Assignment 2 - XAI

Diego Sourbag, Reinier Bos
s3699420, s3703177

1. Task definition

This assignment focuses on the evaluation of machine learning models to classify chest radiographs, which are focused on classifying them as normal or showing signs of pneumonia. The goal it to find the most effective model in terms of predictive performance. It should also be able to interpret the models well, using Explainable AI (XAI) techniques.

We will begin by exploring and visualizing the data set to understand the distribution of the class and the characteristics of the image. A baseline Convolutional Neural Network (CNN) model will be implemented, followed by performance evaluation through accuracy, precision, recall, F1 score, and confusion matrix. Subsequently, hyperparameter optimization will be performed, and an alternative model may be tested for comparison.

To gain insights into model behavior, XAI methods such as Grad-CAM will be applied to visualize which image regions most influence the model's predictions, especially in misclassified cases. The deliverables include a structured report with a comparison of model performance and an in-depth analysis using XAI, culminating in a conclusion on the most effective modeling approach for this task.

2. Data exploration

The dataset used for this assignment is the Chest X-Ray Images (Pneumonia) dataset available on Kaggle (Mooney, 2018). It consists of grayscale X-ray images categorized into two classes: Normal and Pneumonia. The dataset is structured as:

- Train set: Contains the majority of the images used to train the model.
- Validation set: Used during model development for performance monitoring and tuning.
- Test set: Used to evaluate the final model performance.

An initial analysis revealed that the dataset is imbalanced, with significantly more samples labeled as "Pneumonia" than "Normal." This imbalance may influence the model's ability to generalize well and will be taken into account during evaluation.
The images in the dataset are all grayscale, and they are generally consistent in resolution. The images are rescaled to a 150×150 dimension for compatibility with the baseline CNN and Bayesian optimization. For the ResNet model, the images are rescaled to a 224×224 dimension, to make sure the match the input requirements of the model.

Set	Normal	Pneumonia	Total
Train	1341	3875	5216
Validation	8	8	16
Test	234	390	624

Table 1: Class distribution of the chest X-ray dataset.

To get an understanding of the data, we analyzed a few pictures from both classes. The Pneumonia images show noticeable white spots or cloudy areas, which is an indication of Pneumonia, because fluid builds up in the lungs. The normal X-ray images show a clear image of lungs.

2.1 Data augmentation

To avoid overfitting, the dataset is expanded to be even bigger. This is done by making small transformations on the existing dataset, to simulate variations. Some techniques that were applied are: grayscale, horizontal and vertical flips, random crops and much more. By doing this, the dataset is expanded to create a better model.

3. Baseline Method

To establish a performance benchmark for the task, we implemented a Convolutional Neural Network (CNN) for binary classification: *Normal* vs. *Pneumonia*. The architecture is based on the example from the Kaggle notebook.

Architecture Overview

The model consists of three convolutional blocks with increasing filters (32, 64, 128 and 256), each followed by max pooling. A flatten layer prepares the data for a dense layer with 128 units, followed by dropout and a sigmoid output layer for binary classification.

Preprocessing and Training Setup

All images were resized to 150×150 pixels and normalized to the [0, 1] range. The model was trained using binary crossentropy loss and the RMSprop optimizer. A batch size of 32, and 12 epochs were used.

4. Hyperparameter Optimization

To enhance the performance of the baseline Convolutional Neural Network (CNN), we applied Bayesian Optimization. To do this, we used **BayesSearchCV** from the **skopt** library. This method systematically explores the hyperparameter space and is more sample-efficient than traditional grid or random search techniques. This is especially valuable given the limited computational resources.

Search Space and Parameters

The search was conducted over three key hyperparameters that influence model capacity and regularization:

- **Learning Rate:** Log-uniformly sampled between 10^{-5} and 10^{-2} to balance convergence speed and stability.
- **Dropout Rate (Dense Layers):** Uniformly sampled between 0.05 and 0.4 to mitigate overfitting by randomly dropping neurons in fully connected layers.
- **Dropout Rate (Convolutional Layers):** Uniformly sampled between 0.1 and 0.4 to regularize the convolutional filters during training.

The model was wrapped using **KerasClassifier** to enable compatibility with **BayesSearchCV**, and trained for 12 epochs with a batch size of 32. The search procedure evaluated 10 combinations over 3-fold cross-validation, using accuracy as the scoring metric.

Best Configuration and Performance

After conducting the hyperparameter optimization, the best model configuration was identified as follows:

- **Learning Rate:** 0.0007435894120716282
- **Dropout Rate (Dense):** 0.28898246126575944
- **Dropout Rate (Conv):** 0.2924782969628286

5. Explainable AI analysis

Model explainability is important, especially in medical diagnosis. It must be possible to understand why a model made a decision. To achieve this, we implemented an Explainable AI technique to make the model more transparent. We used Grad-CAM (Gradient-weighted Class Activation Mapping) to interpret the model's predictions (Keras Team, 2021).

Grad-CAM produces heatmaps as visual explanations. It does this by using the gradients of the target class with respect to the last convolutional layer. The heatmaps show which regions of the input image influenced the model's decision the most. We followed these steps to generate Grad-CAM heatmaps:

1. The input X-ray image is passed through the trained CNN model. It predicts a class (Pneumonia or Normal).
2. Grad-CAM identifies the most influential regions in the final convolutional layer.
3. These regions are combined into a heatmap using a weighted sum of the feature maps.
4. The heatmap is resized and overlaid on the original image.
5. Negative values are removed, so only positive, supporting evidence appears in the visualization.

To study model behavior, we applied Grad-CAM to five false positive cases (predicted class 1, true class 0) and five false negative cases (predicted class 0, true class 1) for each model. The heatmaps are shown in Appendix A. Each one includes the model's prediction, confidence score, and the true label.

6. Results

Baseline CNN Model

The baseline Convolutional Neural Network (CNN) was evaluated on the test set using standard classification metrics. The model achieved a high overall performance, particularly in distinguishing between normal and pneumonia cases.

Loss and Accuracy

The model's final performance on the test set is detailed below:

- **Loss:** 0.3370
- **Accuracy:** 91.99%

Confusion Matrix

Table 2: Confusion Matrix – Baseline CNN Model

	Predicted Pneumonia (1)	**Predicted Normal (0)**
Actual Pneumonia (1)	372 (TP)	18 (FN)
Actual Normal (0)	32 (FP)	202 (TN)

Evaluation Metrics

The performance of the baseline model was assessed using precision, recall, and F1-score per class. The results are summarized below:

Table 3: Classification Report – Baseline CNN Model

Class	**Precision**	**Recall**	**F1-score**	**Support**
Pneumonia (1)	0.92	0.95	0.94	390
Normal (0)	0.92	0.86	0.89	234
Accuracy	-	-	0.9199	624
Macro Avg	0.92	0.91	0.91	624
Weighted Avg	0.92	0.92	0.92	624

These results indicate that the baseline CNN performs well overall, especially in detecting pneumonia, with a slightly lower recall for the Normal class, which may reflect a tendency to over-predict pneumonia cases.

6.1 Bayesian Hyperparameter Optimization

To improve upon the baseline CNN, we applied Bayesian hyperparameter optimization. The optimized model was evaluated using the same metrics and test set as the baseline.

Loss and Accuracy

- **Loss:** 0.2818
- **Accuracy:** 90.54%

Confusion Matrix

Table 4: Confusion Matrix – Optimized CNN Model

	Predicted Pneumonia (1)	**Predicted Normal (0)**
Actual Pneumonia (1)	362 (TP)	28 (FN)
Actual Normal (0)	31 (FP)	203 (TN)

Evaluation Metrics

The optimized model showed slightly improved balance in precision and recall across both classes, though the overall accuracy decreased slightly. The detailed metrics are shown in Table 5.

Table 5: Classification Report – Optimized CNN Model

Class	**Precision**	**Recall**	**F1-score**	**Support**
Pneumonia (1)	0.92	0.93	0.92	390
Normal (0)	0.88	0.87	0.87	234
Accuracy	-	-	0.9054	624
Macro Avg	0.90	0.90	0.90	624
Weighted Avg	0.91	0.91	0.91	624

Despite a slight drop in overall accuracy compared to the baseline, the optimized model demonstrated better recall for pneumonia and more balanced performance across both classes. This suggests improved generalization, especially in reducing false negatives for pneumonia.

6.2 ResNet50 Model

To explore an alternative architecture, we trained a ResNet-based model on the same dataset. The results indicate a different trade-off between precision and recall compared to the CNN models.

Loss and Accuracy

The model's performance on the test set is as follows:

- **Loss:** 0.4621
- **Accuracy:** 75.96%

Confusion Matrix

Table 6: Confusion Matrix – ResNet Model

	Predicted Pneumonia (0)	**Predicted Normal (1)**
Actual Pneumonia (0)	355 (TP)	35 (FN)
Actual Normal (1)	115 (FP)	119 (TN)

Evaluation Metrics

The ResNet model's classification performance is presented below:

Table 7: Classification Report – ResNet Model

Class	**Precision**	**Recall**	**F1-score**	**Support**
Pneumonia (0)	0.76	0.91	0.83	390
Normal (1)	0.77	0.51	0.61	234
Accuracy	-	-	0.7596	624
Macro Avg	0.76	0.71	0.72	624
Weighted Avg	0.76	0.76	0.75	624

While the overall accuracy of the ResNet model is lower than the baseline and optimized CNNs, it demonstrates high recall for pneumonia cases. This suggests that the model is more sensitive to detecting pneumonia but at the cost of more false positives for normal cases. The imbalance between recall and precision for the Normal class indicates room for improvement in handling class-specific features.

6.3 XAI analysis

6.3.1 Baseline Model

Misclassifications Analysis
The visualization is visible in Figure 1. In the false positive cases, the Grad-CAM heatmaps consistently highlight regions such as the heart and spine. These is anatomically dense but not typically associated with disease. This suggests the model may be focusing on visually dominant structures rather than pathology. In the false-negative cases, the model shows weak or misplaced activations. These were sometimes outside the lungs, indicating that is misses the subtle signs that it should focus on.

Interpretation of Results
The Grad-CAM results show that the baseline model generally focuses on consistent regions across samples. This may support its high classification performance. In several cases, the highlighted areas align with parts of the lung, indicating that the model can learn relevant

patterns. However, especially with false negatives, the model fails to activate on key areas. This is suggesting room for improvement in sensitivity and localization.

Overall, while the model performs well and exhibits reasonable behavior in many cases, explainability analysis highlights occasional reliance on non-causal features and difficulty with subtle findings. These insights suggest potential benefits from incorporating lung-field masking or fine-grained annotations to further enhance both performance and intepretability.

6.3.2 HPO Model

Misclassifications Analysis
The Grad-CAM visualizations for the HPO model's misclassified samples are shown in Figure 2. Similar to the baseline model, the false positive cases highlight regions such as the mediastinum, diaphragm, and central thoracic area. However, the activations in the HPO model appear more localized and often extend into the lower lung fields, which are more clinically relevant for detecting pneumonia.

In the false negative cases, the HPO model shows improved attention to pulmonary regions compared to the baseline model. Although some activations are still weak or diffuse, several examples show meaningful focus on peripheral and upper lobe areas where abnormalities are visible. This suggests the HPO model has developed a more refined internal representation and is better at identifying subtle pathological features.

Interpretation of Results
The Grad-CAM heatmaps of the HPO model reveal more focused and medically relevant activations than those observed in the baseline model. In many false positives, the model appears to be influenced by ambiguous regions that might share texture characteristics with pneumonia but are not pathological. In false negatives, while the model occasionally misses smaller or atypical findings, it often highlights parts of the lungs where subtle signs are present.

Overall, the HPO model demonstrates improved spatial awareness and stronger alignment with clinical features. The use of optimized hyperparameters has impact on determining features .

6.3.3 ResNet50 Model

Misclassifications Analysis
Figure 3 show the heatmaps of the ResNet50's model performance. In false positive cases, the model tends to activate around strong structural features such as the spinal column and diaphragm, with some attention leaking into the lower lung zones. While these areas are often rich in anatomical detail, they are not necessarily indicative of pneumonia. These heatmaps suggest that the model may occasionally mistake normal anatomical patterns for pathological signals.

In false negative cases, the model shows moderate but scattered activation within lung fields. Compared to the baseline model, ResNet displays a broader spatial focus, sometimes identifying subtle features on the periphery of the lungs. However, it occasionally under-activates or highlights regions lacking clinical relevance, leading to missed pneumo-

nia predictions.

Interpretation of Results
The Grad-CAM visualizations for ResNet reveal a model that is more spatially aware than the baseline, but not as precise as the HPO model. In several false positive samples, activations concentrate on regions that appear visually complex but are not associated with pneumonia, such as the heart border or bony structures. This may indicate the model's sensitivity to texture but insufficient refinement in distinguishing between anatomical and pathological cues.

In the false negatives, the model's focus does include relevant lung regions, though it sometimes misses finer details. This behavior may reflect an overconfidence determining a non-pneumonia case.

Overall, the ResNet model demonstrates improved interpretability and more clinically aligned attention compared to the baseline. However, it still struggles in differentiating between subtle pathology and non-pathological complexity.

7. Conclusion

Three models for pneumonia detection from chest X-ray images were compared. We compared a baseline CNN, a Bayesian-optimized CNN (HPO model), and a ResNet50 model. The baseline CNN achieved the highest accuracy (91.99%) and showed strong precision and recall for both classes. However, as this is focused on the medical world, accuracy is not the only thing that is important. False negatives, missing pneumonia cases, can have bad outcomes.

The Bayesian-optimized CNN has a slightly lower overall accuracy (90.54%). However, it has a more balanced recall across both classes. Especially improving recall for pneumonia. The Grad-Cam visualizations also revealed more clinically relevant activations. This shows that the model learned more meaningful patterns, so that it could better identify pathological features. This is important, as sensitivity to pneumonia is very important in the medical world.

The ResNet50 model had the lowest overall accuracy (75.96%) and weak precision for normal cases. This resulted in a high number of false-positives. Grad-CAM showed that the model often focused on visually complex anatomical structures rather than pathological signs. There could be some reasons for this underperfomance, like architectural differences, lack of hyperparameter tuning, or overfitting.

In conclusion, we see that the baseline CNN has the highest accuracy. However, the best overall performance is by the Bayesian-optimized model. It has a good balance between predictive performance and intepretability. It has improved recall for pneumonia, and more focused attention maps make it the most suitable model for this task.

8. Apendix

8.1 Misclassified images - Baseline model

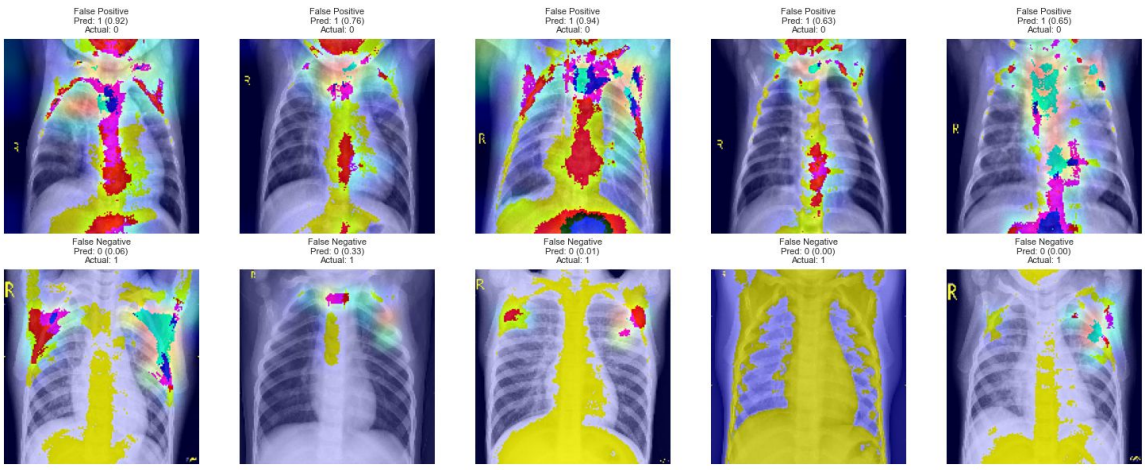

Figure 1: Misclassified images for the baseline model

8.2 Misclassified images - HPO model

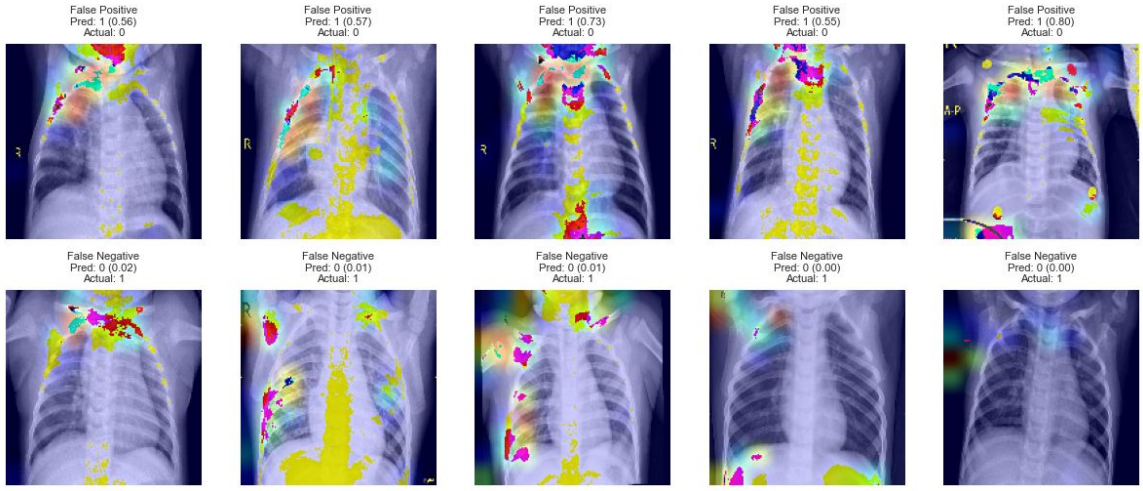

Figure 2: Misclassified images for the hpo model

8.3 Misclassified images - ResNet50 model

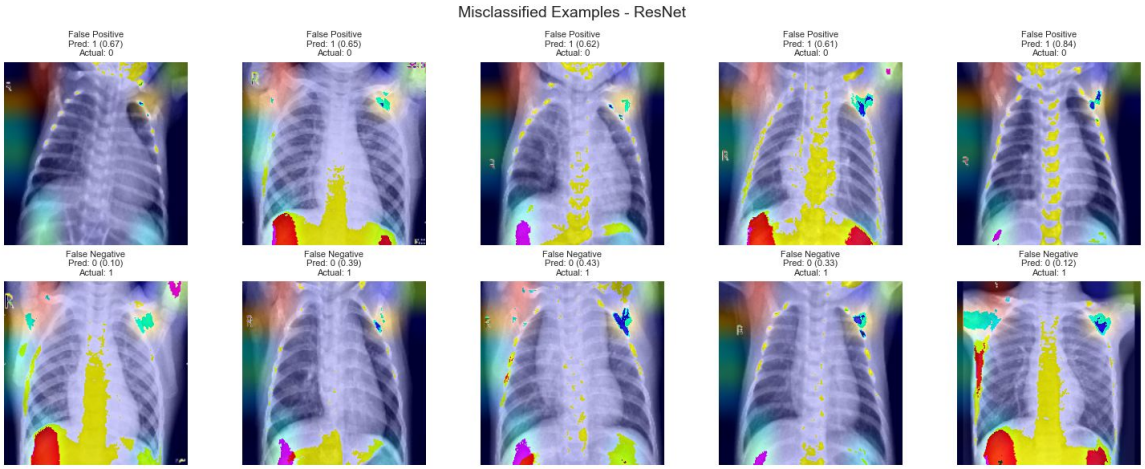

Figure 3: Misclassified images for the ResNet50 model

References

Keras Team. Grad-cam class activation visualization, 2021. URL `https://keras.io/examples/vision/grad_cam/`. Accessed: 2025-04-16.

Paul Mooney. Chest x-ray images (pneumonia). `https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia`, 2018. Accessed: 2025-04-16.