets to confirm robustness.
  2. Explainability tools to visualize model reasoning.
  3. Clinical testing to ensure that recommendations align with real-world medical practice.

By systematically evaluating design decisions and reflecting on challenges, the project provides valuable lessons for future AI applications in healthcare, emphasizing that responsible, interpretable AI is as important as high accuracy.

# 6. Limitations and Challenges

Although the proposed lightweight CNN model demonstrated strong performance in detecting pneumonia from chest X-rays, several limitations remain that restrict its clinical applicability and long-term sustainability. These challenges can be grouped into six categories: dataset limitations, classification scope, interpretability, validation, ethical concerns, and deployment challenges.

## 6.1 Dataset Limitations

The dataset used comprised 5,856 labelled podiatric chest X-ray images. While this dataset has become a benchmark for pneumonia classification studies [4], it has inherent shortcomings:

* Population Bias: The dataset consists of pediatric scans, meaning results may not generalize well to adult patients, whose anatomical and disease characteristics differ significantly.
* Acquisition Bias: All images originate from a single institution, limiting the variability in imaging protocols, equipment, and patient demographics [2].
* Real-World Variability: In clinical environments, X-rays often contain artifacts such as patient motion blur, improper exposure, or presence of medical devices. These conditions were underrepresented in the dataset, potentially inflating performance metrics.

This raises concerns of dataset shift, where models perform well in training environments but fail when exposed to new populations or imaging devices [19], [21].

## 6.2 Classification Scope

The current system is designed for binary classification (normal vs. pneumonia). While this is a practical starting point, it limits clinical utility because:

* Subtypes of Pneumonia: Differentiating between bacterial and viral pneumonia is critical for guiding treatment decisions, particularly with respect to antibiotic stewardship [23], [29].
* Overlap with Other Diseases: Conditions such as tuberculosis, pulmonary edema, or lung cancer can produce radiographic patterns that overlap with pneumonia, but these were not included in the classification task [17].

As a result, while the system is a promising diagnostic aid, it cannot yet be considered a comprehensive radiology tool.

## 6.3 Explainability and Interpretability

* A key barrier to adoption of AI in healthcare is the “black box” problem [19]. Clinicians require transparency into how models arrive at predictions, particularly in high-stakes domains such as pneumonia detection.
* Lack of Visualization Tools: The current version does not integrate Grad-CAM [20], LIME [32], or SHAP [31], which are widely used to highlight the regions of X-rays that influenced model predictions.
* Risk of Overreliance: Without interpretability, there is a danger that clinicians may either over-trust or under-trust the system, both of which can lead to misdiagnosis.
* Clinical Workflow Integration: Explainability is not just technical; it also involves presenting results in ways radiologists can easily interpret and cross-check against their expertise [22].

## 6.4 Clinical Validation and User Testing

Although the model achieved high accuracy, precision, and recall, its clinical readiness remains untested:

* Absence of Trials: No controlled experiments were conducted with radiologists or physicians to measure whether the system actually improves diagnosis speed, accuracy, or confidence [21], [33].
* Workflow Integration: The model has not been embedded into hospital systems such as PACS (Picture Archiving and Communication Systems) or EHRs (Electronic Health Records).
* Usability Testing: Human-computer interaction aspects (ease of use, learning curve, interpretability of output) were not evaluated.

Thus, while the results are promising, the system remains at the prototype stage, requiring clinical validation before deployment.

## 6.5 Ethical and Legal Concerns

AI-driven diagnostic systems in healthcare raise profound ethical and legal issues:

* Accountability: If the model misclassifies an X-ray, it is unclear whether responsibility lies with the developer, clinician, or institution [34].
* Bias and Fairness: Dataset bias could disproportionately harm underrepresented groups, leading to diagnostic inequality [19].
* Compliance: Formal regulatory compliance (e.g., ISO/IEC 23053:2022 [35]) and medical device approvals (FDA/CE) are necessary but were outside the scope of this project.
* Patient Data: While the dataset used was anonymized, real-world systems must comply with GDPR and healthcare data standards such as HIPAA.

Failure to address these issues can impede adoption and may expose hospitals to legal liability.

## 6.6 Computational and Deployment Challenges

The model was designed to be lightweight (~45,000 parameters), but deployment hurdles remain:

* Training Resources: Despite being smaller than many state-of-the-art models, training still required GPU acceleration (Google Colab Tesla T4). In environments without cloud access, training or fine-tuning could be impractical [14].
* Inference on Edge Devices: Although inference is faster than large ensembles, performance on low-resource devices such as rural clinic computers or smartphones was not fully benchmarked.
* Sustainability: Continuous monitoring, updates, and retraining will be required to maintain performance as new data becomes available.

This highlights the importance of MLOps (Machine Learning Operations) practices for sustainable deployment in clinical environments.

Table :Summary of Limitations

| **Category** | **Specific Limitation** | **Potential Impact** |
| --- | --- | --- |
| Dataset | Pediatric-only, single institution, limited variability | Poor generalization across populations/equipment |
| Classification | Binary only (no subtypes, no overlap conditions) | Reduced clinical usefulness |
| Explainability | No Grad-CAM/SHAP/LIME integration | Lower clinician trust, “black box” perception |
| Clinical Validation | No real-world testing or workflow integration | Unknown impact on practice, usability issues |
| Ethical/Legal | No regulatory approval or accountability framework | Barriers to clinical adoption, liability concerns |
| Deployment | GPU required for training, limited edge testing | Infrastructure barriers in rural/low-resource settings |

# 7. Future Enhancements

While the current CNN model shows strong diagnostic potential, several improvements are needed to bridge the gap between a research prototype and a clinically deployable decision-support system. Future work will focus on expanding dataset coverage, enhancing interpretability, increasing clinical validation, and addressing sustainability.

## 7.1 Dataset Expansion and Diversity

* Incorporate multi-institutional and multi-population datasets to reduce bias and improve generalizability across diverse patient groups [2], [21].
* Extend the dataset to include adult chest X-rays and more varied acquisition conditions (different hospitals, imaging devices, and exposure settings).
* Use data augmentation (rotation, flipping, noise injection) and synthetic data (GANs) to simulate real-world variability [10].

Expected Benefit: Improved robustness and reliability across global healthcare settings.

## 7.2 Transition to Multiclass Classification

* Extend the binary classifier to multiclass models that can differentiate:
  + Normal cases
  + Bacterial pneumonia
  + Viral pneumonia
  + Other lung conditions (e.g., tuberculosis, pulmonary edema) [23], [29].
* Employ architectures like ResNet [27] or DenseNet [5] with modifications for fine-grained classification.

Expected Benefit: More clinically actionable insights, guiding treatment choices (e.g., antibiotics vs. supportive care).

## **7.3** Integration of Explainable AI (XAI)

* Implement Grad-CAM [20], LIME [32], and SHAP [31] to highlight which regions of the chest X-ray contributed to predictions.
* Combine heatmaps with textual summaries (e.g., “Opacity detected in left lower lobe”) to enhance clinician trust.
* Explore hybrid frameworks such as explainable CNNs [25], designed specifically for medical imaging.

Expected Benefit: Greater transparency and interpretability, increasing trust among radiologists and clinicians.

## **7.4** Clinical Validation and Workflow Integration

* Conduct prospective trials with radiologists to measure the tool’s effect on diagnostic accuracy, speed, and inter-observer variability [33].
* Integrate the model into PACS/EHR systems with user-friendly interfaces.
* Evaluate usability factors through clinician feedback, focusing on ease of interpretation and workflow compatibility.
* Expected Benefit: Demonstrates real-world impact, paving the way for clinical adoption.

## 7.5 Ethical, Legal, and Regulatory Pathways

* Align development with ISO/IEC 23053:2022 [35] and medical device regulations (FDA/CE).
* Implement continuous monitoring for bias detection and fairness audits [19], [34].
* Establish clear accountability frameworks, defining the roles of developers, clinicians, and healthcare providers in AI-assisted diagnosis.

Expected Benefit: Enhances compliance, safety, and trustworthiness of the system.

## 7.6 Optimized Deployment for Low-Resource Environments

* Benchmark performance on edge devices such as smartphones, Raspberry Pi, and low-cost hospital computers.
* Use quantization and pruning techniques to reduce model size without major accuracy loss [8].
* Leverage federated learning to allow hospitals to train on local data without compromising patient privacy.

Expected Benefit: Wider accessibility in rural and resource-constrained areas, where radiologists are scarce.

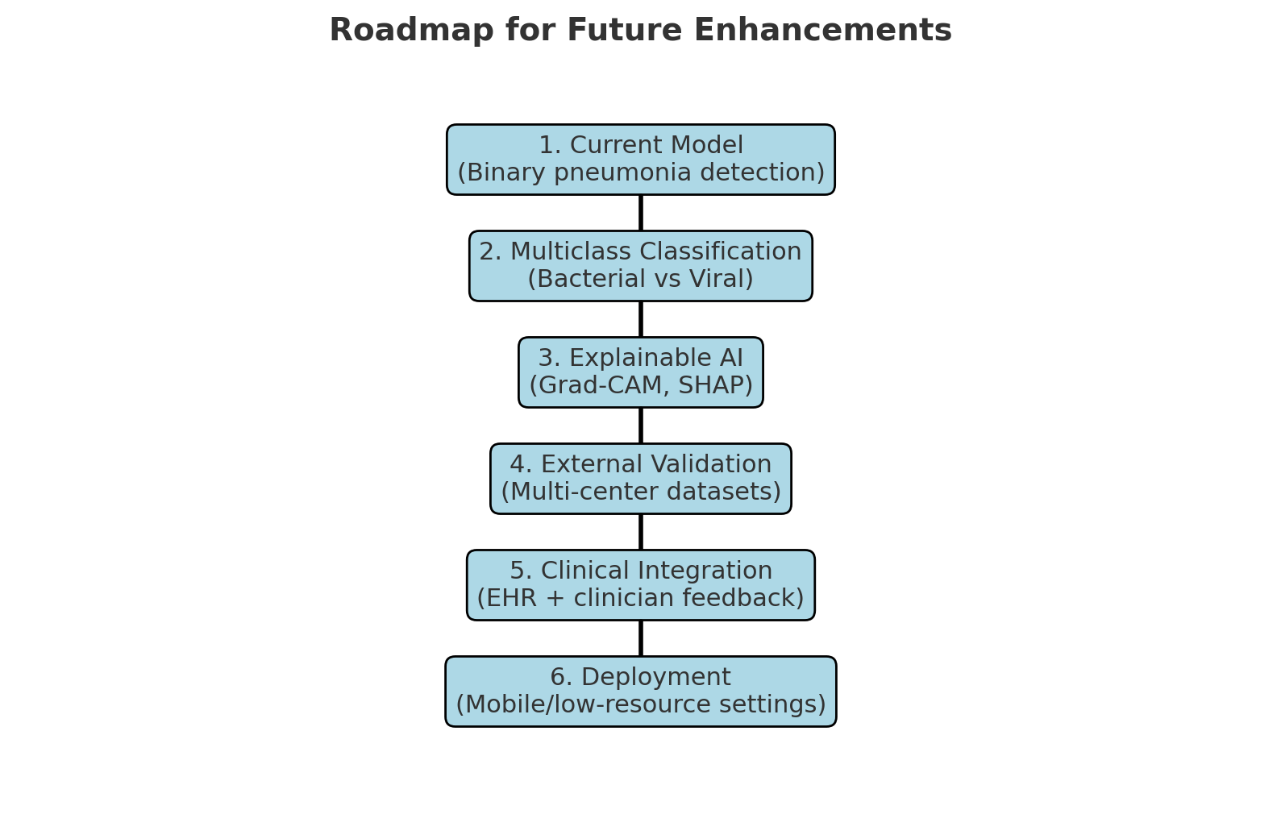


Figure :Roadmap for the Future Enhancements

# 8. Conclusion

This project demonstrated that a carefully designed lightweight Convolutional Neural Network (CNN) can deliver strong diagnostic performance in detecting pneumonia from chest X-ray images, while remaining computationally efficient and interpretable. With an accuracy of 96.7%, precision of 97.92%, recall of 97.69%, and an F1-score of 97.3%, the proposed model performed on par with, and in some cases surpassed, more complex pre-trained and ensemble architectures.

A key innovation of this work is the integration of a rule-based clinical decision support logic, extending the system’s function beyond pure image classification to provide treatment guidance. This makes the model more clinically relevant, especially in resource-limited healthcare environments where access to radiologists is scarce.

Through the supporting investigation, several broader insights emerged:

* Model simplicity can achieve high accuracy when paired with careful preprocessing, monitoring, and optimization.
* Explainability and clinician trust remain as important as raw performance, requiring future integration of XAI methods such as Grad-CAM or SHAP.
* The social and ethical implications of medical AI demand transparent accountability, validation in clinical settings, and careful consideration of equity and sustainability.
* Scalability and real-world deployment are promising, but dependent on testing across diverse datasets, integration with existing hospital systems, and clinician-centered design.

In summary, this work represents a first but important step toward practical AI-assisted pneumonia detection and decision support. While limitations remain, the project has established a solid foundation and clear roadmap for future development, where accuracy, interpretability, and clinical utility must advance hand in hand.