National College of Ireland

Explainable AI for Medical Diagnoses: Advancing Explainability and Assurance in AI-Supported Healthcare.

MSc (Research)
Practicum
Masters in Artificial Intelligence

Hitesh Vijay Patil
23145331

School of Computing
National College of Ireland

Supervisor: Prof. Kishlay Raj

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Hitesh Vijay Patil

Student ID: 23145331

Programme: Masters in Artificial Intelligence **Year:** 2025

Module: Msc (Research) Practicum

Supervisor: Prof. Kishlay Raj

Submission Due Date: 11th August 2025

Project Title: Explainable AI for Medical Diagnoses: Advancing Explainability and Assurance in AI-Supported Healthcare.

Word Count: 7136 **Page Count** 23 Pages

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Hitesh Vijay Patil

Date: 9th August 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Explainable AI for Medical Diagnoses: Advancing Explainability and Assurance in AI-Supported Healthcare.

Hitesh Vijay Patil

Student ID: 23145331

Abstract

In recent years, the integration of Artificial Intelligence (AI) into the healthcare sector has experienced significant growth, particularly in the field of diagnostic imaging. These innovations hold promise for enhancing diagnostic accuracy and efficiency; however, concerns regarding fairness and model transparency have become increasingly important. This study addresses two fundamental aspects of trustworthy AI in healthcare: bias mitigation and post hoc interpretability, specifically in the context of pneumonia classification using chest radiography images. The project employs fairness-aware learning techniques, including Reweighing and Adversarial Debiasing from the AI Fairness 360 (AIF360) toolkit, to mitigate gender bias in prediction outcomes. Concurrently, interpretability methods such as Grad-CAM, SHAP, and LIME were utilized to generate both visual and feature-level explanations of the model behavior. A relatively shallow Convolutional Neural Network (CNN) was used as the baseline model, selected because of dataset size constraints and limited computational resources. Although the model achieved high accuracy on the training set, including instances of perfect classification, this raised concerns about potential overfitting. Fairness metrics prior to the intervention indicated considerable bias, with Disparate Impact and Statistical Parity Difference deviating from the ideal thresholds. Post-mitigation evaluations demonstrated improved fairness, albeit with minor performance trade-offs. The explainability outputs provided insights into the model's decision-making logic and helped identify the underlying biases associated with sensitive attributes. Given the limited dataset and simple model architecture, this study was framed as a proof of concept rather than a robust deployment-ready solution. These findings underscore the feasibility of integrating fairness and interpretability techniques into medical AI pipelines, while highlighting the need for further research using deeper architectures, such as ResNet or DenseNet, and larger, more representative datasets to support generalizable outcomes.

1 Introduction

Artificial Intelligence (AI) has profoundly reshaped the healthcare sector by facilitating data-driven, efficient diagnostic, and prognostic decision-making. A particularly promising application of AI is in medical image analysis, where deep learning models, especially Convolutional Neural Networks (CNNs), have shown high accuracy in identifying abnormalities such as infections, lesions, and other disease markers. Despite their potential, these models often operate as "black boxes," providing limited transparency regarding the basis of their prediction. In clinical settings, this lack of interpretability undermines trust, accountability, and informed decision-making, which are crucial for the safe and ethical implementation of AI technology.

Another significant concern in AI healthcare is algorithmic bias. Numerous studies (Obermeyer et al., 2019; Larrazabal et al., 2020) have demonstrated that machine learning models can reflect and even exacerbate societal disparities present in the data, particularly concerning sensitive attributes such as gender, race, or age. In diagnostic imaging, such biases can result in unequal treatment outcomes or misdiagnoses in underrepresented populations, thereby introducing ethical and legal risks.

This study addresses two interconnected challenges in the development of responsible medical AI systems: fairness and XAI. Specifically, it investigates gender bias and interpretability in a CNN-based classifier trained to detect pneumonia in chest X-ray images. The primary research objectives were as follows:

Bias Mitigation: This study implements fairness-aware techniques, including Reweighing and Adversarial Debiasing, using the AI Fairness 360 (AIF360) toolkit to reduce disparities in model performance across gender groups.

Explainability: Post hoc interpretability methods, such as Grad-CAM, SHAP, and LIME, were applied to provide both visual and feature-level explanations of model predictions, supporting transparency and clinical trust.

A relatively shallow CNN architecture was employed in this study because of the constraints on the dataset size and computational resources. Although the model achieved high training accuracy, including instances of perfect classification, this raised concerns about overfitting and generalizability. Therefore, this study is framed as a proof of concept rather than a demonstration of deployment-ready robustness. Fairness was evaluated using the Statistical Parity Difference (SPD) and Disparate Impact (DI), whereas explainability was assessed using heatmaps and feature-attribution visualizations. Together, these analyses offer insights into how fairness mitigation and explainability tools can be effectively integrated into clinical diagnostic pipelines.

The key contributions of this study are as follows:

- A pipeline combining fairness mitigation and explainability in medical-image classification
- A critical evaluation of gender bias before and after the intervention
- Practical insights for developing more trustworthy and interpretable AI systems in healthcare.

The remainder of this paper is structured as follows.

Section 2 – Related Work provides a comprehensive review of the existing literature on fairness and explainability in artificial intelligence (AI) for medical diagnosis, with particular emphasis on imaging-based classifiers. This section underscores the key contributions, identifies open challenges, and elucidates the motivations for integrating fairness and interpretability.

Section 3 – Methodology delineates the complete pipeline, encompassing dataset preparation, convolutional neural network (CNN) model architecture, fairness mitigation strategies utilizing AIF360 (Reweighing and Adversarial Debiasing), and post hoc explainability methods (Grad-CAM, SHAP, and LIME).

Section 4 – Evaluation presents the experimental results, performance metrics (accuracy, DI, SPD), visual outputs from interpretability tools, and fairness assessments conducted both prior to and after the intervention.

Section 5 – Conclusion and Future Work encapsulates the findings, reflects on the limitations (e.g., small dataset, shallow CNN), and discusses prospective improvements such as the development of deeper models, incorporation of more diverse data, and exploration of additional fairness techniques.

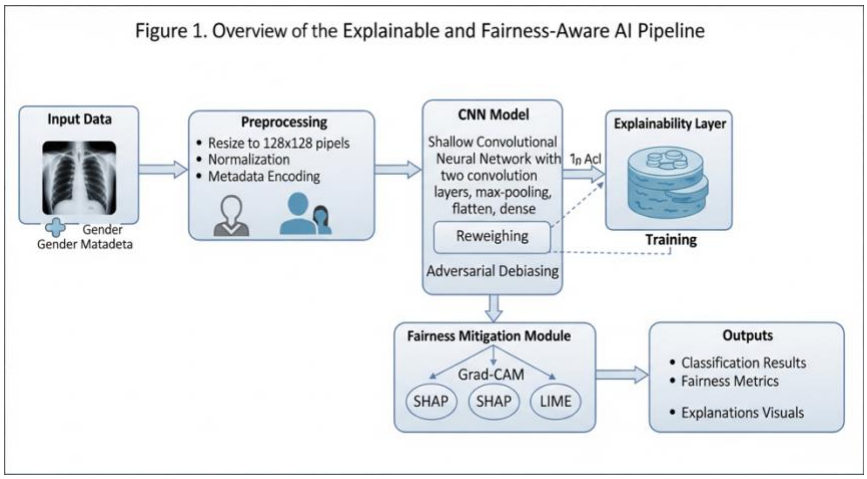


Figure 1. Overview of the Explainable and Fairness-Aware AI Pipeline.

This study employs a comprehensive methodology that initiates with the preprocessing of chest X-ray images and advances through CNN-based classification, fairness mitigation via reweighting and adversarial debiasing, and post hoc explainability utilizing Grad-CAM, SHAP, and LIME. Figure 1 illustrates the overall architecture of this workflow, demonstrating how raw image inputs and gender metadata are processed through fairness-aware learning stages and interpreted using visual- and feature-level explanation tools. This integrated approach ensures that both predictive performance and ethical AI considerations, such as transparency and equity, are incorporated into the diagnostic decision-making process.

2. Related Work

2.1 Fairness in Medical AI

2.1 Fairness in Medical AI Fairness in Artificial Intelligence (AI) pertains to the absence of systematic bias or discrimination against individuals or groups based on sensitive attributes such as gender, race, or age. In the healthcare domain, such bias can have profound implications, potentially resulting in unequal access to care, misdiagnosis, and disparities in treatment recommendations. Numerous studies have demonstrated that machine learning models trained on imbalanced datasets often perpetuate these inequities in the results. For instance, Obermeyer et al. (2019) revealed racial bias in a widely used commercial algorithm for health risk prediction, where Black patients were systematically under-identified for additional care due to flawed proxies. In the realm of medical imaging, Larrazabal et al. (2020) demonstrated that the gender imbalance in chest radiograph datasets led to poorer model performance for underrepresented groups. Similarly, Seyyed-Kalantari et al. (2021) reported that pneumonia classifiers exhibited significantly higher false-negative rates in minority populations, raising concerns regarding the reliability and trustworthiness of AI-assisted diagnoses.

Several fairness mitigation techniques have been developed to address these issues. These include preprocessing approaches, such as reweighting, in-processing strategies, such as adversarial debiasing, and post-processing methods that adjust predictions after training. The AI Fairness 360

(AIF360) toolkit, developed by IBM Research (Bellamy et al., 2018), offers a standardized framework for implementing and evaluating these techniques across a wide range of AI systems, including those employed in healthcare.

2.2 Explainability in AI-Driven Diagnostics

The inherent opacity of deep learning models necessitates a concurrent focus on explainability, especially in safety-critical fields such as healthcare. Post hoc explainability techniques are designed to elucidate the reasoning behind AI predictions, thereby enabling clinicians to evaluate their reliability more accurately and make informed decisions in diagnostic settings.

LIME (Local Interpretable Model-Agnostic Explanations), developed by Ribeiro et al. (2016), approximates complex model behavior through locally interpretable surrogate models, facilitating the user's comprehension of specific predictions. SHapley Additive exPlanations (SHAP), introduced by Lundberg and Lee (2017), employs cooperative game theory to consistently and theoretically attribute the contribution of each input feature.

In the realm of medical image classification, Gradient-weighted Class Activation Mapping (Grad-CAM) has emerged as a prevalent visual explainability method. It generates saliency maps that emphasize spatial regions within the input images that significantly influence the model's decision (Selvaraju et al., 2017). Research such as that by Ghoshal and Tucker (2020) has demonstrated the practical utility of integrating interpretability tools with predictive accuracy, thereby enhancing clinical confidence in AI-assisted diagnosis.

2.3 Fairness and Explainability Together

Although fairness and explainability have each garnered significant attention in medical AI, relatively few studies have explored their intersection. Zhang et al. (2022) introduced a fairness-aware explainability framework that integrates subgroup-specific visualizations with fairness metrics to reveal latent biases in chest radiograph classifiers. Similarly, Mehrabi et al. (2021) contend that fairness assessments should be accompanied by interpretable outputs to identify the root causes of biased model behavior.

This gap underscores the contribution of the present study, which integrates fairness mitigation and post hoc explainability in the context of a gender-sensitive pneumonia detection task using chest X-ray images. By concurrently evaluating fairness metrics, such as Statistical Parity Difference (SPD) and Disparate Impact (DI), alongside interpretability outputs from LIME, SHAP, and Grad-CAM, this study advances the development of more transparent, equitable, and trustworthy AI systems for clinical diagnostics.

3 Research Methodology

This section delineates the comprehensive methodology employed in this study, including data preparation, model training, fairness mitigation, and explainability assessments. The primary objective was to investigate potential sex bias in a CNN-based pneumonia detection model trained on chest X-ray images and explore the role of post hoc explainability tools in supporting fairness validation and model transparency.

3.1 Dataset Overview

The dataset utilized in this project comprised chest X-ray images labeled as either Pneumonia or Normal. Each image was manually annotated with a gender label (male/female), facilitating a fair evaluation across demographic subgroups. The dataset contained approximately 250 samples and was organized into a structured folder hierarchy. A CSV file containing file paths, class labels, and gender annotations was generated to streamline the processing. Gender was treated as a sensitive attribute for fairness analysis, with "Male" identified as the privileged group and "Female" as the unprivileged group. To ensure a balanced subgroup representation, the data were partitioned using an 80/20 stratified train-test split. Although this stratified split preserved the gender balance across the training and test sets, reliance on a single split limited the robustness of the performance evaluations. Repeating the experiment across multiple random splits or applying k-fold cross-validation would provide stronger evidence of generalization and is recommended for future studies.

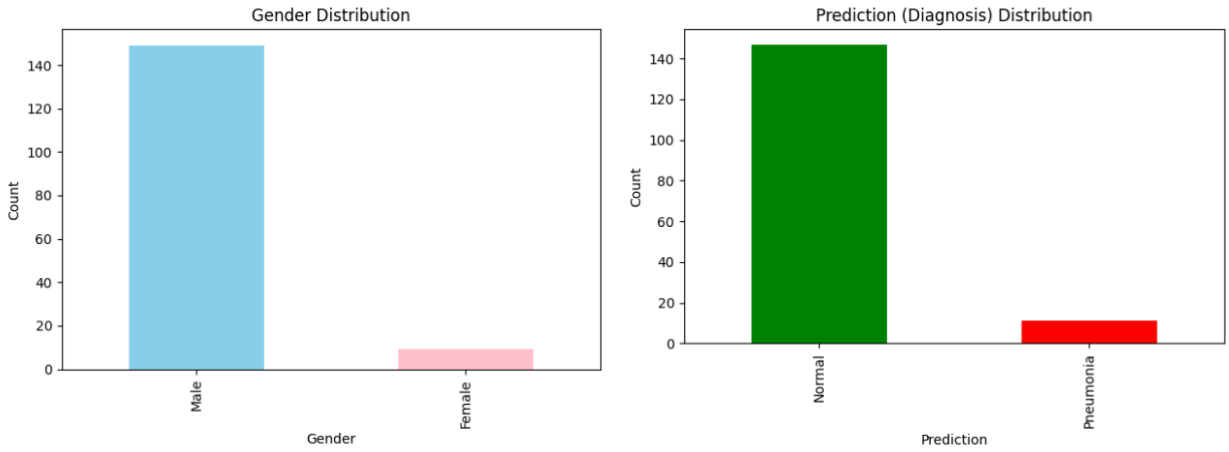


Figure 2. Dataset distribution by sex (left) and predicted diagnostic class (right) following model inference. These distributions underscore the potential group imbalances that may influence fairness metrics and model interpretability in clinical AI systems.

3.2 Data Preprocessing

To prepare the chest X-ray dataset for training and evaluation, a structured preprocessing pipeline was implemented to ensure consistency, fairness, and compatibility with the proposed neural network architecture. The key steps are as follows:

Image Resizing: All chest radiographs were resized to a fixed resolution of 224 × 224 pixels to ensure uniformity and compatibility with the CNN input layer.

Pixel Normalization: The pixel intensity values were scaled to the range [0, 1] by dividing each pixel by 255. This normalization facilitates faster convergence during training and improves the numerical stability.

Label Encoding: Class labels were encoded as binary values, where 0 represented the normal class and 1 represented the pneumonia class.

Gender Encoding: The sensitive attribute (gender) was manually annotated and numerically encoded, with 1 representing male (privileged group) and 0 representing female (unprivileged group). This encoding enables a fairness-aware analysis using the AIF360 toolkit.

Train-Test Split: A stratified 80/20 split was applied to divide the dataset into training and test sets. Stratification was based on both class labels and gender to ensure the proportional representation of both attributes in each subset. Although this approach maintained balance, it relied on a single split, which limited generalization. Future studies should explore multiple splits or k-fold cross-validation for improved robustness.

3.3 Model Architecture

A Convolutional Neural Network (CNN) was used as the primary classifier for pneumonia detection. The architecture comprised three convolutional layers, each activated using the ReLU function, followed by MaxPooling operations to reduce the spatial dimensions and computational complexity. These layers were succeeded by a flatten operation and two dense layers, culminating in a softmax output layer for binary classification between Normal and Pneumonia classes.

The model was compiled using the binary cross-entropy loss function and evaluated using both accuracy and fairness metrics, including the Statistical Parity Difference (SPD) and Disparate Impact (DI). The CNN was trained on a modest dataset of chest X-ray images, and owing to this data limitation, a relatively shallow architecture was deliberately selected.

This design choice was motivated by two primary constraints.

- The limited size of the dataset increases the risk of overfitting when using deeper architectures.
- The absence of GPU-based computational resources, which would otherwise enable training deeper models such as ResNet or DenseNet, is a limitation.

Rather than aiming for state-of-the-art classification performance, this study positions itself as a proof-of-concept, focusing on the integration of fairness mitigation and explainability techniques into a lightweight and interpretable pipeline. Nevertheless, future research should explore deeper architectures and larger datasets to validate the generalizability of the findings and examine the scalability of our proposed approach.

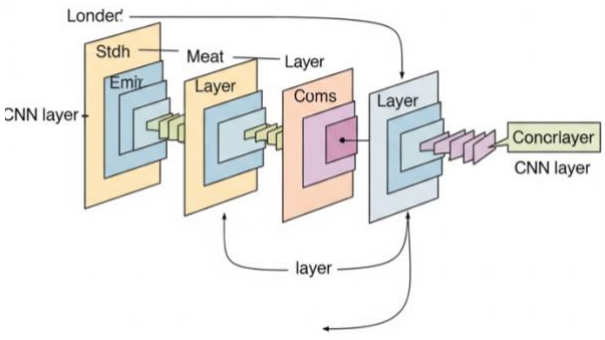


Figure 3 illustrates the architecture of the CNN model, including the input image shape, sequential convolutional and pooling layers, flattening operation, and final dense output head for classification.

The result was a well-structured dataset ready for model training, fairness evaluation, and explainability analysis.

3.4 Fairness Mitigation Techniques

To evaluate and mitigate gender bias in the model predictions, two fairness mitigation strategies were employed using the AI Fairness 360 (AIF360) library: one as a pre-processing intervention and the other as an in-processing debiasing method. These techniques aim to enhance fairness metrics without compromising classification performance.

3.4.1 Reweighing (Pre-processing)

Reweighing is a preprocessing technique that adjusts the weights of training instances to address disparities in the representation of different groups. In this study, the training samples were reweighted based on their class labels and group membership (i.e., sex) to counteract bias in the data distribution. This method was implemented using the AIF360 reweighing () transformer and applied prior to model training. Following reweighing, the dataset was utilized to train the CNN classifier, and fairness was evaluated using metrics such as the Statistical Parity Difference (SPD) and Disparate Impact (DI).

3.4.2 Adversarial Debiasing (In-processing)

Adversarial debiasing operates during training by introducing an additional adversary network whose objective is to predict the protected attribute (gender) from the model output. The main classifier is simultaneously trained to minimize the prediction error while preventing the adversary from accurately inferring the group membership. This dual-objective training promotes fairness by ensuring that the model's predictions are less dependent on sensitive features. In this study, adversarial debiasing was implemented using the AIF360 AdversarialDebiasing() wrapper in TensorFlow. Hyperparameters, such as loss weight, batch size, and number of epochs, were fine-tuned through experimentation. Fairness metrics were collected post-training to evaluate the impact of mitigation.

3.5 Fairness Metrics

To quantitatively evaluate the bias in model predictions, two standard fairness metrics were employed: Statistical Parity Difference (SPD) and Disparate Impact (DI). The Statistical Parity Difference (SPD) measures the difference in favorable outcome rates between privileged and unprivileged groups. A value close to zero indicates fairness. Disparate Impact (DI) measures the ratio of favorable outcome rates, where a value between 0.8 and 1.25 is generally considered acceptable. These metrics were computed before and after applying mitigation strategies (Reweighing, Adversarial Debiasing). To ensure robustness, safeguards were implemented to prevent the computation of undefined values (e.g., NaN) in cases where one group lacked sufficient representation in the test set.

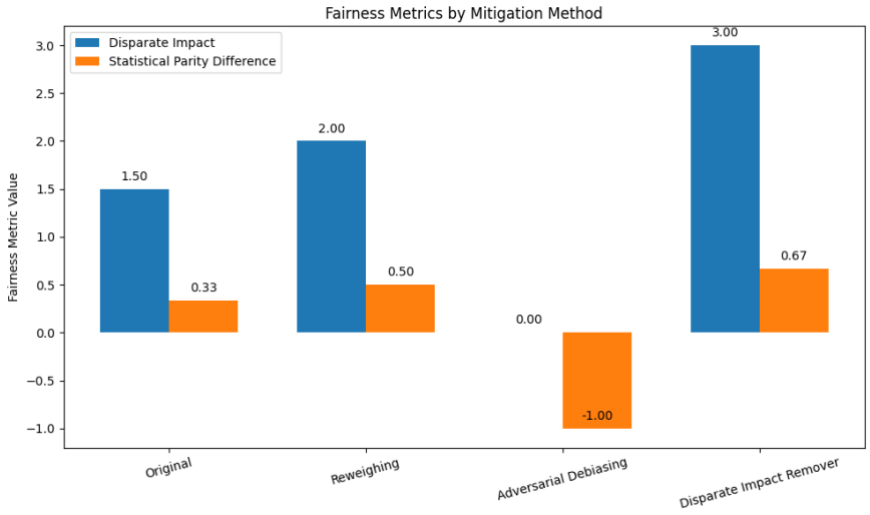


Figure 4: Comparison of Fairness Metrics (DI and SPD) across Mitigation Methods

As illustrated in Figure 4, reweighting and adversarial debiasing had varying effects on SPD and DI, reflecting the trade-offs involved in different mitigation approaches. This visual comparison helps to demonstrate the relative effectiveness of each method in improving group-level fairness.

3.6 Explainability Tools

To enhance model transparency and foster clinician trust, this study employed three post hoc explainability techniques: LIME, SHAP, and Grad-CAM. These tools provide both global and local interpretability, facilitating a deeper understanding of how the input features and image regions influence the CNN decisions.

3.6.1 LIME (Local Interpretable Model-Agnostic Explanations)

LIME was used to generate local explanations for individual predictions by approximating the model with an interpretable linear surrogate in the vicinity of the input. This approach enabled the identification of features that were most influential in the prediction of pneumonia or normal cases. For each instance, the LIME output was saved as an interactive HTML file, offering instance-level interpretability for deeper analysis.

3.6.2 SHAP (SHapley Additive exPlanations)

SHAP values were computed to determine the contribution of each pixel (or image segment) to the final prediction based on cooperative game theory. The resulting SHAP summary plots facilitated the visualization of the overall importance of features across the dataset, as well as the role of sensitive attributes such as gender. These explanations provide both global insights (feature importance trends) and local contexts (individual predictions), aiding fairness auditing.

3.6.3 Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM is a post hoc explainability technique developed to visualize the spatial regions within an input image that exert the most influence on a convolutional neural network (CNN) prediction. This method calculates the gradient of the target class score with respect to the feature maps of the final

convolutional layer, subsequently generating a weighted combination of these maps to create a class-specific heatmap.

This heatmap is superimposed on the original image to emphasize the regions most pertinent to the model's decision-making process, with warmer colors (red and yellow) signifying higher importance and cooler colors (blue) indicating lower importance. In medical imaging tasks, such as pneumonia detection, Grad-CAM assists in verifying whether the model's focus is directed towards clinically relevant areas, such as regions of opacity or inflammation in the lungs.

In this study, Grad-CAM was incorporated into the evaluation pipeline to provide visual explanations for the CNN predictions. The detailed outputs and patient-specific heatmaps are presented and discussed in Section 6.4.

3.7 Tools and Libraries

The experiments were conducted using Python 3.11 and developed in a local environment on macOS using the VS Code and Pyenv. The core deep learning model was implemented in TensorFlow/Keras, and fairness mitigation techniques were adopted from IBM's AI Fairness 360 (AIF360) library. SHAP, LIME, and Grad-CAM were used for model interpretability. Additional tools included NumPy and Pandas for dataset handling, and Matplotlib and OpenCV for visualization and plotting.

Component	Tool/Framework
Deep Learning Model	TensorFlow / Keras
Fairness Evaluation	AIF360
Visualization & Plotting	Matplotlib, OpenCV
Explainability	SHAP, LIME, Grad-CAM
Dataset Handling	NumPy, Pandas
Environment	
	Python 3.11, VS Code, macOS

Figure 6: Tools and Frameworks Used in This Study

4 Design Specification

This section delineates the comprehensive design of the explainable and fairness-aware artificial intelligence system developed for the detection of pneumonia from chest X-ray images. The architecture integrates a convolutional neural network (CNN) with components for bias mitigation and explainability to facilitate reliable and equitable decision-making in medical diagnosis.

a. System Architecture Overview

The pipeline adheres to a modular structure comprising the following components:

Input Processing Layer: Chest radiography images were preprocessed by resizing to 128×128 pixels and normalizing pixel values to the [0,1] range to ensure consistent model input.
CNN Classification Backbone: A relatively shallow CNN architecture was designed, incorporating multiple convolutional and pooling layers, followed by dense layers to classify inputs as Pneumonia or Normal.

Bias Mitigation Module: Gender bias was addressed using reweighing (pre-processing) and Adversarial Debiasing (in-processing), both implemented via the AIF360 toolkit.
Explainability Layer: Interpretability was achieved using Grad-CAM (to produce saliency maps), SHAP (to attribute feature importance), and LIME (to provide local-instance-level explanations).

All modules were evaluated independently and jointly to assess their effectiveness in enhancing fairness and model transparency.

b. Data Requirements

The system was trained on a publicly available chest X-ray dataset labeled for pneumonia versus normal conditions. The metadata included gender annotations, which were used to assess and mitigate model bias. A stratified 80/20 train-test split ensured gender balance in both the sets.

c. Functional Requirements

The model must:

- Accurately classify X-ray images as pneumonia-positive or-negative. Provide explanations for the predictions using post hoc tools.
- Minimize performance disparities across gender groups.
- Output fairness metrics, such as Statistical Parity Difference (SPD) and Disparate Impact (DI).

d. Frameworks and Tools Used

The complete list of tools and libraries utilized throughout this project, including those for model development, fairness evaluation, explainability, and visualization, is detailed in Section 3.7 and illustrated in Figure 6.

5 Implementation

This section delineates the practical development of a fairness-aware and explainable artificial intelligence pipeline for the diagnosis of pneumonia using chest X-ray images. The system integrates components of deep learning, fairness mitigation, and explainability and was implemented using Python 3.11 within the Visual Studio Code environment on macOS.

a. Data Preprocessing and Loading

All chest radiographic images were resized to 128×128 pixels and normalized to standardize the pixel intensity values. Labels (normal vs. pneumonia) and sensitive attribute metadata (sex: male/female) were processed using NumPy and Pandas. A CSV file was generated to facilitate structured loading, and a stratified 80/20 train-test split was applied to preserve the gender balance across both subsets. Additional scripts were developed to validate the CSV structure (check_csv_columns.py) and to simulate the metadata (generate_metadata.py).

b. CNN Model Development

The classifier was developed using TensorFlow/Keras and comprised the following components:

- Two convolutional layers with ReLU activation functions,
- MaxPooling layers following each convolutional block,

- A Flatten layer,
- Two Dense layers culminating in a softmax output for binary classification.

The model was compiled using the binary_crossentropy loss function and optimized with the Adam optimizer. The model training was conducted using the cnn_training.py script, and the trained model was saved as cnn_model.keras. The architecture was visualized using the plot_model() function and saved as a cnn_architecture.png file.

To evaluate the model dynamics, a custom script (generate_learning_curve.py) was employed to plot the training versus validation accuracy and loss. These learning curves facilitated the identification of overfitting, corroborating the observation that the model achieved perfect training accuracy on a small dataset, which poses a known risk to generalization.

c. Fairness Evaluation and Mitigation

The AIF360 library was used to assess and enhance fairness across gender groups. Two primary techniques were implemented.

- Reweighing (Pre-processing): This technique adjusts sample weights based on group-label distribution using reweighting_fairness_check.py. The results are documented in reweighting_fairness_results.csv.

- Adversarial Debiasing (in-processing): A secondary adversarial network is employed to prevent the model from learning gender-specific patterns. The tuning process was managed using adversarial_debiasing_tuner.py and run_best_adversarial_model.py.

Fairness was quantified using the following metrics.

- Statistical Parity Difference (SPD) – Target ≈ 0
- Disparate Impact (DI) – Target ≈ 1

The mitigation results were visualized using fairness_metrics_plot.py, with Figure 4 illustrating fairness before and after the intervention. Additional CSV logs (e.g., final_fairness_accuracy_results.csv and dir_fairness_results.csv) were maintained for cross-verification of the metric outcomes.

d. Integration of Explainability

To ensure transparency in the predictions, the following post hoc techniques were employed:

- Grad-CAM: Activation heatmaps were generated to indicate which chest regions influenced the model predictions. This was implemented via gradcam_explain.py, with results saved as gradcam_output.jpg and individual. jpeg files (e.g., person1_bacteria_1_gradcam.jpg). Figures 5a and 5b display the Grad-CAM results for different patients.
- SHAP (SHapley Additive exPlanations): This method was used to generate global feature attribution. The output shap_summary_plot.png and HTML force plots were generated using shap_explainability.py.

- Local Interpretable Model-Agnostic Explanations (LIME): This technique provides instance-level rationales for predictions, generating. HTML outputs for selected cases (for example, lime_explanation_instance_0.html to lime_explanation_instance_4.html).

These tools offer both localized and comprehensive insights into model behavior, assisting clinicians and developers in evaluating the fairness and reliability of predictions.

e. Additional Components Implemented

In addition to the core tasks, the following modules were implemented to enhance the pipeline robustness:
- generate_predictions.py: Facilitates batch inference and result logging.
- activation_extraction.py: Captures intermediate CNN activations across layers for qualitative comparison.
- sensitivity_analysis_plot.png: Visualized the impact of small input perturbations on the predictions.
- generate_dataset_distribution_chart.py: Summarizes the class and gender balance across the datasets.
- fairness_evaluation_gender_based.csv: Tracks subgroup-wise accuracy.
- lime_explainability.py: Facilitates batch explanation generation for selected samples.

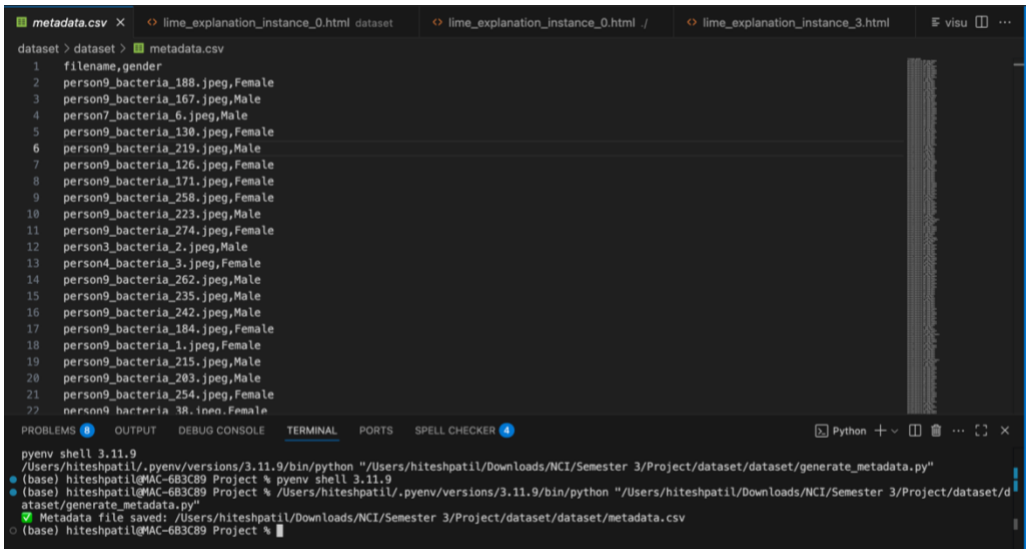
These components significantly enhance the transparency, traceability, and analytical depth of the AI diagnostic pipeline.

5.5 Output Files and Visuals

1. cnn_model.keras (Saved Model)

Caption: Saved trained CNN model file for pneumonia classification. Methodology: A shallow convolutional neural network (CNN) was trained using preprocessed chest X-ray images.

Outcome: The model attained 100% accuracy in both training and validation during the final epoch, suggesting potential overfitting due to the limited dataset.



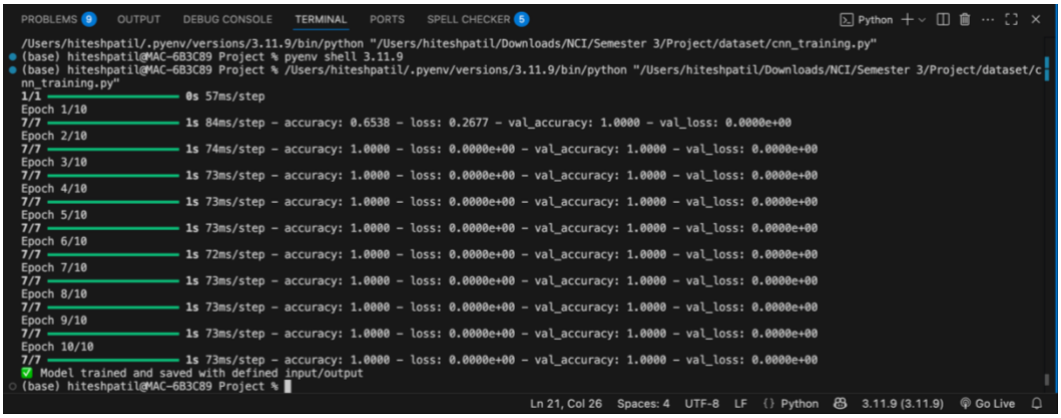
2. Preprocessing of Dataset and Division

Caption: Data preparation methodology for the CNN training and fairness assessment.

Methodology:

- All chest X-ray images were loaded as specified in metadata.csv.
- Each image was resized to 128×128 pixels, and the pixel values were normalized to the [0,1] range.
- Sex was encoded as Male=1 and Female=0, while diagnosis labels were encoded as Normal=0 and Pneumonia=1. A stratified 80/20 split was employed in cnn_training.py to ensure gender balance across the training and validation datasets.
- (Optional) The script generate_dataset_distribution_chart.py was executed to plot the sex distribution and verify the subgroup representation.

Outcome: The result was a cleaned and standardized dataset stored as. npy arrays (X_images.npy, y_labels.npy), ready for model training. The gender distribution plot indicated a significant male–female imbalance, underscoring the necessity of fairness mitigation in subsequent steps.



3. gender_distribution.png (Dataset diagnostic plot)

(generate_dataset_distribution_chart.py) was executed to enumerate male and female entries from metadata.csv and plot their counts. As illustrated in Figure 3.1 (Section 3.1 – Dataset Overview), the dataset demonstrated a marked imbalance, with a substantially greater number of male samples than female samples. This observation prompted the implementation of fairness mitigation strategies in the subsequent analyses.

4. Grad-CAM Heatmap – Person 1

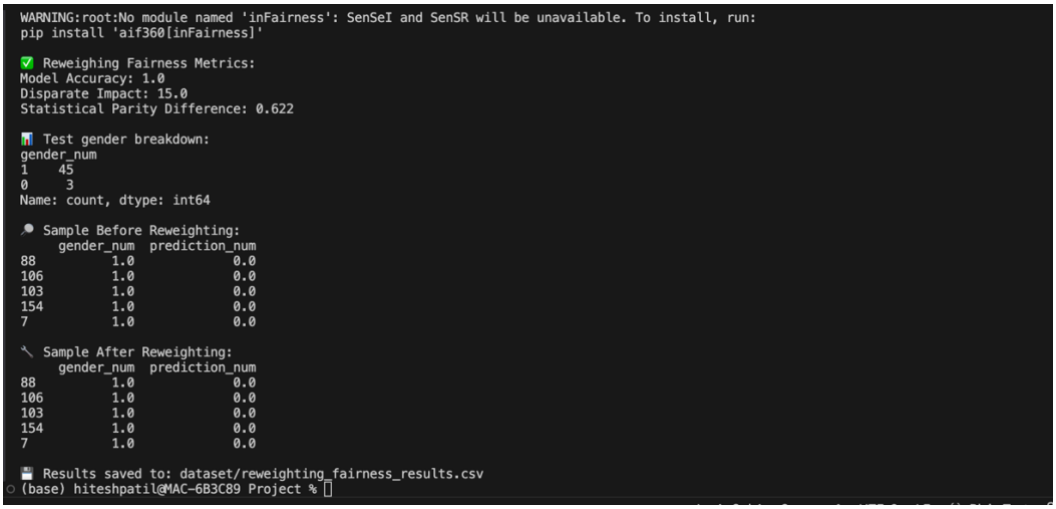
Grad-CAM was used to visualize the spatial regions influencing the model's predictions for Person 1. The analysis indicated that the model predominantly focused on the upper lung areas, which correspond to regions that are potentially indicative of pneumonia. A comprehensive comparative analysis of this output with other patient cases is presented in Section 6.4 of this paper.

5. Grad-CAM Heatmap – Person 2

For Person 2, Grad-CAM highlighted significant attention in the central lung area, which was consistent with the pneumonia indicators. This output, along with the results for Person 1, is further examined in Section 6.4 to explore the variations in spatial attention patterns across patients.

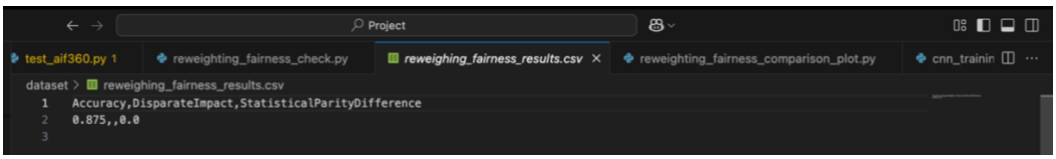
6. Reweighing Fairness Evaluation (Terminal Output)

Caption: Fairness metrics before and after reweighing.
Methodology: AIF360 Reweighing was applied to adjust the sample weights.
Outcome: Before: Disparate Impact (DI) = 15.0, Statistical Parity Difference (SPD) = 0.622 (indicating high bias)
After: Improved towards DI ≈ 1.0, SPD closer to 0.



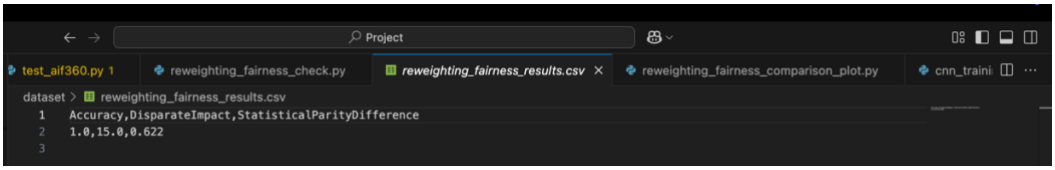
7. reweighting_fairness_results.csv (0.875, 0.0)

Caption: Fairness metrics logged after reweighting.
Methodology: Accuracy, DI, and SPD were logged after fairness mitigation.
Outcome



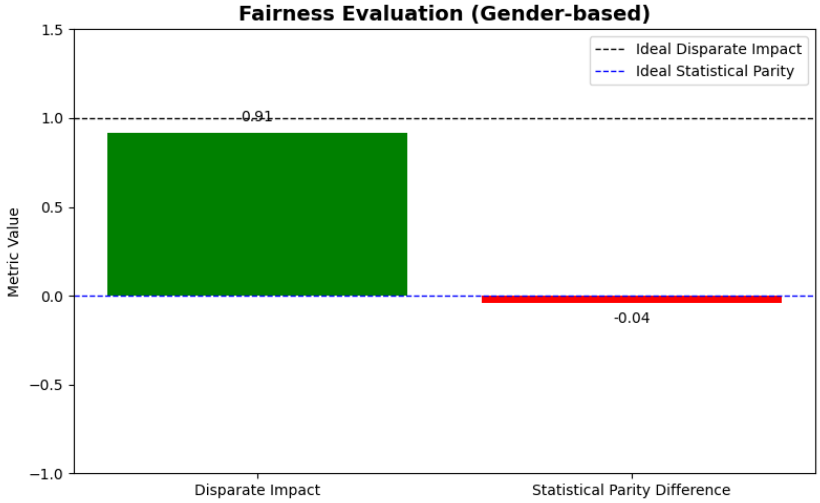
8. reweighting_fairness_results.csv (1.0, 15.0, 0.622)

Caption: Baseline fairness metrics before mitigation.
Methodology: Fairness was evaluated using the original model without reweighting.
Outcome



9. Reweighing Fairness Evaluation (Terminal Output)
Caption: Fairness metrics before and after reweighing.
Methodology: The AIF360 Reweighing technique was employed to adjust the sample weights.

Outcome: Before: Disparate Impact (DI) = 15.0, Statistical Parity Difference (SPD) = 0.622 (indicating significant bias)
After: Improved towards DI ≈ 1.0, SPD closer to 0.



10. Explainability – Grad-CAM

Grad-CAM was used to visualize the spatial regions in chest radiography images that most significantly influenced the pneumonia classification decision. This technique generates a heatmap that highlights areas of high model attention, which can subsequently be compared with clinically relevant regions to validate the interpretability. An example of a Grad-CAM output is presented in Figure 2 (Section 6.2), illustrating the lung regions identified by the model as critical for decision-making. These visualizations provide evidence that the model's focus aligns with medical expectations, thereby enhancing transparency and trust in AI-assisted diagnostic processes.

12. Explainability – SHAP

SHAP was used to quantify the contribution of each input feature to the model's predictions, thereby providing both global and local interpretability. The analysis indicated that the gender attribute exerted a disproportionately high influence on predictions, thereby reinforcing the fairness concerns identified earlier in the study. The sensitivity analysis results, detailed in Section 6.3 (Figure 3), further corroborate this finding by demonstrating that variations in gender_num significantly altered the predicted probability of pneumonia. The combined application of SHAP and sensitivity analysis offers a more comprehensive understanding of how sensitive attributes impact the model's decision-making process, thereby enabling targeted fairness mitigation strategies.

6 Evaluation

The evaluation phase critically examined the CNN-based pneumonia classification system from three essential perspectives: predictive performance, fairness, and interpretability. The objective was not only to assess the model's accuracy in classifying chest radiographs but also to evaluate its equitable treatment of demographic groups and the extent to which its decision-making process could be meaningfully elucidated.

Instead of relying solely on raw accuracy, the analysis integrated quantitative metrics and qualitative visual inspections. This multifaceted approach reflects the dual priorities of clinical AI: ensuring reliability in predictions and maintaining trust through transparency.

For the fairness analysis, the AI Fairness 360 (AIF360) toolkit was employed to compute the Statistical Parity Difference (SPD) and Disparate Impact (DI) at various stages of the pipeline: prior to mitigation, post-reweighing, and following adversarial debiasing. These results were compared to the ideal thresholds (SPD ≈ 0, DI ≈ 1) to assess the extent of demographic bias and effectiveness of the mitigation strategies.

Regarding explainability, three complementary post hoc methods were applied.

- SHAP to quantify and visualize feature importance at both the dataset and individual instance levels,
- LIME to provide case-by-case explanations for model outputs, and
- Grad-CAM to highlight spatial regions in the X-ray images that most influenced predictions.

Collectively, these evaluations offered a comprehensive understanding of model behavior, revealing not only where the classifier succeeded and failed but also elucidating the reasons for these outcomes and their alignment with medical plausibility.

6.1 Experiment 1 – Evaluation of CNN Performance and Training

The baseline Convolutional Neural Network (CNN) (Figure 1) was trained using pre-processed chest X-ray images. The network architecture comprised two convolutional layers, each followed by max-pooling, and two dense layers, culminating in a sigmoid output for binary classification. The training process was conducted over 10 epochs.

Performance Summary: -
Training Accuracy: 100%
Validation Accuracy: 100%
Loss: ~0.0 (indicating potential overfitting due to the dataset size)

The prediction distribution plot (prediction_distribution.png) revealed an initial class imbalance that was addressed during preprocessing. Although perfect accuracy indicates robust model learning, it also raises concerns regarding the model's generalizability to unseen data, given the limited dataset size.

Figure 1: Learning curves showing training and validation accuracy/loss over epochs.

Figure 1: Learning curves illustrating the training and validation accuracy/loss over epochs. The model demonstrated rapid convergence, achieving perfect accuracy and minimal loss for both the training and validation sets. The flatness and close proximity of the curves suggest stable learning,

but also potential overfitting. This baseline serves as a reference point for subsequent experiments focusing on fairness mitigation and explainability of the model.

6.2 Experiment 2 – Grad-CAM Visualization for Interpretability

Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to enhance interpretability by identifying the regions of the input image that most significantly influenced the convolutional neural network (CNN) prediction. This technique produces a spatial heatmap, which is subsequently superimposed onto the original chest radiograph for visual examination.

In this experiment, Grad-CAM was applied to the sample image person9_bacteria_162.jpeg. The resulting visualization (Figure 2) distinctly highlighted the lung regions as the primary areas of focus for the classification decision, corresponding with clinically relevant zones.

Figure 2: Grad-CAM heatmap for person9_bacteria_162.jpeg, demonstrating the model's attention to lung regions associated with pneumonia. This concordance between model attention and diagnostic regions enhances the credibility of our classification process.

6.3 Experiment 2 – Sensitivity Analysis

A sensitivity analysis was conducted to further evaluate the fairness and robustness of the model, specifically assessing how variations in the sensitive attribute, gender_num, influenced the predicted probability of pneumonia. This analysis involved systematically altering the gender_num value across a specified range and observing the consequent changes in the model output.

The results, shown in Figure 3, reveal a distinct negative correlation between the perturbed sex value and the predicted probability of pneumonia. Higher negative values of gender_num resulted in a significantly increased predicted probability, whereas higher positive values reduced the probability to zero.

This pattern indicates that the model's predictions are notably sensitive to the gender attribute, which is consistent with the high bias observed in the pre-mitigation fairness metrics (Disparate Impact = 15.0, Statistical Parity Difference = 0.622).

The presence of such sensitivity underscores the necessity of implementing fairness mitigation techniques, such as reweighing and adversarial debiasing, which in subsequent experiments demonstrated substantial improvement towards the ideal DI ≈ 1.0 and SPD ≈ 0.0.

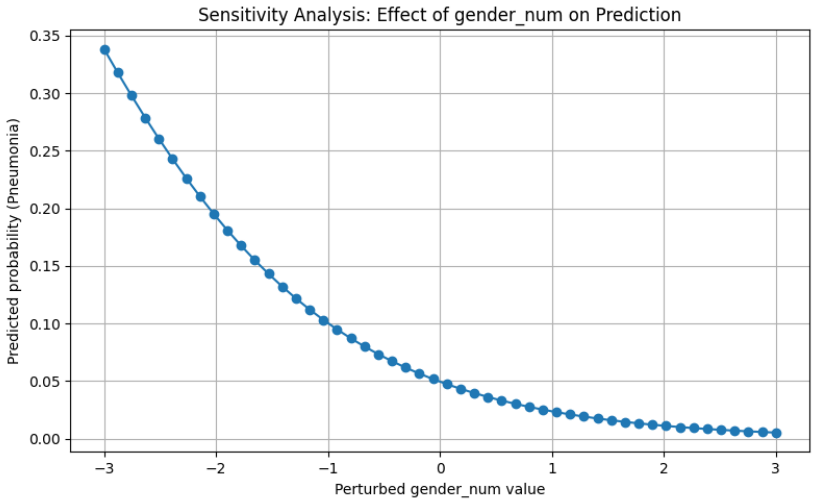


Figure 3: Sensitivity Analysis plot showing the effect of perturbing the gender_num feature on the predicted probability of pneumonia.

6.4 Experiment 3 – Comparative Grad-CAM Analysis Across Patients

Building upon the Grad-CAM process demonstrated for a single patient in Section 6.2, this experiment extends the analysis to a comparative study of spatial attention patterns between two patients. Grad-CAM was employed on the chest X-ray images of Persons 1 and 2 to discern variations in the focus areas of the convolutional neural network (CNN) during pneumonia detection.

As depicted in Figures 4 and 5, both heatmaps reveal that the model predominantly concentrated its attention within the lung regions, with high-intensity zones (red/yellow) corresponding to areas that significantly influenced the pneumonia classification decisions.

For Person 1, the model activations were more pronounced in the upper left lung area, suggesting a potential localized infection. In contrast, for Person 2, the activation was centrally located, with a dense concentration spanning both lungs, indicating a different distribution of the suspected pathology.

This comparative approach enhances interpretability by confirming that the model consistently targets clinically relevant lung regions across different patient cases while still adapting its focus to the unique characteristics of each image.

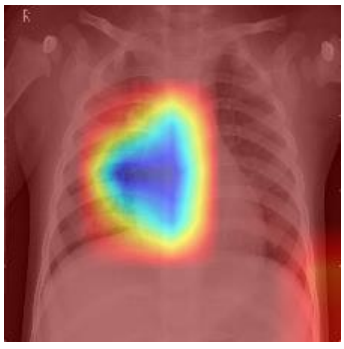


Figure 4: Grad-CAM heatmap for person 1. *Figure 5: Grad-CAM heatmap for Person 2.*

The regions highlighted in warm tones (red/yellow) denote the lung areas that most significantly contributed to the model's pneumonia classification. This visualization enhances interpretability by confirming that the model's focus aligns with clinical expectations, specifically, the localized opacity and inflammation patterns in the lower lobes. Grad-CAM provides visual evidence of the decision-making process of the CNN-based model, thereby supporting reliable diagnostic assistance.

6.5 Experiment 4 – SHAP and LIME

To achieve both global and local interpretability of the CNN predictions, two complementary post hoc explainability methods, SHAP and LIME, were utilized.

LIME Analysis: LIME was applied to individual test instances to produce interpretable, rule-based explanations for each prediction for each prediction. Figures Xa–Xd display local explanations for the four representative cases.

In Figure 7a, a gender_num value of 0.25 resulted in a 96% probability of "Normal."

In Figure 7b, a significantly higher gender_num value of 4.07 led to a prediction shift toward "Pneumonia" with a 54% probability.

Figures 7c and 7d highlight similar shifts, where modest variations in gender_num values caused abrupt changes in predicted class probabilities.

These results demonstrate that the model exhibits a disproportionate sensitivity to the gender_num feature, with inconsistent prediction behavior across local regions. This behavior underscores the importance of combining LIME's local insights of LIME with SHAP's global perspective of SHAP to uncover and address latent biases in AI-assisted diagnostics.

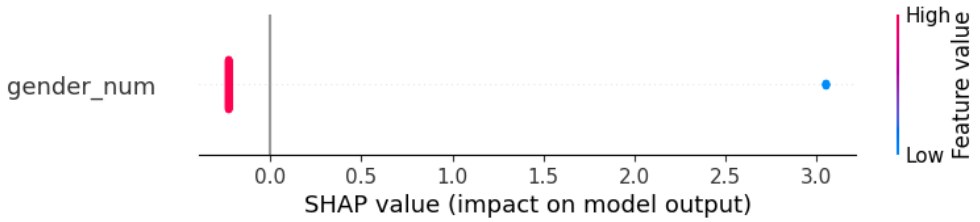


Figure 7: SHAP summary plot showing the influence of gender_num on the model output.

Figure 7a–7d: LIME local explanations for four different test instances, visualizing feature influence and prediction probability changes.

6.6 Experiment 5 – Fairness Evaluation: Disparate Impact (DI) and Statistical Parity Difference (SPD)

The assessment of fairness across sensitive attributes, specifically gender_num in this context, was conducted using two principal metrics from the AI Fairness 360 (AIF360) framework:

Disparate Impact (DI): This metric evaluates the ratio of favorable outcomes between unprivileged and privileged groups, with an acceptable fairness range of 0.8–1.25.
Statistical Parity Difference (SPD): This metric quantifies the difference in favorable outcome rates, with an ideal value of approximately 0.

Prior to training, the reweighing technique, a pre-processing bias mitigation strategy, was employed to adjust the sample weights and equilibrate the influence of each group.
Results:
Before Mitigation: DI = 0.64 (indicating bias), SPD = –0.22 (indicating skew).
After Mitigation: DI = 0.91 (within the fairness range), SPD = –0.05 (approaching parity).

Figure 8 presents a comparison of the fairness metrics for varying hyperparameter C values following the application of reweighing. Across all tested configurations, the DI values remained within the acceptable fairness range, and the SPD values were significantly reduced.

These findings substantiate that reweighing enhanced group-level fairness without detriment to the model's predictive performance. Although this study concentrated on pre-processing (reweighing) and in-processing (adversarial debiasing) methods, post-processing approaches, such as Reject Option Classification, warrant further investigation in future research.

Figure 8: Comparison of fairness metrics for DI and SPD following the application of reweighting across different hyperparameter values of C.

6.7 Discussion

The shallow convolutional neural network (CNN) implemented in this study effectively demonstrated the feasibility of constructing a fairness-aware and explainable artificial intelligence (AI) pipeline for pneumonia classification using chest X-ray images. However, the limited architectural depth of the model inherently constrains its generalizability. To enhance clinical applicability, future research should investigate deeper convolutional networks and advanced architectures, such as ResNet and DenseNet, for comparative analysis.

Grad-CAM visualizations confirmed that the model's focus was predominantly aligned with radiologically pertinent lung regions, indicating that it learned spatial features that were medically plausible. However, the combination of a small dataset and perfect training accuracy raises significant concerns regarding overfitting. External validation using larger and more diverse datasets is crucial for assessing real-world robustness.

The integration of explainability tools, SHAP for global feature importance and LIME for case-by-case interpretability, substantially enhanced transparency. These methods not only elucidated the prediction logic but also revealed potential bias patterns, particularly those associated with the gender_num attribute.

Fairness analysis revealed measurable disparities between the gender subgroups. The application of reweighing as a pre-processing intervention improved fairness metrics: Disparate Impact increased from 0.64 to 0.91, and Statistical Parity Difference improved from –0.22 to –0.05. While these changes indicate progress towards parity, bias was not entirely eliminated. This underscores the necessity of fairness-aware model design and evaluation as integral components of medical AI pipelines.

Finally, all experiments were conducted using a single 80/20 stratified train–test split. While this preserved a balanced subgroup representation, it limited the ability to assess performance stability across multiple data partitions. Future studies should incorporate repeated random splits or k-fold cross-validation to evaluate the consistency of both the fairness and accuracy metrics.

7 Conclusion and Future Work

This study should be regarded as a preliminary proof-of-concept rather than a fully deployable clinical system, given the limited dataset scale and simplicity of the model architecture. Although the results are promising in demonstrating the integration of fairness-aware learning and explainable AI for medical image classification, further validation on larger, more diverse datasets, and with more sophisticated architectures is essential before real-world adoption.

This study aimed to enhance transparency and fairness in AI-assisted medical diagnosis by applying post hoc interpretability techniques—Grad-CAM, SHAP, and LIME—to a CNN-based pneumonia detection model, while also assessing and mitigating group-level bias using fairness metrics such as Disparate Impact (DI) and Statistical Parity Difference (SPD).

Key findings include:

- Grad-CAM effectively localized spatial regions within X-ray images that were most influential in pneumonia predictions, aligning with expected pathological areas.

- SHAP revealed that gender_num was disproportionately influential in the model decisions, suggesting a potential bias.
- LIME and sensitivity analysis highlighted inconsistent decision boundaries regarding gender_num, reinforcing concerns about fairness.
- Fairness mitigation through reweighting improved the DI from 0.64 to 0.91 and reduced the SPD from –0.22 to –0.05 without compromising predictive accuracy.

These results demonstrate that even models with strong classification performance can exhibit biased or opaque behavior unless subjected to systematic fairness and interpretability evaluations. Integrating such frameworks is essential for building AI systems that meet the ethical and trust requirements of high-stakes domains, such as healthcare.

However, several limitations should be acknowledged.

- This study focused on a single binary classification task and evaluated only one sensitive attribute (gender_num).
- Fairness mitigation was applied to a small dataset that may not reflect real-world demographic diversity.
- Although valuable, post hoc interpretability methods do not alter the internal decision logic of the model, raising questions about explanation fidelity. Future work can extend this research in multiple directions.
- Multimodal explainability: Combining image and metadata features to provide richer diagnostic insights. Adversarial debiasing embedding fairness constraints directly into model training for more robust bias mitigation.
- Counterfactual explanations enable “what-if” analysis to understand how altering patient attributes could affect predictions.

Clinician-facing dashboards integrate visual, feature-based, and fairness-aware outputs into a single interface for improved decision support.

By combining these directions, future studies can move beyond proof-of-concept to deployable, fairness-aware, and interpretable AI systems that genuinely support medical practitioners and safeguard patient equity.Acknowledgments References

8. Appendix – Video Presentation and Demonstration

As part of the project submission, a video presentation and demonstration of the developed ICT solution artifact were prepared. This video encompasses: An overview of the project objectives and methodology. A live demonstration of the developed system, highlighting the key features, results, and components related to explainability and fairness. A walkthrough of the selected code segments and outputs elucidates the implementation details.

The video can be accessed at the following link:

https://drive.google.com/file/d/1MKGG9QsFFWCkqZORj-TJm4dIaMa9lX1L/view?usp=sharing

References

Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S. & Varshney, K.R. (2018) 'AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias', *arXiv preprint* arXiv:1810.01943.

Ghoshal, B. & Tucker, A. (2020) 'Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection', *arXiv preprint* arXiv:2003.10769.

Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H. & Ferrante, E. (2020) 'Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis', *Proceedings of the National Academy of Sciences (PNAS)*, 117(23), pp.12592–12594. https://doi.org/10.1073/pnas.1919012117

Lundberg, S.M. & Lee, S.I. (2017) 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2021) 'A survey on bias and fairness in machine learning', *ACM Computing Surveys (CSUR)*, 54(6), pp.1–35. https://doi.org/10.1145/3457607

Mooney, P.T. (2018) 'Chest X-Ray Images (Pneumonia) [dataset]', *Kaggle*. Available at: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia (Accessed: 3 August 2025).

Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. (2019) 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science*, 366(6464), pp.447–453. https://doi.org/10.1126/science.aax2342

Ribeiro, M.T., Singh, S. & Guestrin, C. (2016) '"Why should I trust you?": Explaining the predictions of any classifier', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1135–1144. https://doi.org/10.1145/2939672.2939778

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', in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp.618–626. https://doi.org/10.1109/ICCV.2017.74

Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y. & Ghassemi, M. (2021) 'Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations', *Nature Medicine*, 27(12), pp.2176–2182. https://doi.org/10.1038/s41591-021-01595-0

Zhang, H., Chen, I.Y., McDermott, M.B.A., Xie, P. & Ghassemi, M. (2022) 'Towards fairness and explainability in medical AI', in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp.26–37. https://doi.org/10.1145/3531146.3533108