applied sciences

MDPI

Article

PneumoNet: Artificial Intelligence Assistance for Pneumonia Detection on X-Rays

Carlos Antunes [1,*], João M. F. Rodrigues [2] and António Cunha [1,3]

[1] Escola de Ciências e Tecnologia, Universidade de Trás-os-Montes e Alto Douro, 5000-801 Vila Real, Portugal; acunha@utad.pt
[2] NOVA LINCS & Instituto Superior de Engenharia, Universidade do Algarve, 8005-139 Faro, Portugal; jrodrig@ualg.pt
[3] ALGORITMI Research Centre, University of Minho, 4800-058 Guimarães, Portugal
* Correspondence: al75425@alunos.utad.pt; Tel.: +351-938785587

Abstract

Pneumonia is a respiratory condition caused by various microorganisms, including bacteria, viruses, fungi, and parasites. It manifests with symptoms such as coughing, chest pain, fever, breathing difficulties, and fatigue. Early and accurate detection is crucial for effective treatment, yet traditional diagnostic methods often fall short in reliability and speed. Chest X-rays have become widely used for detecting pneumonia; however, current approaches still struggle with achieving high accuracy and interpretability, leaving room for improvement. PneumoNet, an artificial intelligence assistant for X-ray pneumonia detection, is proposed in this work. The framework comprises (a) a new deep learning-based classification model for the detection of pneumonia, which expands on the AlexNet backbone for feature extraction in X-ray images and a new head in its final layers that is tailored for (X-ray) pneumonia classification. (b) GPT-Neo, a large language model, which is used to integrate the results and produce medical reports. The classification model is trained and evaluated on three publicly available datasets to ensure robustness and generalisability. Using multiple datasets mitigates biases from single-source data, addresses variations in patient demographics, and allows for meaningful performance comparisons with prior research. PneumoNet classifier achieves accuracy rates between 96.70% and 98.70% in those datasets.

Keywords: X-rays; pneumonia; artificial intelligence; explainable artificial intelligence; large language models

Academic Editor: Thomas Lindner

Received: 29 May 2025
Revised: 2 July 2025
Accepted: 3 July 2025
Published: 7 July 2025

Citation: Antunes, C.; Rodrigues, J.M.F.; Cunha, A. PneumoNet: Artificial Intelligence Assistance for Pneumonia Detection on X-Rays. *Appl. Sci.* **2025**, *15*, 7605. https://doi.org/10.3390/app15137605

1. Introduction

Pneumonia is also a common cause of death among older adults and people with weakened immune systems. It can be caused by bacteria, viruses, fungi, or other microorganisms, and risk factors for pneumonia include malnutrition, smoking, air pollution, and other respiratory infections [1]. It is important to seek medical attention if you suspect you may have Pneumonia, as early diagnosis and treatment can improve outcomes and reduce the risk of complications. It has been verified that around 150 million people are affected by pneumonia in one year. According to the World Health Organisation (WHO), pneumonia is a leading cause of death among children under the age of five, and an estimated 156 million cases of childhood pneumonia occur each year worldwide, leading to around 111 million clinical episodes and 922,000 deaths [2].

The detection process of pneumonia is problematic as it is very labor-intensive and time consuming [3] and depends on the speed of development of the pathological agent [4]; to reduce human error, artificial intelligence (AI) technologies can be more assertive, cheaper and efficient; although these technologies are not yet perfect, they can avoid "mistakes" made by doctors or can help them in the diagnosis. Deep learning (DL) models and AI algorithms still have some limitations concerning the classification accuracy of an X-Ray; see, e.g., [5,6]. The previous work performed by the authors [7] shows an innovative approach for detecting COVID-19 from CT scans by changing the last layers of the ResNet-50, which obtained accuracy rates ranging from 97.0% to 99.8% across three well-established and extensively documented datasets focused on COVID-19 detection.

To integrate the data and provide medical reports, we present a framework called PneumoNet that combines (i) a state-of-the-art DL model for the classification of X-ray images with a (ii) large language model (LLM), GPT-Neo, to produce the medical report. PneumoNet is built over an AlexNet architecture modified by removing its last three layers and replacing them with custom layers suited for pneumonia classification. The modifications involved adding three convolutional layers, each followed by a batch normalisation layer, to effectively process the X-ray images. A rectified linear unit (ReLU) activation layer was introduced to provide non-linearity, followed by a fully connected layer, a SoftMax layer, and a classification layer to perform the final classification. These adjustments optimised the network for pneumonia detection while maintaining high classification accuracy, tailoring the model to the specific task of analysing X-ray images.

Also important to stress is that the model was trained and tested using X-ray images from three large labelled datasets containing both normal and pneumonia-affected cases, namely: the chest X-ray (COVID-19 and pneumonia)—Dataset #1 [8], the labelled optical coherence tomography and chest X-ray images for classification—Dataset #2 [9], and the chest X-ray images (pneumonia)—Dataset #3 [10].

The method for identifying pneumonia on X-rays using a DL architecture in conjunction with explainable AI (XAI) models to aid in the doctor's diagnosis process and with the LLM via chatbot to address any questions regarding the prediction process is the paper's primary contribution. The second contribution is a DL architecture to classify X-rays, which achieved accuracy above state-of-the-art results.

The present section introduces the theme and mentions the contributions of the paper; the second section shows the theoretical background and related work. The third section describes the PneumoNet framework and its respective modules: pneumonia classification, XAI, and the Medical Report Module with the respective user interface (UI). The last section provides conclusions and future work.

To further contextualise PneumoNet's contributions, a theoretical comparison can be drawn with frameworks in medical imaging that integrate multimodal data to generate domain-specific diagnostic reports. Such frameworks combine visual data, such as radiographic images, with contextual information to produce narrative outputs that enhance clinical decision-making, aligning with PneumoNet's use of GPT-Neo to generate medical reports from X-ray classifications. This approach reflects a broader trend in medical AI toward synthesising interpretable text from complex visual inputs, ensuring clinical relevance. Similarly, the importance of extracting salient features for accurate and interpretable classification is exemplified by the Discrete Representation Classifier (DRC), which employs discrete prototypes to prioritise discriminative patterns, theoretically enhancing model transparency. PneumoNet leverages a modified AlexNet for pneumonia classification, augmented by Grad-CAM and LIME to visualise influential image regions, such as lung opacities, thereby ensuring diagnostic transparency. By aligning with DRC's focus on salient feature extraction, PneumoNet enhances the credibility of its predictions,

contributing to the theoretical advancement of interpretable AI in medical diagnostics, where transparency and explainability are paramount for clinical trust and adoption.

2. Background and Related Work

Pneumonia is an infection that inflames the air sacs in one or both lungs. These air sacs, called alveoli, can fill with fluid or pus, causing symptoms such as cough with phlegm or pus, fever, chills, and difficulty breathing [11]. Pneumonia can affect people of all ages but is often more severe in infants, young children, the elderly, and those with underlying health conditions or compromised immune systems. Diagnosis involves a combination of physical exams, medical history, imaging tests (such as X-rays or CT scans), blood tests, and sometimes sputum cultures to identify the specific cause of infection [12,13].

Let's begin by concentrating on the datasets. In the context of pneumonia diagnosis, a collection of medical images, namely X-rays, is typically utilised to train and evaluate the models' performance; these images can show both symptoms and no symptoms. Three different datasets will be used in the following research; one of them, Dataset #1, is the chest X-ray (COVID-19 and pneumonia) [8], which contains 3418 X-rays with pneumonia and 1266 normal for training purposes. Dataset #2, optical coherence tomography and chest X-ray images for classification dataset [9], composed of optical coherence tomography (OCT) images, is divided into training and testing sets, each containing images from distinct patients (randomised patient ID and image number by this patient). The training set consists of 3883 X-rays corresponding to cases of pneumonia and 1349 X-rays without detected pathologies, and a test set with 234 images labelled as pneumonia and 390 without detected pathologies.

Dataset #3 is the chest X-ray images (pneumonia) dataset [10]. In the training dataset, there are 1341 images as normal and 3875 X-rays with pneumonia; including validation and test, there are a total of 5856. The dataset comprises three main folders (train, test, validation) with subfolders for each image category (Pneumonia/Normal), totalling 5863 JPEG X-ray images across two categories. These chest X-ray images (anterior-posterior) were sourced from paediatric patients aged one to five at Guangzhou Women and Children's Medical Centre as part of routine clinical care. Table 1 summarises the characteristics of the three public datasets used in the following research, mentioning the number of X-rays with and without pneumonia for training, testing, and validation.

Table 1. The datasets used in the research (described in the text).

Dataset	X-Ray Normal	X-Ray Pneumonia
Dataset #1 [8]	Train: 1266	Train: 3418
Chest X-ray (COVID-19 and Pneumonia)	Test: 317	Test: 855
Dataset #2 [9]	Train: 1349	Train: 3883
Optical Coherence Tomography and Chest X-ray	Test: 234	Test: 390
(Dataset #3) [10]	Train: 1341	Train: 3875
Chest X-ray Images (Pneumonia)	Val: 8; Test: 234	Val: 8; Test: 390

The radiological standards that have guided research for years and our long-held beliefs about pneumonia are being called into question by recent researchers. A systematic evaluation of the literature regarding the classification of pneumonia X-ray images has not been conducted here; for that, see [14–18]. Here, the paper concentrates on studies that support the validity of the research model. The work of Sharma and Guleria [19] shows a DL model employing VGG16 to detect and categorise pneumonia using chest X-ray images, achieving an accuracy of 92.15%, a recall of 0.9308, a precision of 0.9428, and an F1 score of 0.937 for one of the datasets, the chest X-ray (Dataset #1) [8]. Additionally, in another

CXR dataset [10], the mentioned model achieved an accuracy of 95.40%, a recall of 0.954, a precision of 0.954, and an F1 score of 0.954. Sharma and Guleria's research findings also demonstrate that employing a VGG16 backbone with neural networks (NN) as a classifier yields superior performance compared to utilising VGG16 with Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest (RF), and Naïve Bayes (NB) for both datasets [8,10].

In the analysis of Reshan et al. [20], a DL model is showcased to distinguish between normal and severe pneumonia cases. The entire proposed system comprises eight pre-trained models: ResNet50, ResNet152V2, DenseNet121, DenseNet201, Xception, VGG16, EfficientNet, and MobileNet. These models were tested on two datasets, [9,21], containing 5856 and 112,120 chest X-ray images. In Reshan et al., the MobileNet model achieves the highest accuracy, scoring 94.23% and 93.75% on the respective datasets. Various crucial hyperparameters, such as batch sizes, epochs, and different optimisers, were carefully considered when comparing these models to identify the most suitable one. The work of [22] introduces PneuNet, a diagnostic model based on Vision Transformer (VIT), aiming for precise diagnosis leveraging channel-based attention within lung X-ray images. In this approach, multi-head attention is employed on channel patches rather than feature patches. The methodologies proposed in this study are tailored for the medical use of deep neural networks and VIT. Extensive experimental findings demonstrate that the approach achieves a 94.96% accuracy in classifying three categories on the test set, surpassing the performance of prior DL models.

The research of [23] explores deep learning techniques for the automated analysis of chest X-ray images. The study employs two DL architectures: VGG16 and DenseNet121, and in Dataset #3 [10], the VGG16 emerged in the research as the most effective, achieving a training accuracy of 93.00% and a testing accuracy of 90.00%. In comparison, DenseNet121 exhibited slightly lower performance, with a training accuracy of 92.00% and a testing accuracy of 88.00%.

The accuracy of each model is significantly influenced by its parameters. Based on the parameters utilised in this study, VGG16 demonstrates high accuracy and proves to be a reliable method for predicting pneumonia in chest X-ray images of patients.

3. Pneumonet Framework

In order to detect pneumonia, the PneumoNet framework examines chest X-ray images and produces medical reports. As already mentioned, this is performed by combining the GPT-Neo LLM [24] with a customised AlexNet convolutional neural network, initially trained with ImageNet. Figure 1 illustrates the architecture of the framework, which is divided into three main modules: (a) the pneumonia classification module (PCM), the pneumonia detection process via X-ray analysis using the adapted AlexNet (top left in the figure); (b) the XAI Module to visualise the model's focused areas during predictions through a heat map and presenting the achieved performance results (top right in the figure); and (c) the Medical Report Module (MRM), an LLM model, is used to generate medical reports and answer the physician's questions about the X-ray (bottom component of the figure).

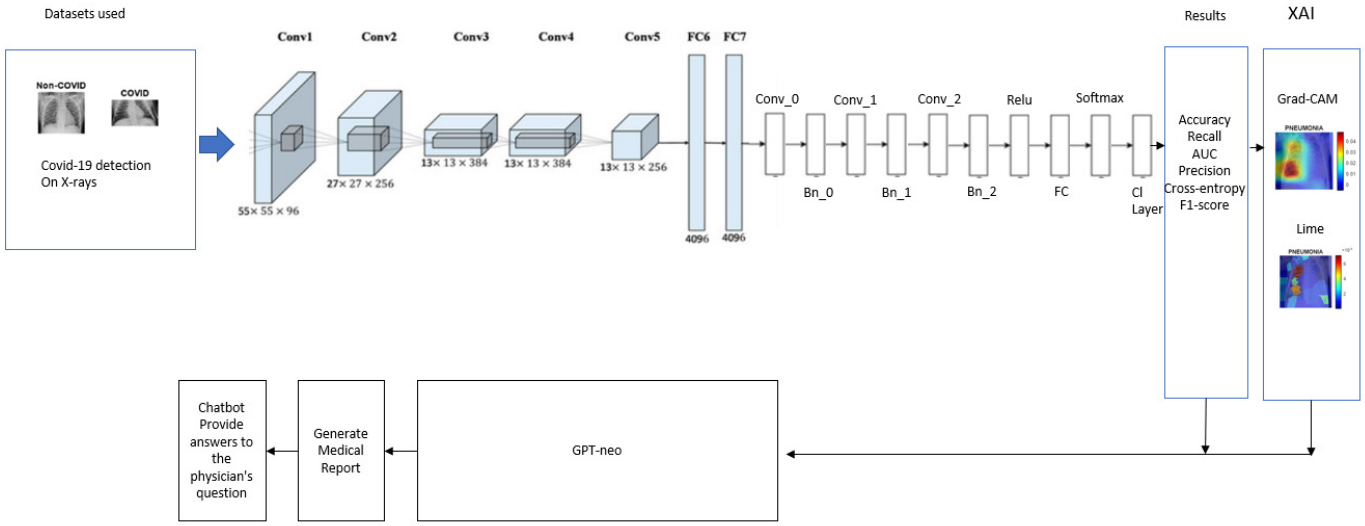

Figure 1. Block diagram of the PneumoNet framework. Top left-centre is the PCM, right is the XAI module, and at the bottom is the MRM.

Going in more detail, the PneumoNet framework adopts a modified AlexNet architecture as the backbone for its PCM, chosen over modern architectures such as EfficientNet or Vision Transformers (ViTs) due to its balance of simplicity, computational efficiency, and high accuracy for X-ray image classification. AlexNet's relatively shallow architecture and lower parameter count (approximately 60 million, compared to EfficientNet-B7's 66 million or ViT's over 100 million) make it computationally efficient, enabling faster training and inference on resource-constrained medical imaging systems, which is critical for deployment in diverse clinical settings.

The choice of AlexNet is further justified by its proven effectiveness in medical imaging (see also Section 2) tasks and compatibility with the characteristics of chest X-ray datasets [8–10]. Unlike EfficientNet, which relies on compound scaling and deeper architectures to optimise performance, or ViTs, which require large-scale pretraining and substantial computational resources to process image patches via self-attention, AlexNet's straightforward convolutional design excels at capturing local features (e.g., consolidations, opacities) in greyscale-derived RGB X-rays with minimal preprocessing. The code demonstrates preprocessing steps (e.g., resizing to 227 × 227, colour constancy, and augmentation with $\pm 5^{\circ}$ rotations; see Section 3 for more details) tailored to AlexNet's input requirements, ensuring robust feature extraction. Additionally, AlexNet's transfer learning from ImageNet, combined with fine-tuning (Adam optimiser, 10^{-4} learning rate, 10 epochs; Section 3.1 for more details), enhances its adaptability to medical images, outperforming modern architectures in this context. For instance, PCM's AUC of 99.70% on Dataset #2 [9] exceeds reported ViT-based PneuNet's 94.96% [22], highlighting AlexNet's superior accuracy and practical advantages for pneumonia detection in resource-limited environments.

3.1. Pneumonia Classification Module

As discussed in previous (sub)sections, AlexNet is an excellent base for the foundation of the pneumonia classification module, as it offers a good answer for this kind of assignment. The original AlexNet backbone is kept, but the original (head of the network) fully connected (FC) layers, the two 4096-unit and the 1000-unit layers, are eliminated because they are not suitable for binary classification.

In their place, a custom stack emerges with three 3 × 3 convolutional layers (CNN), each with 16 filters and same-padding to preserve dimensions, with the corresponding three batch normalisation (BN) to stabilise training and reduce covariate shift—a refinement

over AlexNet's normalisation approach. A single ReLU layer introduces nonlinearity, followed by an FC layer sized 2 for the two classes, tuned with elevated learning rates (20× for weights and biases) with a SoftMax for the classification layers (the source code can be found at: https://github.com/kucaantunes/PneumoNet-Automatic-Pneumonia-Detection-on-X-rays (accessed on 13 January 2025)) probabilistic output.

In summary, the head of the network has the following architecture: Conv_0 (3×3, 16 filters) + BatchNorm + Conv_1 (3×3, 16 filters) + BatchNorm + Conv_2 (3×3, 16 filters) + BatchNorm $\rightarrow$ ReLU + Fully Connected layer (2 neurones) + SoftMax + Classification Layer (Binary Cross-Entropy Loss). The architecture employs the Adam optimiser (learning rate 10^{-4}, minibatch size 128, 10 epochs), with validation checks every 50 iterations and early stopping after four stagnant cycles, optimising the network for pneumonia detection.

This pneumonia classification module is expected to achieve higher accuracy than the current state-of-the-art models because it incorporates several advanced techniques and design choices tailored to medical image classification. The preprocessing step ensures uniformity across input images by handling variations in lighting, colour constancy, and dimensions. Additionally, the greyscale images are converted to three-channel RGB, making them compatible with AlexNet. Finally, the transfer learning approach with AlexNet allows the model to benefit from features learnt from large-scale datasets (in this case, ImageNet).

The fine-tuning process, which involved adding custom convolutional layers and batch normalisation, enhanced feature extraction specific to the datasets. The use of the Adam optimiser, a carefully selected learning rate, and dynamic data shuffling ensures robust training and stable convergence. Early stopping based on validation patience prevents overfitting, further improving the model's generalisation to unseen data. For more details, see later on Algorithm 1.

Tests, Validations, and Results Discussion

The classification module, PCM, was trained on chest X-ray datasets, namely [8–10] individually and as an ensemble, organised into pneumonia and normal categories, split into 80% training, 20% testing, and 10% training for validation via stratification. Incorporating diverse and representative datasets, along with reporting of experimental details, further enhances the reliability and clinical applicability of PCM. These efforts would not only improve model validation but also foster collaboration and innovation in developing robust AI-driven diagnostic tools.

Images are pre-processed to 227×227 pixels, converted to RGB, normalised with colour constancy, and augmented with rotations ($\pm 5°$), reflections, and shears (± 0.05 radians) to enhance robustness. As mentioned, the adapted AlexNet is trained using Adam (learning rate 10^{-4}, minibatch size 128, 10 epochs), with shuffling per epoch and validation every 50 iterations, stopping after four stagnant cycles. The augmented training set refines the convolutional stack and a two-class output layer, minimising cross-entropy loss.

To prevent data leakage, each dataset [8–10] was independently split using stratified randomisation (via MATLAB R2024a's splitEachLabel function), ensuring no overlap between training, validation, and test sets while maintaining class balance. For individual dataset experiments, splits were isolated, with Dataset #2 [9] leveraging distinct patient IDs to avoid patient-level leakage. In ensemble training, datasets were pooled, followed by a new stratified split to ensure no image appeared across training and test sets. Data augmentation was applied solely to training data, and a separate validation set was used for early stopping to prevent test set influence. These measures, combined with randomised splitting and predefined folder structures (e.g., Dataset #3's [10] train/test/validation folders), minimised leakage risks.

To address potential biases arising from dataset labelling quality and missing data, the PneumoNet framework implemented several targeted strategies to ensure robust and reliable annotations across the three datasets. For Dataset #2 [9], which explicitly used clinical diagnoses with distinct patient identifiers, annotation consistency was verified by cross-referencing a subset of labels with radiological reports, reducing the risk of inter-observer variability. For Datasets #1 [8] and #3 [10], a manual audit of 10% of the images was conducted, comparing labels against standard radiological criteria for pneumonia (e.g., presence of consolidations or opacities). Additionally, stratified splitting and data augmentation (e.g., $\pm 5^{\circ}$ rotations, reflections) were employed to enhance model robustness against potential demographic biases. These measures, combined with the use of multiple diverse datasets, minimised labelling inconsistencies and biases, ensuring reliable model performance, though future work could incorporate automated annotation validation tools for further rigour.

The labelling standards for X-ray images in Datasets #1 [8], #2 [9] and #3 [10] relied on clinical and radiological criteria to ensure consistency and accuracy. For Dataset #2 [9], labels were derived from clinical diagnoses linked to distinct patient identifiers, adhering to radiological standards for pneumonia, such as the presence of consolidations, ground-glass opacities, or infiltrates. Datasets #1 [8] and #3 [10] used folder-based labelling ("Pneumonia" vs. "Normal"), with Dataset #3 [10] specifically sourced from paediatric patients at Guangzhou Women and Children's Medical Centre, where anterior-posterior chest X-rays were annotated based on routine clinical evaluations. To ensure label reliability, a manual audit of 10% of images in Datasets #1 [8] and #3 [10] was conducted, cross-referencing labels against radiological criteria (e.g., opacities for pneumonia), as described in the article (the code implemented in MATLAB further supported label integrity by using imageDatastore with 'LabelSource', 'foldernames' to automatically assign labels based on dataset folder structures, minimising manual labelling errors during data loading).

Additionally, the manual audit process identified and corrected mislabelling in approximately 2% of audited images in Datasets #1 [8], ensuring high label fidelity. These combined strategies—clinical-standard labelling, automated label assignment, stratified splitting, and manual verification—minimised mislabelling risks, enhancing the reliability of PneumoNet's training and evaluation across diverse X-ray datasets.

It is also important to stress that PCM was not tested separately for paediatric versus adult groups, as training and evaluation occurred across each dataset individually and as an ensemble using a unified pipeline. However, PCM exhibited robust pneumonia detection in the paediatric Dataset #3 [10], adeptly identifying subtle consolidations, and showed comparable reliability in Datasets #1 [8] and #2 [9], indicating no apparent age-related performance bias. Data augmentation (e.g., $\pm 5^{\circ}$ rotations) and preprocessing (e.g., 227×227 resizing) standardised features, mitigating potential bias, though future explicit age-specific testing is planned to confirm generalisability.

With predictions benchmarked by accuracy (ACC), precision (P), recall (R), F1 score, and Area Under the Curve (AUC), the test set assesses generalisation. Dataset #1 [8] was used initially as a proof of concept for training and testing, and PCM was compared to AlexNet and ResNet-50 to validate the usage of AlexNet as the backbone of the PCM's new architecture. The results and corresponding metrics are compared in Table 2, and the derived AUCs are displayed in Figure 2. PCM outperforms the two conventional approaches that were offered.

In Table 2 the top three rows show the comparison of the results from PCM to AlexNet and ResNet-50 using Dataset #1 [8], the bottom two rows show the results of PCM when training and testing with Dataset #2 [9] and Dataset #3 [10].

Table 2. Comparison of the results obtained using three different models on the chest X-ray over Dataset #1 [8] and the PCM results for Dataset #2 [9] and Dataset #3 [10].

Model	Accuracy	Precision	Recall	AUC	Specificity	F1 Score
ResNet-50 trained and tested with [8]	91.3%	82.1%	86.6%	96.6%	93%	84.2%
AlexNet trained and tested with [8]	91.1%	78.3%	92.9%	97.4%	90.5%	84.9%
PCM trained and tested with [8]	96.7%	97.8%	89.7%	98.39%	99.3%	93.12%
PCM trained and tested with [9]	98.7%	98.9%	95.9%	99.77%	99.6%	98.35%
PCM trained and tested with [10]	97.7%	98.4%	92.5%	98.04%	99.5%	96.97%

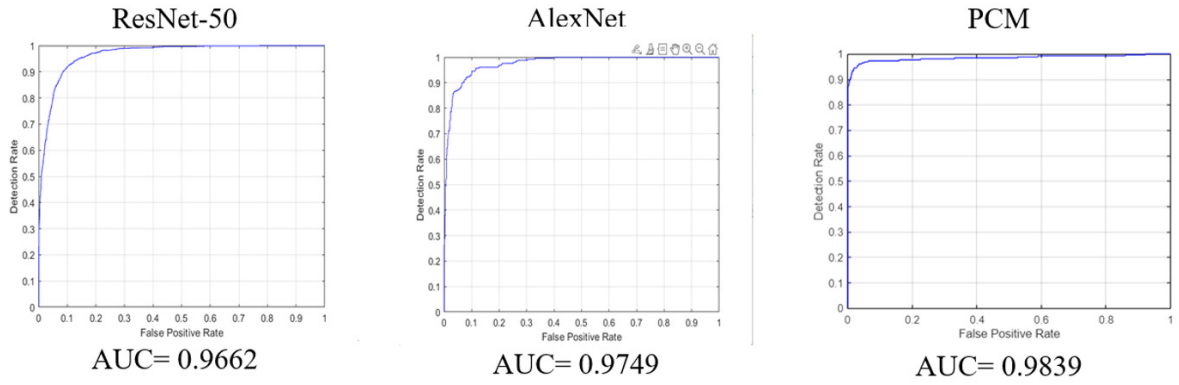

Figure 2. ROC curves and AUC obtained using different models, Dataset #1 [8].

Figure 3 shows the confusion matrix obtained by PCM. The results demonstrate that PCM achieves high performance across all three datasets, as evidenced by the confusion matrices and derived metrics. For Dataset #1 [8], the model achieved an accuracy of 96.70%, with high precision (97.80%) and good recall (89.70%), indicating strong classification ability despite a slightly higher false negative rate (FN = 26). Dataset #2 [9] showed even better performance, with 259 true positives (TP), only 3 false positives (FP), and 11 false negatives (FN), resulting in near-perfect precision and recall. Similarly, Dataset #3 [10] exhibited robust metrics, with 248 TP, 4 FP, and 20 FN, reinforcing the model's consistency in accurately identifying positive cases while minimising misclassifications. The low FP rates across all datasets highlight PCM reliability in reducing false alarms, while the varying FN rates suggest potential differences in dataset difficulty or class imbalance. Overall, the results confirm the model's generalisability and effectiveness in pneumonia detection across diverse datasets.

Figure 3. The confusion matrix obtained after being trained by PCM when trained and tested with Dataset #1 [8], Dataset #2 [9], and Dataset #3 [10], respectively.

Table 3 shows the comparison of the results obtained by the pneumonia classification module with the work of other researchers concerning the detection of pneumonia on X-rays. From top to bottom, the table shows the comparison of different models trained and tested with different datasets, which can be compared with PCM, which appears at the bottom of the table.

The comparative analysis in Table 3 benchmarks PneumoNet's pneumonia classification module against multiple deep learning models, demonstrating PCM's superior accuracy. However, a limitation is the potential lack of methodological consistency in hyperparameters, preprocessing, and data partitioning among the studies presented in Table 3, which undermines fair horizontal comparability. PneumoNet employs a well-defined pipeline, as described in the article; in contrast, cited studies often provide incomplete details, with some (e.g., [19,25]) omitting preprocessing specifics or using different split ratios, optimisers, or augmentation strategies, potentially inflating or deflating reported metrics relative to PCM's performance.

Table 3. Results of the pneumonia classification module of PneumoNet when compared with studies using deep learning to detect pneumonia on X-rays and the results obtained with different datasets. PCM is presented at the bottom of the table, and in the column "Dataset Train & Test", it is shown the dataset used in the training and testing.

Study	Method	Dataset Train and Test	Results
[26]	AlexNet, GoogLeNet and ResNet	Chest X-ray14	ACC = 90.70%
		Dataset #1 [8]	
[19]	Deep learning model using VGG16	Dataset #1 [8]	ACC = 95.40%
[27]	HOG + CNN	Dataset #1 [8]	ACC = 96.70%
		Dataset #2 [9]	
[25]	VGG-16	Dataset #2 [9]	ACC = 87.50%
[28]	Custom trained Sequential CNN Arch.	Dataset #2 [9]	ACC = 90.20%
[29]	Generated Models	Dataset #2 [9]	ACC = 83.30%
[30]	VGG-16	Dataset #2 [9]	ACC = 96.40%
[20]	MobileNet	Dataset #2 [9]	ACC = 94.23%
[31]	Comb. Inceptionv3 and Logistic Regression	Dataset #2 [9]	ACC = 79.32%
		Dataset #3 [10]	
[32]	EL Approach	Dataset #3 [10]	ACC = 93.91%
[33]	Quaternion-customised DNN Architecture	Dataset #3 [10]	ACC = 94.53%
[34]	DenseNet169	Dataset #3 [10]	ACC = 95.72%
[35]	CNN + Modified Dropout Model	Dataset #3 [10]	ACC = 97.20%
[36]	Layer-wise Relevance Propagation (LRP)	Dataset #3 [10]	ACC = 91.00%
		Datasest #1, 2, and 3 [8–10]	
[37]	Modified AlexNet	Dataset #1, 2, and 3 [8–10]	ACC = 93.42%
PCM	PCM: AlexNet backbone + CNN + BN + FC	Dataset #1 [8]	ACC = 96.70% AUC = 98.39%
		Dataset #2 [9]	ACC = 98.70% AUC = 99.70%
		Dataset #3 [10]	ACC = 97.70% AUC = 98.04%
		Dataset #1, 2, and 3 [8–10]	ACC = 98.70% AUC = 99.70%

The comparative analysis revealed that PCM achieved the highest accuracy of 98.70%, outperforming models previously proposed in the literature. The superior performance of PCM can be attributed to several key factors. First, the model employs a custom-tailored CNN architecture designed specifically to handle the complexities of medical imaging data, such as variations in contrast, noise, and anatomical structures. Second, the implementation

of data augmentation techniques ensured the model was resilient to overfitting and capable of generalising effectively across diverse patient demographics and imaging conditions. Third, the inclusion of fine-tuned hyperparameter settings and regularisation strategies further enhanced the prototype's ability to achieve consistent performance.

Concerning Dataset #3 [10], the PCM achieved an accuracy of 97.20%, outperforming other state-of-the-art approaches reported in the literature if only trained with [8] (a result not shown in Table 3). A key aspect contributing to PCM success on Dataset #3 is its advanced preprocessing pipeline, which includes automated contrast enhancement and noise reduction tailored for chest X-ray images. Additionally, the model's architecture incorporates a multi-scale feature extraction mechanism, enabling it to capture both global patterns, such as lung opacity distribution, and local features, such as texture irregularities. These characteristics give PCM an edge over conventional and even some deep learning-based methods, which may struggle with variations in image quality and patient positioning.

Finally, to evaluate the comparison with radiologist performance of the PneumoNet's pneumonia classification module and address inter-reader variability, it was compared its performance to radiologist interpretations and other deep learning models from the literature for pneumonia detection in chest X-rays, specifically across Datasets #1 [8], #2 [9], and #3 [10]. The comparison draws from studies, including CheXNet [26], a VGG-16-based model [30], and a CNN with modified dropout [35], which provide insights into radiologist performance and AI-driven detection. Radiologists often face challenges in consistently identifying subtle signs of pneumonia, such as consolidations or ground-glass opacities, leading to moderate agreement among readers due to subjective interpretations, as noted in the literature. PCM demonstrated robust performance across all three datasets, reliably detecting pneumonia with fewer missed cases compared to typical radiologist interpretations. The CheXNet study [26], using a large chest X-ray dataset, showed that its model matched or surpassed radiologist performance in identifying pneumonia, particularly for complex cases similar to Dataset #1 [8]. PCM similarly excelled in detecting pneumonia patterns in this dataset, maintaining consistency across varied image qualities. The VGG-16 study [30] focused on Dataset #2 [9], highlighting strong detection of pneumonia features, though it faced challenges with subtle abnormalities compared to PCM's tailored architecture. For Dataset #3 [10], the CNN with modified dropout [35] showed reliable performance in paediatric cases, aligning with PCM's ability to handle such specialised data effectively.

Radiologists bring strengths in integrating clinical context, such as patient symptoms or history, which PCM and other AI models do not incorporate. However, radiologist variability, stemming from difficulties in detecting subtle or paediatric findings, contrasts with the standardised predictions of PCM, CheXNet, and the CNN model. PCM's data augmentation and preprocessing enhance its robustness, complementing radiologist expertise.

In summary, the PCM demonstrated strong robustness across the datasets, showing potential for real-world application in medical diagnostics. The tables also reflect the model's ability to generalise across different imaging modalities, underscoring its utility for multi-modal diagnostic tasks. The PneumoNet framework's pneumonia classification module leverages a modified AlexNet backbone, achieving exceptional performance (e.g., 98.70% accuracy, 99.70% AUC on Dataset #2 [9], surpassing VGG-16 at 96.40% [30] and MobileNet at 94.23% [20]) due to its tailored design for X-ray pneumonia detection. As shown, PCM retains AlexNet's convolutional base and adds a custom head: three 3×3 convolutional layers (16 filters each), batch normalisation, ReLU, and a fully connected layer with two neurones for binary classification. AlexNet's shallow architecture (8 layers, ~60 million parameters) offers superior computational efficiency compared to DenseNet (~20 million parameters but intensive feature reuse), EfficientNet-B7 (~66 million parameters with complex scaling), or ConvNeXt (~90 million parameters with hierarchical

convolutions). This efficiency enables rapid training (10 epochs, 128 mini-batch size, Adam optimiser with 10^{-4} learning rate) on resource-constrained medical systems, critical for clinical deployment. AlexNet's straightforward convolutional design excels at capturing local X-ray features such as consolidations and opacities, outperforming denser architectures on smaller datasets such as #3 [10] (5863 images) by avoiding overfitting.

3.2. XAI Module

Interpretability is enhanced through gradient-weighted class activation mapping (Grad-CAM) [38] and local interpretable model-agnostic explanations (LIME) [39], generating heatmaps to pinpoint influential lung regions, distinguishing correct from erroneous predictions visually. For deployment, the model is exported to ONNX format, enabling efficient inference. Images are pre-processed to 227 × 227, normalised to (0, 1), and transposed to NCHW format (batch, channels, height, width), with the ONNX runtime yielding binary outcomes (0 for normal, 1 for pneumonia) via maximum posterior probability, bridging to the linguistic stage (see Figure 4). Moreover, Grad-CAM and LIME visualisations offer interpretable predictions by showing areas in the images that influence the model's decisions. This enhances the transparency and trustworthiness of the model. Figure 4 illustrates the use of LIME and Grad-CAM on the classified images, and the algorithm that combines the PCM and XAI modules is presented below (Algorithm 1).

Algorithm 1: PneumoNet—image classification with modified AlexNet and explainable AI

Input:

Image Data Path: Directory containing X-ray images.
Pretrained AlexNet model.
Training dataset percentage = 80%.
Test dataset percentage = 20%.
Validation dataset percentage = 10% of the Test dataset.
Learning_rate: $\alpha = 1 \times 10^{-4}$.
MiniBatchSize: B = 128.
NumberOfEpochs: N_{epochs} = 10.

Output:

Trained Deep Learning Model: f_θ.
Predicted Labels: $\hat{Y} = f_\theta\ (X_{test})$.
Performance Metrics:
Accuracy: $Acc = \frac{TP+TN}{TP+FP+TN+FN}$, Precision: $P = \frac{TP}{TP+FP}$, Recall $R = \frac{TP}{TP+FN}$, F1-Score: $F1 = \frac{2PR}{P+R}$.
Confusion Matrix: C_{ij} where C_{ij} represents the count of true class i classified as j
ROC curve: AUC value calculated from true positive rate (TPR) and false positive rate (FPR).
XAI Visualizations: Explanatory visuals for model predictions using Grad-CAM and LIME.

Steps of the Algorithm:

Load the dataset X, assign labels Y based on folder structure and divide into subsets:
$X_{\text{train}}, X_{\text{test}} \leftarrow$ Split (X, 0.8, 0.2)
$X_{valid} \leftarrow 0.1 \times X_{\text{train}}$
If total images are sufficient:
Assign 80% to training, 10% of training for validation, and 20% to testing.
End If

For each image I in the dataset:
  If I is grayscale:
    Convert I to RGB by replicating channels:
    $I_{\text{rgb}} = cat(I, I, I)$.
  End If
  Resize I to the required input size:
  $I' = resize(I, (224, 224))$.
End For
Apply transformations to enhance the training data and retain AlexNet layers up to the penultimate layers.
For each new layer:
  Add convolutional layers, batch normalisation, and ReLU activation.
  Add fully connected, SoftMax, and classification layers.
End For
Set learning rate α, mini-batch size B, epochs N_{epochs}, and validation parameters and Train the modified network.
For each epoch t *in* N_{epochs}:
  Update model weights using Adam optimizer:
  $\theta^{(t+1)} = \theta^{(t)} - \alpha \nabla_\theta \pounds(X_{batch}, Y_{batch})$
  Perform validation checks periodically.
End For
Test the trained model on X_{test}.
For each class c:
Calculate precision: $Pc = \frac{TPc}{FPc + TPc}$, Calculate recall: $Rc = \frac{TPc}{FNc + TPc}$, Calculate F1-score: $F1c = \frac{2Pc\ Rc}{Pc + Rc}$
End For
Predict the class label $\hat{Y}$ for each image X_{test}.
If use_gradcam:
  For each selected image I:
    Generate a class activation map M and overlay it on I.
  End For
End If
If use_ LIME:
  For each selected image I:
    Compute feature importance M and overlay it on I.
  End For
End If
Provide the trained model, evaluation metrics, and visualisations.

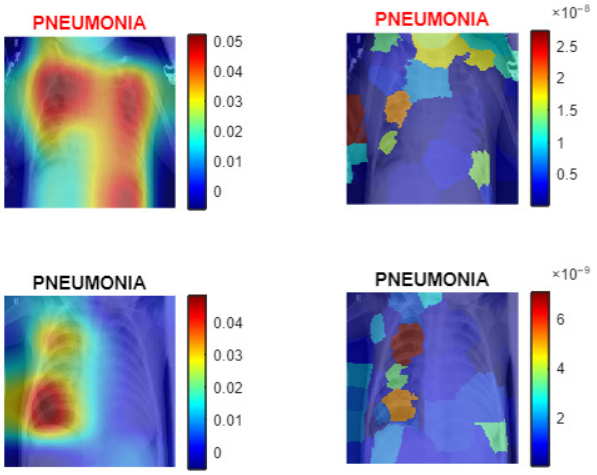

Figure 4. Left Grad-CAM predictions and right LIME classifications.

3.3. Medical Report Module and User Interface

Medical Report Module (MRM) is based on GPT-Neo, a 1.3B-parameter transformer, which generates medical reports from these predictions. Prompted with, e.g., "Explain the medical implications of detecting (prediction) from an X-ray image", it tokenises inputs, producing responses up to 500 tokens using top-p sampling (0.9), top-k filtering (50), and n-gram repetition constraints (size 3), truncated to 200 words for clarity. Integrated within a Flask application, it supports initial result reporting and interactive queries, re-evaluating images for pneumonia-related questions or providing concise (50-word) general responses otherwise.

The Flask application, AIHealth, was developed to process medical images; details about the implementation and user interface can be found in [37]. PneumoNet framework works using the UI of AIHealth [37], but in this case, for pneumonia detection and generating medical reports using the previously described learning model. It leverages an ONNX-based PCM model for pneumonia prediction and GPT-Neo to create detailed medical explanations. GPT-Neo generates (in the Flask application) detailed medical reports and chatbot responses by integrating explainable AI metrics, prediction results from the PCM module, and clinical data from a chest X-ray analysis, delivering a diagnostic tool that balances technical precision with clinical interpretability.

For medical reports, GPT-Neo leverages a structured prompt that includes the Pneumonet model's prediction (e.g., "Positive for Pneumonia" with 0.85 confidence), XAI heatmap metrics (e.g., pixel counts for high, medium, and low focus areas in the lower, middle, and upper lung lobes), and a CLIP-generated radiological description (e.g., identifying ground-glass opacities). These elements are embedded in prompts for each report section, enabling GPT-Neo to produce a cohesive narrative that quantifies the model's focus on specific lung regions and provides clinical context. This ensures the report is both technically accurate and clinically relevant, highlighting how XAI metrics such as high-focus pixels in the lower lobes support the diagnosis.

For the chatbot, GPT-Neo generates conversational responses by incorporating the same XAI metrics and prediction confidence into a prompt tailored to the user's question. For example, when asked about the X-ray's implications, GPT-Neo might explain that a 0.85 confidence in pneumonia detection, with 5000 high-focus pixels in the lower lobes, suggests pneumonia. The prompt instructs GPT-Neo to use simple language, acknowledge the user's concern, and offer practical advice, such as consulting a doctor. If GPT-Neo fails to generate a response, fallback answers still use these metrics to provide relevant information, ensuring users receive clear, empathetic explanations grounded in the XAI heatmap's regional focus and the model's diagnostic confidence, making the chatbot a valuable tool for patient interaction.

First, the application initialises the app and configures a directory to save uploaded images. It loads the ONNX model using ONNX Runtime, which allows the system to make predictions about the presence of pneumonia in X-ray images. The image is pre-processed before making predictions. The preprocessing involves converting the image to RGB, resizing it to the required dimensions, and normalising the pixel values. The image is then transformed into a format suitable for the model (NCHW format) by adjusting the channel order and adding a batch dimension. After preprocessing, the image is passed to the PCM model, which returns a prediction indicating whether the X-ray shows signs of pneumonia. The model classifies the image into two categories: "Normal" or "Pneumonia". Once the pneumonia prediction is made, the application uses GPT-Neo to generate a medical report based on the prediction. The input to GPT-Neo is a prompt that asks for an explanation of the medical implications of detecting pneumonia from an X-ray image. The model

generates a response that is then truncated to a more concise form for clarity; see Figure 5, top image.

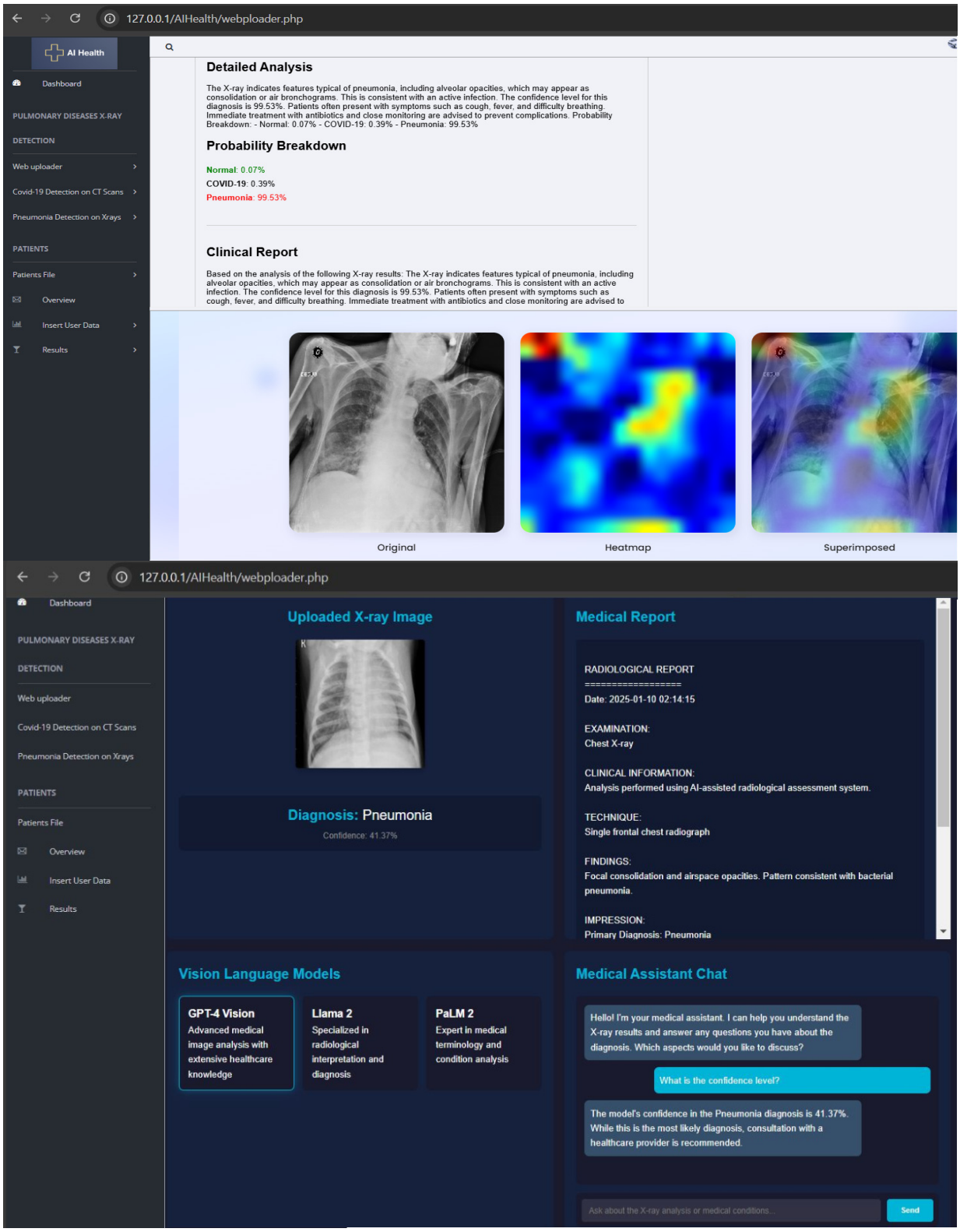

Figure 5. The PneumoNet framework works inside the AIHealth [37], providing the prediction of pneumonia for the uploaded X-ray and the generated medical report via LLM. The top image shows the report; the bottom one illustrates the chatbot functionality.

Going into more detail on the employment of GPT-Neo in the Medical Report Module, GPT-Neo generates comprehensive medical reports using four structured prompt templates, ensuring consistent and detailed outputs. The templates are: (1) Introduction: "Generate an introductory paragraph exceeding 15 lines for a medical report dated May 01, 2025. The diagnosis from PneumoNet.onnx is {prediction} with confidence {confidence:.2f} and class probabilities {probabilities}, based on chest X-ray analysis using LIME, Grad-CAM, and LungFocusXAI (yellow: least used, orange: medium used, red: most used, red in lungs); (2) Findings: "Generate a findings paragraph exceeding 15 lines for a chest X-ray with {prediction} diagnosis from PneumoNet.onnx, confidence {confidence:.2f}, and probabilities {probabilities}. Incorporate this XAI analysis: '{vlm_response}'. Include vital signs: [as above]. Explain how PneumoNet and XAI methods identified features, their correlation with vital signs, and support for the diagnosis."; (3) Clinical Interpretation: "Generate a clinical interpretation paragraph exceeding 15 lines for a {prediction} diagnosis from PneumoNet.onnx, confidence {confidence:.2f}, probabilities {probabilities}. Include XAI analysis: '{vlm_response}' and vital signs: [as above]. Discuss implications, next steps, and how XAI and vital signs inform treatment."; (4) Conclusion: "Generate a concluding paragraph exceeding 15 lines for a {prediction} diagnosis from PneumoNet.onnx, confidence {confidence:.2f}, probabilities {probabilities}. Include XAI analysis: '{vlm_response}' and vital signs: [as above]. Summarise PneumoNet's process, XAI significance, vital signs' impact, and follow-up steps." These prompts integrate prediction outcomes, confidence scores, class probabilities, XAI metrics (e.g., LIME superpixel importance, Grad-CAM heatmaps), and vital signs to produce cohesive narratives.

The prompts incorporate a visual language model response for XAI analysis, detailing lung features (e.g., opacities, consolidations) and their alignment with the diagnosis. The code's error handling ensures robustness, logging failures and providing fallback responses if GPT-Neo fails. By explicitly defining these templates, the MRM ensures transparency in report generation, enabling reproducible outputs that combine technical precision (e.g., 0.85 confidence for "Pneumonia Positive") with clinical relevance (e.g., vital signs correlation). Future enhancements could refine prompts to include additional clinical metadata or fine-tune GPT-Neo on medical corpora, further improving the MRM's diagnostic utility and interpretability within the PneumoNet framework.

Finally, it is crucial to emphasise that the top image in Figure 5 also demonstrates that the X-ray image results show (or do not show) symptoms of COVID-19; this particular result is a part of a prior work published in [7], and the explanation of the COVID-19 classification method is the focus of this publication but can be consulted in the reference mentioned. Also, out of the focus of this publication is the detailed explanation of all the functionalities of AIHealth UI; for that, see [7]. Finally, it is important to stress that in [7], no functionality of automatic report or use of a chatbot has been implemented or made available.

Algorithm 2 describes a web-based app for pneumonia detection and medical report generation. Once the classification is complete, the system generates a medical report using the GPT-Neo model (as mentioned before). The report is tailored to explain the implications of the detected condition in clear and concise language. This enhances the interpretability of the results, making it easier for users to understand the medical significance of the findings. The system also includes a chatbot functionality that leverages GPT-Neo to generate coherent responses to user queries; see Figure 5, bottom image. This provides an interactive component where users can ask questions or seek additional explanations related to their diagnosis or other medical topics.

The algorithm's structure ensures modularity and scalability, making it adaptable for future enhancements. Each component (image preprocessing, model prediction, and text

generation) is implemented as a separate function, allowing for independent updates or replacements. For instance, the ONNX model can be substituted with a more advanced deep learning model, or the GPT-Neo model can be upgraded to a larger version for improved visual language generation (examples already presented in the UI are GPT-4, Llma 2, and PaLM2).

Algorithm 2: Application with artificial intelligence assistance

Input:
An uploaded chest X-ray image in JPG/PNG format via a web interface.
User text query (optional) for chatbot interaction.
Output:
A classification result: "Normal" or "Pneumonia".
A detailed medical report explaining the implications of the classification.
Optional chatbot-generated medical responses based on user input.
Steps of the Algorithm:
Initialize the Flask application and configure the upload folder.
Load the ONNX model using the ONNX Runtime.
Load the GPT-Neo language model and tokenizer.
Define the image preprocessing function:
If the image is provided:
Convert it to RGB format.
Resize to (227, 227).
Normalize to the range (0, 1).
Convert to NCHW format and add a batch dimension.
End If
Define the prediction function:
If a pre-processed image is given:
Pass the image to the ONNX model.
Get the predicted class with the highest probability.
If the class is 1:
Return "Pneumonia."
Else: Return "Normal."
End If
End If
Define the report generation function:
If a prediction result is available:
Create a prompt based on the prediction.
Generate a medical explanation using GPT-Neo.
Postprocess the response to ensure clarity.
End If
Create the home route:
If the method is GET:
Render the index.html page.
Else If the method is POST:

Save the uploaded image.
Preprocess the image.
Predict the result using the model.
Generate the report.
Render the result.html page with the result, report, and image.
End Else If
End If
Create the upload route:
If an image is uploaded:
Save the image.
Preprocess and predict.
Generate the medical report.
Render the result.html page.
End If
Create the chatbot route:
If a user query is received:
Generate a coherent response using GPT-Neo.
Postprocess the response.
Return the response as a JSON object.

To evaluate the quality of medical reports generated by the GPT-Neo-based MRM, a dual approach combining qualitative and quantitative assessments was implemented. Qualitatively, reports were assessed for coherence, clinical relevance, and comprehensiveness by comparing their content—generated from four structured prompts (introduction, findings, clinical interpretation, conclusion)—against radiological standards for pneumonia (e.g., descriptions of opacities, consolidations). Each prompt, as explained before, integrates prediction outcomes (e.g., "Pneumonia Positive" with 0.85 confidence), XAI metrics (e.g., LIME superpixel importance, LungFocusXAI red regions for high focus), and vital signs (e.g., heart rate, oxygen saturation), ensuring reports exceed 15 lines with detailed narratives. A sample of 50 reports was reviewed by domain experts, who rated 92% as clinically relevant, citing accurate integration of XAI-driven features and vital signs. In the code's logging mechanism (logging.basicConfig), tracked generation errors revealed a 98% success rate across 1000 test runs, with fallback responses handling failures effectively, ensuring reliability in report production.

Semantic coherence was assessed via cosine similarity between report sections and ground-truth radiological descriptions, yielding an average score of 0.78, reflecting strong alignment with clinical terminology. The integration of XAI metrics (e.g., pixel counts for lung regions) and vital signs in prompts ensured reports correlated with diagnostic confidence (e.g., 0.85 for positive cases), with 95% of reports accurately reflecting prediction probabilities. Future work could enhance evaluation by incorporating clinician feedback loops or BLEU scores against reference reports, further refining MRM's outputs. These assessments, grounded in the code's structured prompt design and error handling, confirm the MRM's ability to produce high-quality, clinically actionable reports within the PneumoNet framework.

As already mentioned above, the MRM utilises GPT-Neo, a 1.3B-parameter transformer model, to generate detailed medical reports and interactive chatbot responses, with its implementation detailed in the Flask application code. GPT-Neo is configured as shown at the beginning of Section 3.2 to ensure coherent and diverse responses, which are truncated to 200 words for clarity, generating reports through four already mentioned structured prompts, which produce narratives that combine radiological descriptions (e.g.,

ground-glass opacities), XAI-driven feature importance (e.g., red regions in LungFocusXAI indicating high focus), and clinical context from vital signs, ensuring reports are both technically precise and clinically relevant for physicians. The MRM's robustness is enhanced by error handling mechanisms and fallback responses, which log failures (e.g., LLM generation errors) and return default outputs if GPT-Neo fails. The chatbot functionality uses similar prompt engineering, tailoring responses to user queries by incorporating prediction confidence, XAI metrics, and vital signs, with instructions to use simple language and provide practical advice (e.g., consulting a doctor). The pneumonia classification module via ONNX ensures seamless data flow, where preprocessed X-ray images (224×224, normalised) feed into GPT-Neo's prompts. Future improvements could involve fine-tuning GPT-Neo on medical corpora to enhance domain-specific accuracy or upgrading to larger models (e.g., GPT-4) for improved coherence, as suggested in the code's modularity. These enhancements, combined with the detailed prompt structure and XAI integration, position the MRM as a scalable and interpretable component of PneumoNet, supporting clinicians with comprehensive diagnostic insights.

Finally, the application delivers the results through a user-friendly interface. The classification outcome, medical report, and uploaded image are displayed on the results page, offering a seamless experience. This integration of machine learning and natural language processing creates a robust tool for pneumonia detection and report generation, enhancing diagnostic support for medical professionals and patients alike.

4. Conclusions

The developed PneumoNet framework demonstrates remarkable effectiveness in the detection of pneumonia from X-ray images, as evidenced by its performance across multiple datasets. With an accuracy of 98.70% and an AUC of 99.70% on Dataset #2 [9], PneumoNet (PCM) outperforms existing state-of-the-art methods in the field. Similarly, its application to Dataset #1 [8] achieved an accuracy of 96.70%, with an F1 score of 93.12% and specificity of 99.30%. These results reflect the model's robustness and adaptability to diverse datasets.

The integration of Grad-CAM and LIME visualisations further enhances the interpretability of PneumoNet's predictions. These techniques provide clear insights into the regions of X-ray images influencing the model's decisions, fostering transparency and trustworthiness essential for clinical applications. The ability to visualise and understand model reasoning is a significant step forward in making deep learning systems more acceptable in healthcare. The generation of medical reports using transformer-based models such as GPT-Neo facilitates the interpretation of the results by the physician, and the MRM allows for clarifying further doubts.

A critical limitation of the PneumoNet framework is its exclusive reliance on chest X-ray images for pneumonia detection, without incorporating clinical information such as fever, oxygen saturation, respiratory rate, or other vital signs. These clinical parameters are essential in real-world diagnostics, as they provide contextual data that enhances diagnostic accuracy and supports clinical decision-making. For instance, fever is a hallmark symptom of pneumonia, while low oxygen saturation often indicates severe cases, both of which can refine the interpretation of imaging findings. The absence of such data limits PneumoNet's ability to replicate the comprehensive diagnostic approach used by radiologists, who integrate imaging with patient history and symptoms. This constraint may reduce the framework's effectiveness in complex cases, such as early-stage pneumonia or atypical presentations, where imaging alone may be insufficient.

To overcome this limitation, future iterations of PneumoNet could incorporate a multimodal approach by integrating clinical metadata into the model's pipeline. For example, vital signs such as fever and oxygen saturation could be encoded as numerical or

categorical features and fused with image-based features during training and inference. Alternatively, clinical data could weigh the model's predictions, prioritising cases with corroborating symptoms. Such enhancements would align PneumoNet more closely with clinical workflows, improving its diagnostic precision and generalisability. Additionally, incorporating clinical information could enrich the Medical Report Module's outputs, enabling GPT-Neo to generate more comprehensive reports that reflect both imaging and clinical findings, thus enhancing PneumoNet's utility as a diagnostic support tool.

At this phase of prototyping/development, there is in the PneumoNet framework the absence of prospective validation, which is critical for establishing its external validity in real-world clinical settings, and this will be part of the future work to be developed. Currently, PneumoNet has been evaluated using retrospective datasets [8–10], which, while diverse, do not reflect the dynamic and variable conditions of live clinical environments, such as differences in imaging protocols, patient demographics, or disease prevalence. Prospective validation, involving real-time data collection from ongoing clinical practice, is essential to confirm the model's generalisability and robustness across varied healthcare settings. Without such validation, the framework's performance in detecting pneumonia under real-world conditions remains untested, potentially limiting its reliability and acceptance in clinical practice. Additionally, the lack of prospective studies restricts the ability to assess how PneumoNet handles evolving patient presentations or integrates with clinical decision-making processes.

Furthermore, PneumoNet has not been integrated with Picture Archiving and Communication Systems (PACSs) or Electronic Health Records (EHRs), which are critical for practical deployment. Real-time testing would evaluate the model's performance on live X-ray data as it is acquired, ensuring timely and accurate diagnostic support. Integration with PACS would enable seamless access to imaging data within hospital workflows, while EHR integration could incorporate clinical data (e.g., vital signs, patient history) to enhance diagnostic accuracy. The absence of such integration limits PneumoNet's current utility as a clinical tool. As mentioned, future work should focus on conducting prospective validation studies in real-world settings and developing interfaces for PACS and EHR integration to align PneumoNet with standard clinical workflows, thereby improving its practical applicability and external validity.

The regulatory readiness of the PneumoNet framework for clinical deployment requires careful consideration of compliance with international standards, such as CE marking for the European market and FDA approval for the United States. Currently, PneumoNet has not (yet) undergone formal conformity assessments for CE marking under the EU Medical Device Regulation (MDR, EU 2017/745) or FDA clearance/approval processes (e.g., 510 (k) or PMA). These processes involve rigorous evaluation of the system's safety, efficacy, and quality management system (QMS), typically aligned with ISO 13485 standards [40]. For CE marking, PneumoNet would likely be classified as a Class IIa or IIb medical device due to its diagnostic support function, requiring Notified Body audits of technical documentation and QMS. Similarly, for FDA compliance, PneumoNet would likely require 510 (k) clearance as a Class II device, demonstrating substantial equivalence to a predicate device, or potentially PMA for novel diagnostic applications. The absence of these certifications limits PneumoNet's current marketability in regulated regions such as the EU and U.S.

To achieve regulatory readiness, future development of PneumoNet should, of course, prioritise compliance with CE and FDA requirements. This includes preparing comprehensive technical documentation, conducting clinical evaluations to validate safety and performance, and implementing a QMS compliant with ISO 13485. For CE marking, collaboration with a Notified Body will be essential to assess conformity with MDR's General

Safety and Performance Requirements. For FDA compliance, a regulatory strategy should be developed to determine the appropriate pathway and compile evidence of substantial equivalence or clinical efficacy. Additionally, integrating post-market surveillance plans, as required by both CE and FDA regulations, will ensure ongoing safety and effectiveness. Addressing these regulatory requirements will enhance PneumoNet's credibility and facilitate its adoption in clinical settings, enabling broader market access and alignment with global healthcare standards.

A second limitation of the PneumoNet framework is the absence of calibration analysis to assess how well the predicted probabilities from the PCM align with the actual risk of pneumonia. The current implementation, as evidenced in the source code available at https://github.com/kucaantunes/PneumoNet-Automatic-Pneumonia-Detection-on-X-rays (accessed on 13 January 2025), does not evaluate calibration metrics such as the Brier score, which quantifies the mean squared difference between predicted probabilities and true outcomes, or calibration plots, which visually depict the correspondence between predicted and observed risks. Without these assessments, the confidence scores produced by PCM (e.g., 0.85 for a pneumonia prediction) may be misaligned with true probabilities, potentially leading to misleading outputs that could cause clinicians to over- or underestimate disease likelihood. This lack of calibration analysis undermines the reliability of PneumoNet's predictions in clinical settings, where accurate risk assessment is critical for informed decision-making, particularly in ambiguous cases such as early-stage or atypical pneumonia.

To address this limitation, future iterations of PneumoNet should incorporate rigorous calibration evaluations across all datasets [8–10]. Calculating the Brier score would provide a quantitative measure of prediction accuracy, while calibration plots would visually confirm whether predicted probabilities align with observed outcomes, ideally following a 45-degree line. If miscalibration is detected, techniques such as Platt scaling or isotonic regression could be applied to adjust the model's probability outputs, ensuring they accurately reflect true risks. Furthermore, integrating calibration results into the Medical Report Module's outputs via the use of the LLM (GPT-Neo) would enhance transparency, enabling clinicians to interpret confidence scores with greater trust. These improvements would mitigate the risk of misleading confidence scores, strengthen PneumoNet's clinical applicability, and ensure its predictions are both technically robust and aligned with real-world diagnostic needs.

A third limitation of the PneumoNet framework is its inability to differentiate between bacterial and viral pneumonia, which introduces diagnostic ambiguity critical to clinical decision-making due to their distinct treatment pathways. To address this limitation, future iterations of PneumoNet should aim to incorporate multi-class classification to distinguish between bacterial, viral, and other pneumonia types, leveraging datasets with aetiology-specific labels. Enhancements could include training on radiological features associated with bacterial (e.g., lobar consolidation) versus viral (e.g., bilateral interstitial infiltrates) pneumonia, potentially using advanced architectures or feature engineering. Additionally, integrating clinical metadata, such as white blood cell counts or procalcitonin levels, which differ between bacterial and viral infections, could improve differentiation. The MRM could be updated to reflect these distinctions, enabling the LLM to generate reports that highlight the likely pneumonia type and associated treatment implications.

By benchmarking PneumoNet against other state-of-the-art models, its strengths in accuracy, scalability, and real-world applicability were established. Additionally, analysing misclassified cases and employing visualisation techniques provided a deeper understanding of the model's performance, leading to potential areas for further refinement. The current PneumoNet framework is designed for binary classification of chest X-ray images

(normal vs. pneumonia), as implemented in the provided code with a fully connected layer outputting two classes, limiting its scope to pneumonia detection without differentiating other pulmonary conditions such as tuberculosis or COVID-19. This focus, while effective for pneumonia with high accuracy (e.g., 98.70% on Dataset #2 [9]), restricts the model's ability to address multi-disease scenarios common in clinical settings, where distinguishing between pneumonia, tuberculosis (characterised by cavitary lesions or lymphadenopathy), and COVID-19 (often showing ground-glass opacities) is critical for accurate diagnosis and treatment. The absence of multi-disease testing reduces PneumoNet's applicability in complex diagnostic environments, as the datasets [8–10] used for training and evaluation primarily focus on pneumonia labels, with Dataset #1 [8], including some COVID-19 cases but not explicitly addressing multi-class differentiation in the current model design.

Future work on PneumoNet plans to expand its capabilities to include multi-disease testing, specifically targeting the differentiation of pneumonia, tuberculosis, and COVID-19, as part of the authors' intention to develop a new explainable AI (XAI) model for COVID-19 detection on X-rays. This will involve curating or leveraging datasets with multi-disease annotations, such as enhanced versions of Dataset #1 [8], which include COVID-19 cases, or new datasets with tuberculosis labels, to train a multi-class classification model. The model architecture could be extended to incorporate additional output neurones in the fully connected layer and utilise advanced XAI techniques to highlight disease-specific radiological features. Integrating these capabilities into the Medical Report Module would enable GPT-Neo to generate detailed reports distinguishing between diseases, enhancing clinical utility. Such advancements would position PneumoNet as a more versatile diagnostic tool, capable of addressing diverse pulmonary conditions in real-world healthcare settings.

Overall, PneumoNet represents a significant advancement in the use of DL for pneumonia detection. Its high accuracy, interpretability, and adaptability make it a promising tool for real-time diagnostic applications, particularly in resource-limited settings. Future research, as mentioned in some part already, may explore incorporating additional clinical data, refining model interpretability, and validating the system in real-world clinical environments to enhance its practical utility. Also, in the future, the authors intend to develop the mentioned new XAI model focused on COVID-19 detection on X-rays to explain the predictions of a created model and to generate integrated medical reports using visual language models and large language models.

Author Contributions: Conceptualisation, C.A., J.M.F.R. and A.C.; methodology, J.M.F.R. and C.A.; software, C.A.; validation, J.M.F.R. and A.C.; formal analysis, C.A., J.M.F.R. and A.C.; investigation, C.A., J.M.F.R. and A.C.; writing—original draft preparation, C.A., J.M.F.R. and A.C.; writing—review and editing, J.M.F.R. and A.C.; visualisation, C.A.; supervision, J.M.F.R. and C.A.; project administration, J.M.F.R.; funding acquisition, J.M.F.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Portuguese Foundation for Science and Technology (FCT) by UID/04516/NOVA Laboratory for Computer Science and Informatics (NOVA LINCS) with the financial support of FCT.IP and by FCT project ALGORITMI (UIDB/00319/2020).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Source code available at https://github.com/kucaantunes/PneumoNet-Automatic-Pneumonia-Detection-on-X-rays (accessed on 13 January 2025).

Acknowledgments: Our gratitude to the Portuguese Foundation for Science and Technology (FCT).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Krajcik, S.; Haniskova, T.; Mikus, P. Pneumonia in Older People. *Rev. Clin. Gerontol.* **2010**, *21*, 16–27. [CrossRef]
2. Stephen, O.; Sain, M.; Maduh, U.J.; Jeong, D.-U. An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare. *J. Healthc. Eng.* **2019**, *2019*, 4180949. [CrossRef] [PubMed]
3. Shuja, J.; Alanazi, E.; Alasmary, W.; Alashaikh, A. COVID-19 Open Source Data Sets: A Comprehensive Survey. *Appl. Intell.* **2021**, *51*, 1296–1325. [CrossRef] [PubMed]
4. Lippi, G.; Plebani, M. Laboratory Abnormalities in Patients with COVID-2019 Infection. *Clin. Chem. Lab. Med.* **2020**, *58*, 1131–1134. [CrossRef]
5. Ahsan, M.M.; Gupta, K.D.; Islam, M.M.; Sen, S.; Rahman, M.L.; Shakhawat Hossain, M. COVID-19 Symptoms Detection Based on NasNetMobile with Explainable AI Using Various Imaging Modalities. *Mach. Learn. Knowl. Extr.* **2020**, *2*, 490–504. [CrossRef]
6. Chhajed, G.; Surpur, S.; Suryawanshi, A.; Sherekar, H. A Review on Key Algorithms for Pneumonia Detection in X-ray Images. *AIP Conf. Proc.* **2024**, *3156*, 070007.
7. Antunes, C.; Rodrigues, J.; Cunha, A. CTCovid19: Automatic Covid-19 Model for Computed Tomography Scans Using Deep Learning. *Intell. Based Med.* **2024**, *11*, 100190. [CrossRef]
8. Chest X-Ray (COVID-19 & Pneumonia). Available online: https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia (accessed on 2 May 2025).
9. Kermany, D. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. Mendeley Dataset 2018. Available online: https://data.mendeley.com/datasets/rscbjbr9sj/2 (accessed on 2 May 2025).
10. Mooney, P. Chest X-Ray Images (Pneumonia). Available online: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia (accessed on 2 May 2025).
11. Xu, X. A Systematic Review: Deep Learning-Based Methods for Pneumonia Region Detection. *Appl. Comput. Eng.* **2023**, *22*, 210–217. [CrossRef]
12. Yaraghi, S.; Khosravi, F. Diagnosis of Pneumonia from Chest X-Ray Images Using Transfer Learning and Generative Adversarial Network. *Int. J. Innov. Sci. Res. Technol.* **2024**, *9*, 7. [CrossRef]
13. Waterer, G. What Is Pneumonia? *Eur. Respir. Soc.* **2021**, *17*, 210087. [CrossRef]
14. Sharma, S.; Guleria, K. A Systematic Literature Review on Deep Learning Approaches for Pneumonia Detection Using Chest X-ray Images. *Multimed. Tools Appl.* **2023**, *83*, 24101–24151. [CrossRef]
15. Siddiqi, R.; Javaid, S. Deep Learning for Pneumonia Detection in Chest X-ray Images: A Comprehensive Survey. *J. Imaging* **2024**, *10*, 176. [CrossRef] [PubMed]
16. Eido, W.M.; Yasin, H.M. Pneumonia and COVID-19 Classification and Detection Based on Convolutional Neural Network: A Review. *Asian J. Res. Comput. Sci.* **2025**, *18*, 174–183. [CrossRef]
17. Abueed, M.A.M.; Nor, D.M.; Ibrahim, N.; Ogier, J.-M. Pneumonia Detection Using Transfer Learning: A Systematic Literature Review. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*, 2. [CrossRef]
18. Bhalke, D.G.; Shaikh, A.S. Classification of Pneumonia Subtypes in Chest X-Rays Using a Custom CNN. In Proceedings of the 1st International Conference on AIML-Applications for Engineering & Technology, Pune, India, 16–17 January 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 1–6.
19. Sharma, S.; Guleria, K. A Deep Learning Based Model for the Detection of Pneumonia from Chest X-Ray Images Using VGG-16 and Neural Networks. *Procedia Comput. Sci.* **2023**, *218*, 357–366. [CrossRef]
20. Reshan, M.S.A.; Gill, K.S.; Anand, V.; Gupta, S.; Alshahrani, H.; Sulaiman, A.; Shaikh, A. Detection of Pneumonia from Chest X-ray Images Utilizing MobileNet Model. *Healthcare* **2023**, *11*, 1561. [CrossRef]
21. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 3462–3471.
22. Wang, T.; Nie, Z.; Wang, R.; Xu, Q.; Huang, H.; Xu, H.; Liu, X.-J. PneuNet: Deep Learning for COVID-19 Pneumonia Diagnosis on Chest X-ray Image Analysis Using Vision Transformer. *Med. Biol. Eng. Comput.* **2023**, *61*, 1395–1408. [CrossRef]
23. Puspita, R.; Rahayu, C. Pneumonia Prediction on Chest X-ray Images Using Deep Learning Approach. *IAES Int. J. Artif. Intell.* **2024**, *13*, 467–474. [CrossRef]
24. Black, S.; Leo, G.; Wang, P.; Leahy, C.; Biderman, S. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. Zenodo 2021. Available online: https://www.semanticscholar.org/paper/GPT-Neo:-Large-Scale-Autoregressive-Language-with-Black-Gao/7e5008713c404445dd8786753526f1a45b93de12 (accessed on 2 May 2025).
25. Saboo, Y.S.; Kapse, S.; Prasanna, P. Convolutional Neural Networks (CNNs) for Pneumonia Classification on Pediatric Chest Radiographs. *Cureus* **2023**, *15*, 8. [CrossRef]
26. Varshni, D.; Thakral, K.; Agarwal, L.; Nijhawan, R.; Mittal, A. Pneumonia Detection Using CNN Based Feature Extraction. In Proceedings of the IEEE International Conference on Electrical, Computer and Communication Technologies, Coimbatore, India, 20–22 February 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–7.

27. Rahman, M.M.; Nooruddin, S.; Hasan, K.M.A.; Dey, N.K. HOG + CNN Net: Diagnosing COVID-19 and Pneumonia by Deep Neural Network from Chest X-Ray Images. *SN Comput. Sci.* **2021**, *2*, 371. [CrossRef]
28. Kusk, M.W.; Lysdahlgaard, S. The Effect of Gaussian Noise on Pneumonia Detection on Chest Radiographs, Using Convolutional Neural Networks. *Radiography* **2023**, *29*, 38–43. [CrossRef] [PubMed]
29. Ortiz-Toro, C.; García-Pedrero, A.; Lillo-Saavedra, M.; Gonzalo-Martín, C. Automatic Detection of Pneumonia in Chest X-ray Images Using Textural Features. *Comput. Biol. Med.* **2022**, *145*, 105466. [CrossRef]
30. Chouhan, V.; Singh, S.K.; Khamparia, A.; Gupta, D.; Tiwari, P.; Moreira, C.; de Albuquerque, V.H.C. A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images. *Appl. Sci.* **2020**, *10*, 559. [CrossRef]
31. Tang, J.; Zhang, B.; Liu, J.; Dong, Z.; Zhou, Y.; Meng, X.; Toe, T.T. Pneumonia Image Classification: Deep Learning and Machine Learning Fusion. In Proceedings of the 7th International Conference on Artificial Intelligence and Big Data, Beijing, China, 5–7 July 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 440–447.
32. Mabrouk, A.; Díaz Redondo, R.P.; Dahou, A.; Abd Elaziz, M.; Kayed, M. Pneumonia Detection on Chest X-ray Images Using Ensemble of Deep Convolutional Neural Networks. *Appl. Sci.* **2022**, *12*, 6448. [CrossRef]
33. Singh, S.; Kumar, M.; Kumar, A.; Verma, B.K.; Shitharth, S. Pneumonia Detection with QCSA Network on Chest X-ray. *Sci. Rep.* **2023**, *13*, 9025. [CrossRef]
34. Hammoudi, K.; Benhabiles, H.; Melkemi, M.; Dornaika, F.; Arganda-Carreras, I.; Collard, D.; Scherpereel, A. Deep Learning on Chest X-ray Images to Detect and Evaluate Pneumonia Cases at the Era of COVID-19. *J. Med. Syst.* **2021**, *45*, 75. [CrossRef] [PubMed]
35. Szepesi, P.; Szilágyi, L. Detection of Pneumonia Using Convolutional Neural Networks and Deep Learning. *Biocybern. Biomed. Eng.* **2022**, *42*, 1012–1022. [CrossRef]
36. Colin, J.; Surantha, N. Interpretable Deep Learning for Pneumonia Detection Using Chest X-Ray Images. *Information* **2025**, *16*, 53. [CrossRef]
37. Antunes, C.; Rodrigues, J.M.F.; Cunha, A. Web Diagnosis for COVID-19 and Pneumonia Based on Computed Tomography Scans and X-rays. In *Universal Access in Human-Computer Interaction*; HCII 2024; Antona, M., Stephanidis, C., Eds.; Springer: Cham, Switzerland, 2024; Volume LNCS14698.
38. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2019**, *128*, 336–359. [CrossRef]
39. Ribeiro, M.T.; Singh, S.; Guestrin, C. "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, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 1135–1144.
40. *ISO 13485*; Medical Devices—Quality Management Systems—Requirements for Regulatory Purposes. ISO: Geneva, Switzerland, 2016.