# Comparison of binary classification techniques for Pneumonia detection from Chest X-ray images.

**Lorenzo Spataro** spataro.1946590@studenti.uniroma1.it
**Marco Realacci** realacci.1938880@studenti.uniroma1.it
**Federico Raponi** raponi.1963339@studenti.uniroma1.it
**Gioele Maria Zoccoli** zoccoli.1850491@studenti.uniroma1.it
**Alessio Pannozzo** pannozzo.1960374@studenti.uniroma1.it

## Abstract

This work presents a comprehensive evaluation of various approaches for classifying pneumonia from chest X-ray images, emphasizing both traditional machine learning and deep learning methods.

A detailed cross-dataset analysis was conducted to assess the generalizability and robustness of these approaches across different datasets, a crucial aspect in medical image analysis. The inclusion of cross-dataset evaluations highlights the models' ability to adapt to varying data distributions, ensuring their applicability in real-world clinical scenarios. This study provides valuable insights into the strengths and limitations of both traditional and deep learning methods, guiding the development of more reliable AI-driven diagnostic tools for pneumonia detection.

## 1 Introduction

In recent years, the combination of technology and medicine has revolutionised the diagnosis and treatment of many diseases. Pneumonia, the leading cause of infectious death among children under five, caused more than 800,000 deaths in 2017, accounting for 15% of global child mortality (1). The most affected regions are South Asia and sub-Saharan Africa. Prevention and early diagnosis are crucial to significantly reduce fatalities, especially in resource-limited settings where access to vaccines and treatment is often insufficient.

Advanced machine learning methods in medical image analysis hold great promise for improving disease diagnosis, including pneumonia. However, addressing challenges in standardization and adapting these technologies to diverse clinical settings remains critical for their widespread adoption and effectiveness.

## 2 Related Work

In recent years, computer-aided diagnosis has revolutionised medical image analysis. Initially, traditional machine learning methods were based on manually extracted features such as histograms and pixel variations, but the advancement of deep neural networks has enabled significantly better results.

Deep learning-based techniques, particularly convolutional networks (CNNs), have demonstrated superior capabilities in identifying and classifying affected regions. For example, recent studies presented models capable of detecting anomalies with high accuracy, often outperforming human experts (2).

A 2024 systematic review highlights how these approaches have improved efficiency and diagnostic accuracy, discussing advantages and limitations of existing methodologies and proposing future directions, such as the integration of multimodal models to combine clinical and imaging data (3). However, challenges remain, such as model explainability and handling bias in datasets, which affect reliability in clinical applications.

## 3 Proposed method

The study investigates cross-domain generalization in the context of diagnosing pneumonia from chest X-ray images. Specifically, the analysis focuses on evaluating whether models trained on a particular dataset can generalize effectively to unseen datasets, a critical challenge in medical imaging where data distribution shifts are common. To this end, various machine learning and deep learning architectures were employed.

Each model was trained on one dataset designated as the training set and subsequently tested on distinct, unseen datasets to assess its cross-domain generalization capabilities.

Model performance was benchmarked using a

comprehensive suite of evaluation metrics, including *accuracy*, *precision*, *recall*, *F1-score*, and the area under the Receiver Operating Characteristic curve (*ROC_AUC*). These metrics were selected to provide a holistic understanding of each model's classification performance across different aspects.

## 4 Models

### 4.1 Traditional models

The models employed in this study include K-Nearest Neighbors (KNN) (4), Support Vector Machines (SVM) (5), Binary Decision Trees (6), and Logistic Regression (7).

To ensure robust and consistent preprocessing across models, we implemented a pipeline comprising random undersampling to address class imbalance, standard scaling for feature normalization, and Principal Component Analysis (8) for dimensionality reduction. Hyperparameter optimization was performed using a randomized search strategy with k-fold cross-validation(9), aiming at maximizing the precision.

### 4.2 CNN

We used a simple convolutional neural network (CNN), with the architecture described in (10). Hyperparameter optimization was performed using Weight&Biases[1] with Sweeps, selecting the best configuration from 24 different models (Fig. 1).

The final architecture consists of six blocks, each with a 2D convolutional layer (kernel size 3×3) followed by a max-pooling layer (pool size 2×2). The number of filters increases progressively, with sizes 3, 6, 9, 12, 15, and 16 in the final block. ReLU activation is applied to all convolutional layers. The output is flattened and passed through two dense layers (64 neurons each) with ReLU activation. The model was trained for 50 epochs using the Adam optimizer (learning rate 0.0001), a batch size of 64, and a dropout rate of 0.4. The loss function used was the weighted cross-entropy loss.

To enhance generalization to unseen data, augmentation techniques were employed during training.

[1]Weight&Biases site: `https://wandb.ai/site/`

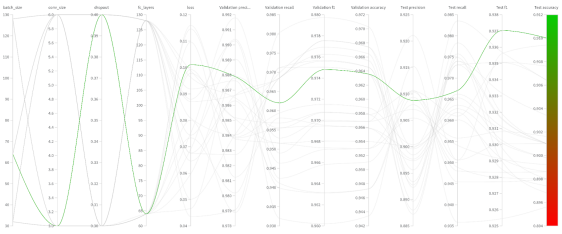

Figure 1: Weight&Biases - CNN

### 4.3 ResNet50

As transfer learning is a good strategy for the analysis of medical images (11), we decided to use the ResNet50 architecture pre-trained on the ImageNet-1K dataset (12). ResNet50 is a deep convolutional neural network characterized by its residual blocks (13), which allow the model to train very deep architectures without the gradient vanishing problem that typically occurs in traditional deep networks. These residual connections help to propagate gradients through the network more effectively, making it possible to train very deep models and achieving remarkable performance on a variety of tasks, including image classification.

The pre-trained ResNet-50 model has been adapted, replacing the original output layer with a custom linear layer that produces two output units, corresponding to the two classes in our binary classification problem: positive or negative to pneumonia. This modification enables the model to output class probabilities for pneumonia detection based on the learned features from the pre-trained model. Furthermore, we froze the model parameters up to the "Stage 3" block (highlighted in red in Figure 2).

The fine-tuning process involved training the modified model on the target datasets using the cross-entropy loss function, with a learning rate of 0.0001 and L2 regularization (penalty of 0.0001) to prevent overfitting. To address the issue of class imbalance, class weights were incorporated into the loss function, giving more importance to the minority class. This adjustment helps the model better learn from the underrepresented class, improving its overall performance in imbalanced classification tasks. Data augmentation techniques were employed during training.

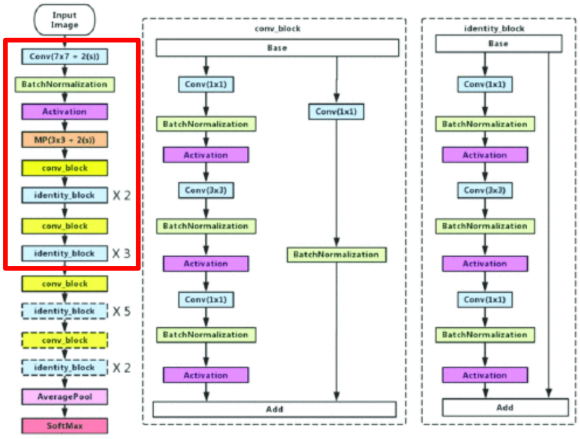


Figure 2: Architecture of ResNet-50

## 5 Datasets

In order to assess the generalization capabilities of the models, three distinct publicly available datasets were selected for this study. Each of these datasets offers unique characteristics in terms of size, label annotations, and image quality, making them suitable for evaluating the performance of machine learning models in the context of pneumonia detection. The following sections provide a detailed description of each dataset, highlighting their key features and the challenges they present for model training and evaluation.

### 5.1 Chest X-Ray Images (Pneumonia)

This dataset, available on Kaggle (14), contains a total of 5,863 chest X-ray images, divided into two categories: *normal* (healthy individuals) and *pneumonia* (patients affected by pneumonia). The dataset is imbalanced, with 74% of the images labeled as *pneumonia*, and the remaining 26% labeled as *normal*.

The dataset is originally split into three subsets: *train*, *test* (624 images), and *validation* (16 images). However, we found this splitting to be insufficient for our purposes, and thus, we chose to re-split the data using an 80%/20% ratio for training/validation and testing.

### 5.2 CheXpert

CheXpert (Irvin et al., 2019 (15)) is a large-scale dataset containing 224,316 chest X-ray images from 65,240 different patients. The dataset includes x-ray images from different perspective (frontal and lateral) and both AP and PA projections. The image label's are NLP-predicted, with the custom chexpert-labeler, based on the NegBio (16) labeler. Images includes labels for different medical conditions, including enlarged cardiomediasintum, cardiomegaly, lung opacity, lung lesion, and pneumonia. Additionally, each label can have four possible values:

- **Empty label** (no reference to the condition),
- **1** (confidently present),
- **0** (confidently absent),
- **-1** (uncertainly present, when reports are ambiguous).

For the purpose of pneumonia detection, we filtered the dataset to include only the images labeled for pneumonia. In cases where the label for pneumonia was not defined, we considered images with the label for *lung opacity* defined as negative for pneumonia if the value of *lung opacity* was 0. This is possible as pneumonia implies lung opacity. In the end, we ended up having 12,509 usable images, with 6,039 images labelled as positive.

**Note:** Images labeled as "negative" do not necessarily correspond to healthy individuals. Rather, they represent patients who are not affected by pneumonia. Therefore, while the "negative" images are used as non-pneumonia cases, they may still exhibit other health anomalies.

### 5.3 RSNA Pneumonia Detection Challenge

The RSNA Pneumonia Detection Challenge dataset (17) is part of a collaboration between the Radiological Society of North America (RSNA), the US National Institutes of Health (NIH), and other organizations, aimed at advancing the automated detection of pneumonia from chest X-ray images. This dataset contains 30,227 images, divided into two categories: *positive* (9,555 images) and *negative* (20,672 images).

Each image was hand-labeled by a single radiologist, with annotations for the presence of lung opacity, which is considered a primary indicator for pneumonia. The dataset aims to foster the development of machine learning models for the early detection of pneumonia, with the goal of automating initial screening processes to prioritize and expedite the review of potential pneumonia cases.

## 6 Experimental Results

The analysis focuses on comparing the performance of traditional machine learning models and

deep learning architectures for the task of pneumonia detection, both in intra-dataset and cross-dataset evaluations. Below, we highlight the key findings:

### 6.1 Performance Metrics

The models were evaluated using standard classification metrics, including Accuracy, Precision, Recall, F1-score, and ROC_AUC. The deep learning architectures outperformed traditional models in every scenario. The ResNet50 models consistently demonstrated the highest scores. We can clearly observe this behavior in Figures 3, 4, 5, 6, 7 and 8.

In the figures, the left bar represents, for each architecture and each metric, the average scores obtained in the intra-dataset scenario, where the models were tested on the test split of the dataset used during training. The right bar represents the average scores obtained in the cross-dataset scenario, where the models were tested exclusively on datasets different from the ones seen during training. Furthermore, scores were also compared with those obtained from a Dummy Classifier, used as a baseline. This classifier always predicts the majority class, making it a simple yet important reference for evaluating the effectiveness of the models. The comparison with this baseline is useful to highlight whether the models are indeed learning meaningful patterns in the data, as opposed to simply memorizing the dominant class distribution, which is important when dealing with class imbalance.

The results of our best architecture are comparable with the ones achieved by a related research (18).

### 6.2 Intra-Dataset vs Cross-Dataset Generalization

Traditional machine learning models, such as Logistic Regression, Decision Tree, and KNN, demonstrated overall lower performance compared to deep learning networks in both intra-dataset and cross-dataset evaluations. However, an interesting observation was that the gap between intra-dataset and cross-dataset performance was smaller on KNN and Logistic Regression, if compared with deep learning models. This suggests that while these models struggled to achieve high absolute performance, their simpler feature representations were less sensitive to domain shifts between datasets.

Regarding deep learning models, ResNet50 significantly outperformed the basic CNN in every scenario, showing that being pretrained on large datasets like ImageNet provides a strong advantage by enabling the model to extract more general and robust features.

### 6.3 Dataset Characteristics and Model Performance

The variability in dataset characteristics, such as the labeling methods and class imbalance, strongly influenced model performance. For example, CheXpert, with its NLP-predicted labels, posed unique challenges for both traditional and deep learning models due to potential labeling noise. RSNA posed challenges for the fact that classes are imbalanced. Chest-X-Ray Images posed challenges for both the size (only 5,8k images) and for the class imbalance, thus being the worst model for cross-dataset generalization. However, Chest-X-Ray Images was the best performing model in the intra-dataset domain. We can observe this behavior in Figures 9, 10, 11, 12, 13, and 14. In these figures, the x-axis represents the dataset used for training, while the y-axis represents the dataset on which the model was tested. The matrices display the performance scores for each model across different training and testing dataset combinations, allowing for a clear comparison of intra- and cross-dataset generalization capabilities.

### 6.4 Threshold optimization

In our classification task for pneumonia detection using X-ray images, we observed that using a thresholding approach on model outputs, rather than relying solely on softmax probabilities, significantly improved model performance. Specifically, for classifiers that output scores, we calculated the Receiver Operating Characteristic (ROC) curve and determined the optimal threshold by minimizing the Euclidean distance between the points on the curve and the ideal point (0, 1). This approach allowed us to fine-tune the decision threshold, balancing the trade-off between false positive rate (FPR) and true positive rate (TPR) for better classification accuracy.

However, this method is not applicable to classifiers like KNN and Decision Tree, which do not directly output scores.

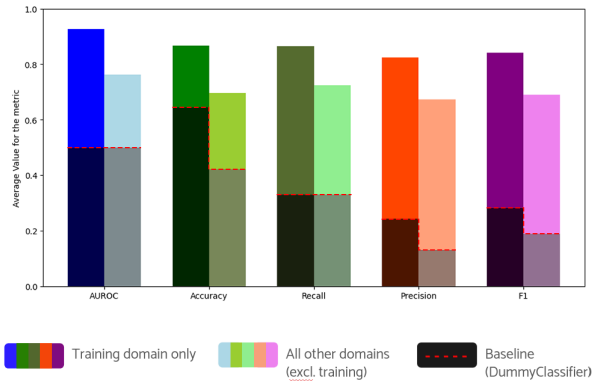


Figure 3: Score averages - ResNet50

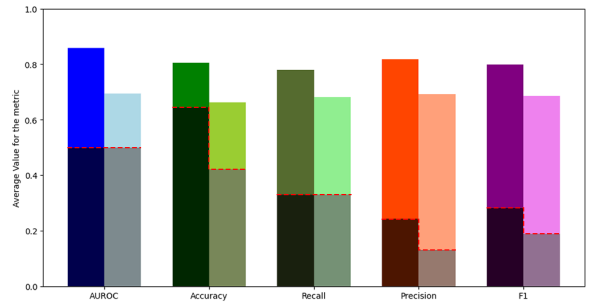


Figure 4: Score averages - CNN

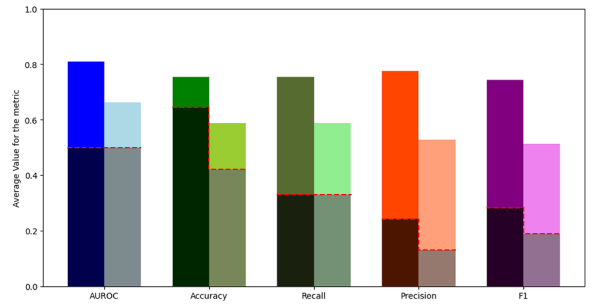


Figure 5: Score averages - SVM

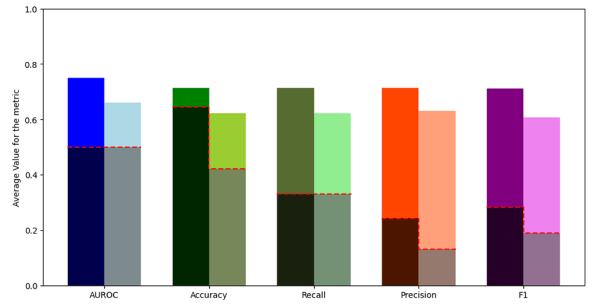


Figure 6: Score averages - KNN

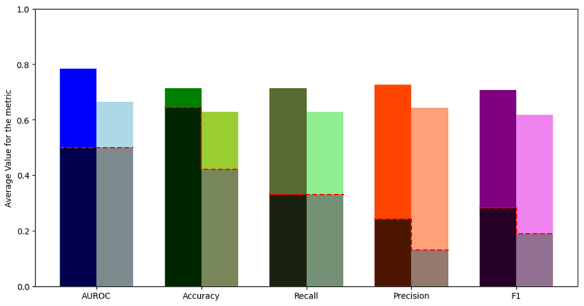


Figure 7: Score averages - Logistic Regression

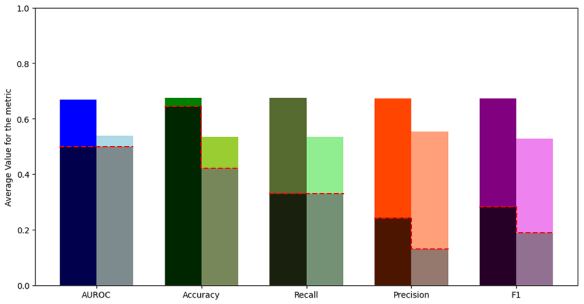


Figure 8: Score averages - Decision Tree

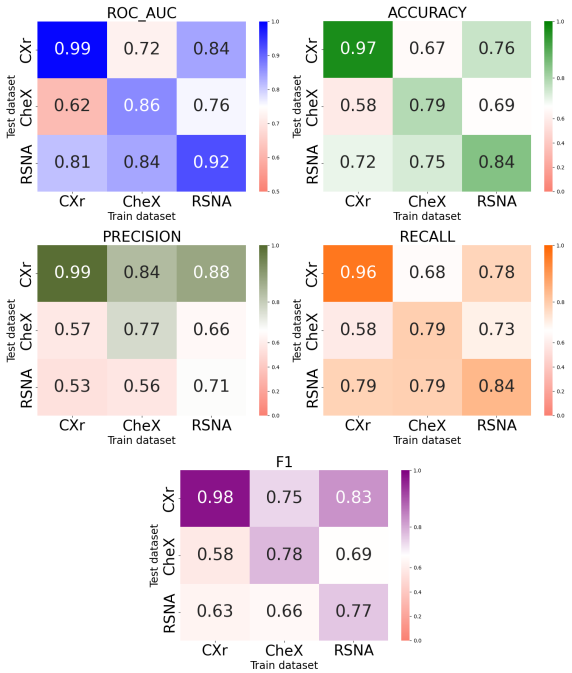


Figure 9: ResNet50

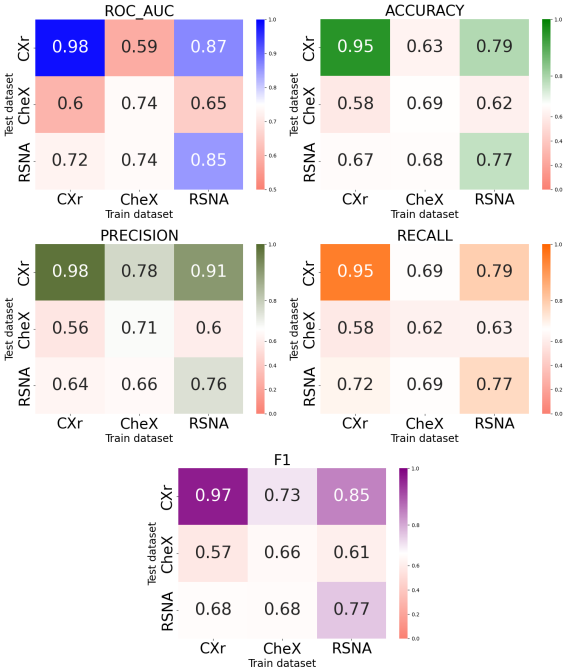


Figure 10: CNN

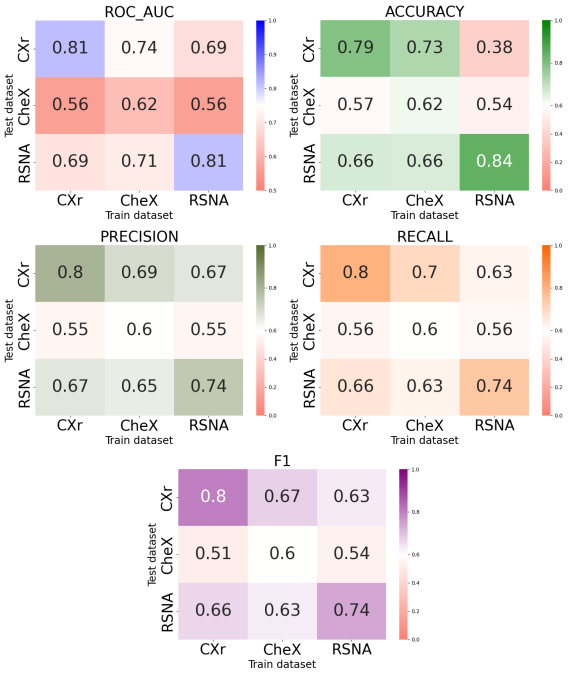


Figure 11: KNN

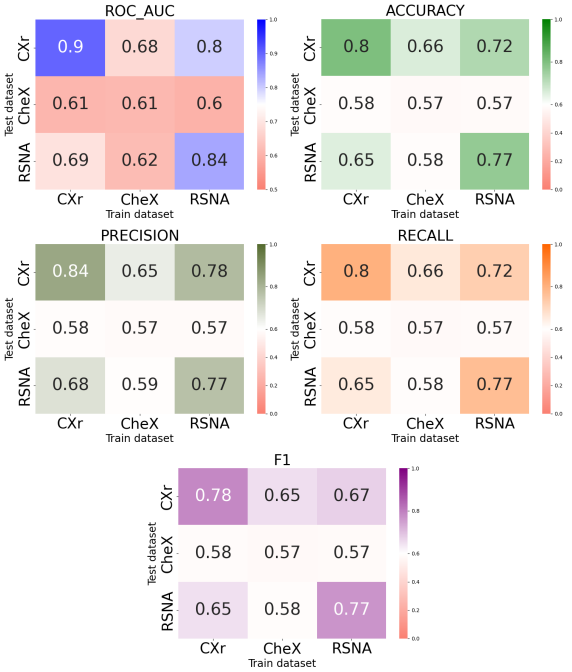


Figure 12: Logistic Regression

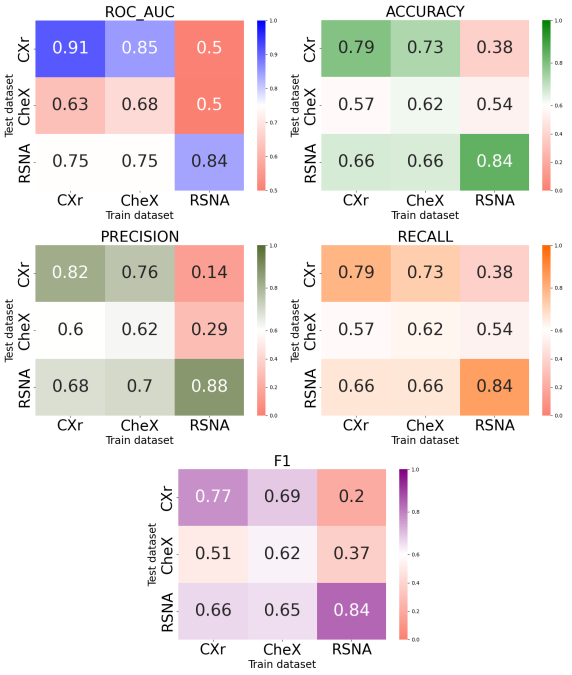


Figure 13: SVM

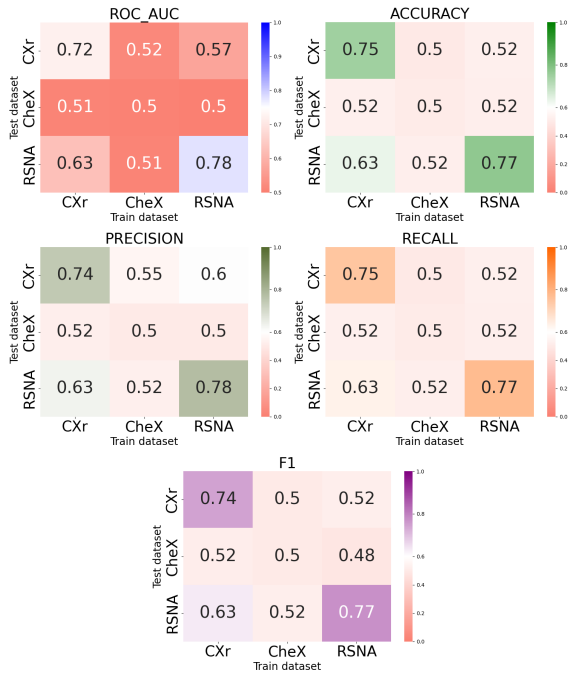


Figure 14: Decision Tree

## 7 Conclusions and Future Work

This study demonstrates the potential of AI-driven tools for pneumonia detection, highlighting the strengths and limitations of traditional and deep learning methods. The findings underscore the importance of considering dataset characteristics and cross-dataset evaluation to ensure model robustness in diverse clinical settings.

Future research directions include:

- **Integration of Multimodal Data:** Combining chest X-ray images with clinical data (e.g., patient history, symptoms, lab results) could provide a more comprehensive diagnostic approach, improving accuracy and reliability.
- **Addressing Dataset Bias:** To tackle bias in medical images datasets, future work should explore methods to handle class imbalance, labelling noise, and image quality variations. This could include Generative Adversarial Networks (GANs) or Transformer-based models leveraging attention mechanisms that can help to focus on the most important parts of the image.
- **Model Explainability:** While deep learning models have demonstrated superior performance, their ”black box” nature remains a concern for clinical applications. Future work should focus on developing techniques to improve the explainability of these models, allowing clinicians to understand why a particular diagnosis is made.

## 8 Leveraged Sources

To train the CNNs, we used some code snippets from here: https://www.kaggle.com/code/teyang/pneumonia-detection-resnets-pytorch

## References


[1] Ministero della Salute Italiana. *Linee guida per la prevenzione e il controllo delle malattie* (2021), https://www.salute.gov.it/portale/malattieInfettive/dettaglioSchedeMalattieInfettive.jsp?id=121&area=Malattie%20infettive&menu=indiceAZ&tab=1

[2] Gabruseva, T., Poplavskiy, D. & Kalinin, A. Deep Learning for Automatic Pneumonia Detection. *2020 IEEE/CVF Conference On Computer Vision And Pattern Recognition Workshops (CVPRW)*. pp. 1436-1443 (2020)

[3] Xu, X. A systematic review: Deep learning-based methods for pneumonia region detection. *Applied And Computational Engineering*. **22**, 210-217 (2023,10), http://dx.doi.org/10.54254/2755-2721/22/20231219

[4] Cover, T. M., & Hart, P. E. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* pp. 21-27. (1967)

[5] Cortes, C., & Vapnik, V. Support-vector networks. *Machine Learning*, **20** pp. 273-297 (1995)

[6] Quinlan, J. R. Induction of decision trees. *Machine Learning*, **1** pp. 81-106 (1986)

[7] Cox, D. R. The regression analysis of binary sequences (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)* pp. 215-242 (1958)

[8] J. Ian, Principal Component Analysis. *International Encyclopedia of Statistical Science* pp. 1094-1096

[9] Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* pp. 111-147 (1974)

[10] Barneih, F., Nasir, N., Kansal, A., Alshaltone, O., Bonny, T., Al-Shabi, M. & Al–Shammaa, A. Pneumonia Detection in Chest X-ray Images using ResNet50 Model. *2023 Advances In Science And Engineering Technology International Conferences (ASET)*. pp. 01-04 (2023)

[11] Rahman, T., Chowdhury, M., Khandakar, A., Islam, K., Islam, K., Mahbub, Z., Kadir, M. & Kashem, S. Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray. *Applied Sciences*. **10** (2020), https://www.mdpi.com/2076-3417/10/9/3233

[12] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. & Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal Of Computer Vision (IJCV)*. **115**, 211-252 (2015)

[13] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 770-778 (2016)

[14] Kermany, D., Zhang, K. & Goldbaum, M. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. (2018), https://api.semanticscholar.org/CorpusID:126183849

[15] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D., Halabi, S., Sandberg, J., Jones, R., Larson, D., Langlotz, C., Patel, B., Lungren, M. & Ng, A. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. (2019), https://arxiv.org/abs/1901.07031

[16] Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R. & Lu, Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. (2017), https://arxiv.org/abs/1712.05898

[17] Anouk Stein, M., Wu, C., Carr, C., Shih, G., Dulkowski, J., Kalpathy, Chen, L., Prevedello, L., Marc Kohli, M., McDonald, M., Peter, Culliton, P., MD, S. & Xia, T. RSNA Pneumonia Detection Challenge. https://kaggle.com/competitions/rsna-pneumonia-detection-challenge (2018), Kaggle

[18] Cohen, J., Hashir, M., Brooks, R. & Bertrand, H. On the limits of cross-domain generalization in automated X-ray prediction. (2020), https://arxiv.org/abs/2002.02497