**XAI Assignment Report**

Group 12

Luca Duţu, Isabella La Mola, Rares Popa

Course: Business Analytics

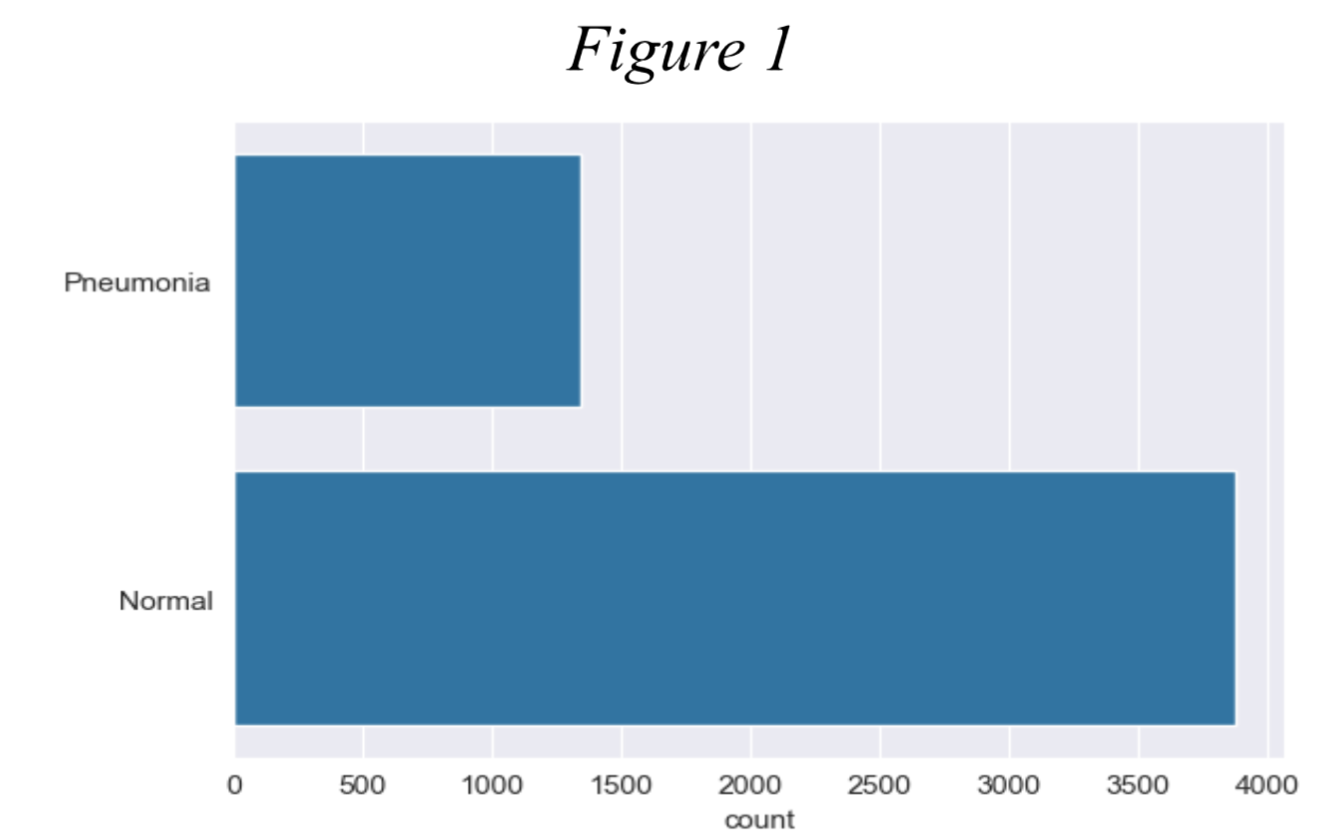
Instructor: Dr. Niki van Stein

Date: 16.12.2024

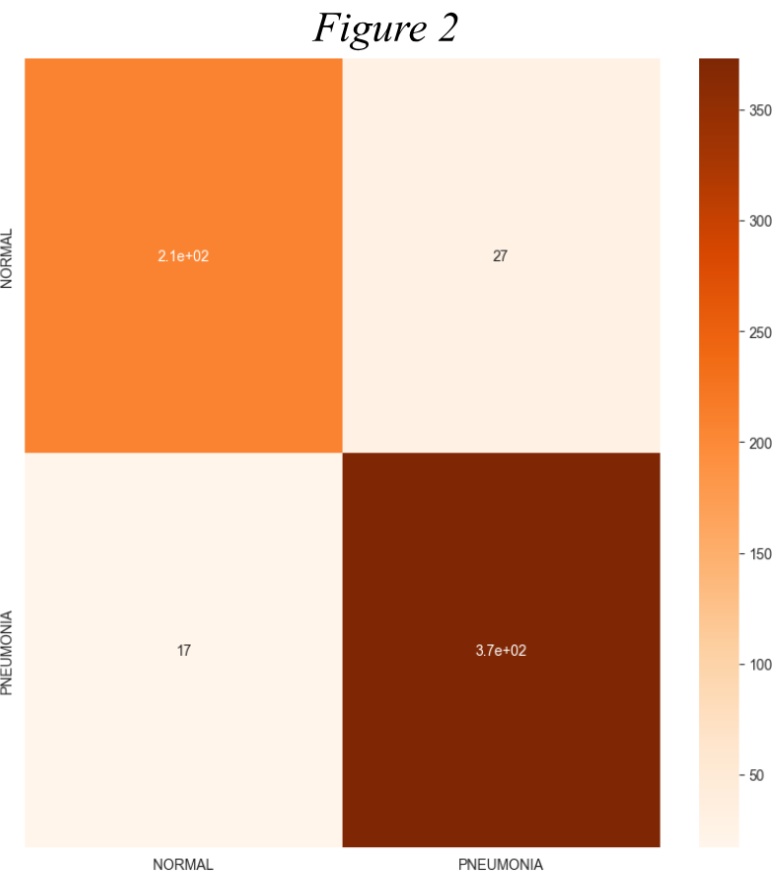
**Task Definition**

The objective of the assignment is to predict whether patients have pneumonia or not based on chest X-Ray images. A conventional neural network (CNN) was the main basis of this project, upon which explainable AI (XAI) methods were applied for analysis. This paper outlines the steps in this project, starting with data exploration to understand the structure of the dataset. The CNN model– sourced from Kaggle– was then implemented, trained, and evaluated to establish its baseline performance metrics. After establishing a baseline, the model altered in attempts to optimize hyperparameters in a way that would benefit its performance and runtime. Finally the last step in this task involved evaluating the CNN using an XAI technique, namely Grad was evaluated using Explainable AI (XAI) methods like Grad-Cam. This method was used to visualize the decision-making process of the CNN, highlighting the areas the model focused on when making predictions. The ultimate goal of this task was to optimize a model so it accurately diagnoses pneumonia, and to apply XAI techniques to explore the decision-making of the model.

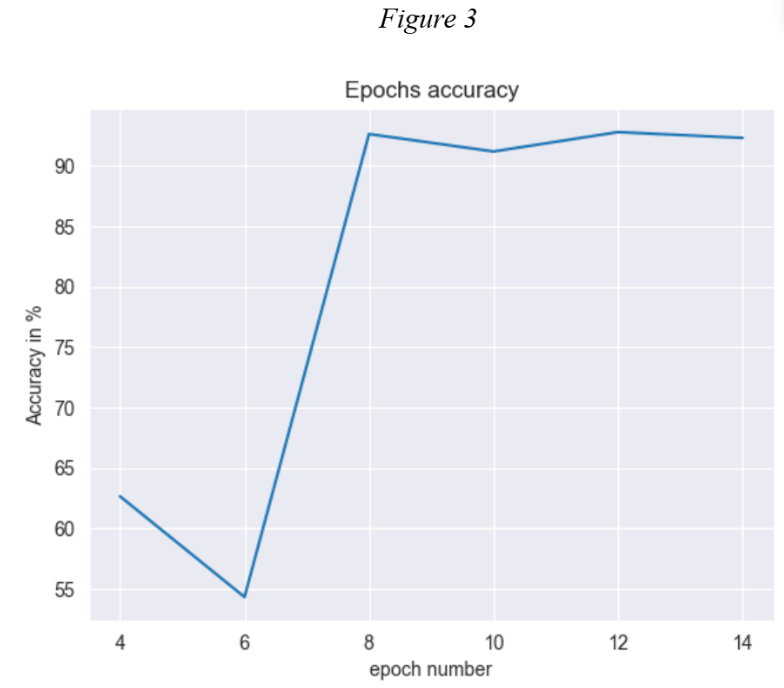
**Data Exploration**

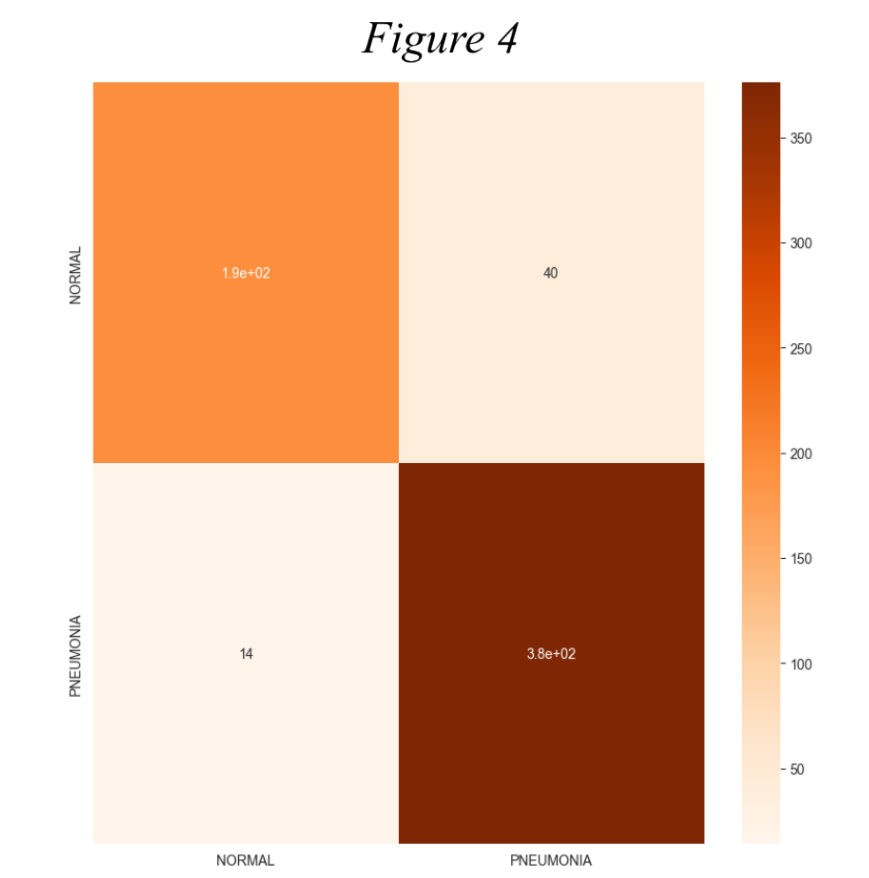
The dataset used was downloaded for Kaggle, and it contained chest X-Ray images categorized into two classes: “*Normal*” and “*Pneumonia*”. Each image was resized to 150x150 pixels for standardization of the inputs for the CNN model. There are three subsets across which the data is divided: training, validation, and testing, each of which is contained in a separate directory. An analysis of the training dataset revealed an imbalance in the sample, with healthy cases significantly outnumbering pneumonia cases. Specifically, the dataset contains 3,875 healthy cases and 1,341 pneumonia cases, as shown in *Figure 1*. This is an important detail to consider, as this imbalance could influence the model's performance, leading to biased predictions favoring the majority class. Nevertheless, the dataset is well-suited for training due to the large quantity of high-resolution, pre-classified medical images.

**Baseline Method**

The baseline model is a Convolutional Neural Network (CNN) designed to classify chest X-rays as either “*Normal*” or “*Pneumonia*”. The model consists of several convolutional layers to find patterns in the images. It also utilizes pooling and dropout layers to make the data smaller and prevent the model from overfitting. The classification is done with a sigmoid function applied to the output layer, since we are dealing with a binary classification task. The output of the final layer which is between 0 and 1 is then rounded to the closest of the two. The baseline model was trained using the Adam optimizer, which is a very efficient algorithm using the stochastic gradient descent method. Training took place over 12 epochs with a batch size of 32 and data augmentation via ImageDataGenerator was employed to enhance model robustness. Lastly, binary cross-entropy was chosen as the loss function because of its proficiency in binary classification tasks. 

As displayed in *Table 1* in the results section below, the original model achieved an overall accuracy of 93%. This demonstrates a strong performance in distinguishing between “*Normal*” and “*Pneumonia*” classes. The precision, which measures the proportion of correct positive predictions, was 0.91 for “*Pneumonia*” and 0.93 for “*Normal*” cases. Recall values were 0.88 for “*Pneumonia*” and 0.96 for “*Normal*”, indicating the model’s effectiveness in identifying true positives. The F1-scores, which balance both precision and recall, were 0.90 and 0.94 for their respective categories. This indicates an overall balance between classes, despite the difference in data points within each class. The confusion matrix in *Figure 2* above, reveals that the model correctly classified around 370 “*Pneumonia*” cases and 200 “*Normal*” cases. However, it misclassified a larger proportion of “*Pneumonia*” cases as “*Normal*” than vice versa. While the model demonstrates strong overall performance, the results suggest an opportunity to reduce false negatives, particularly for “*Pneumonia*” cases.

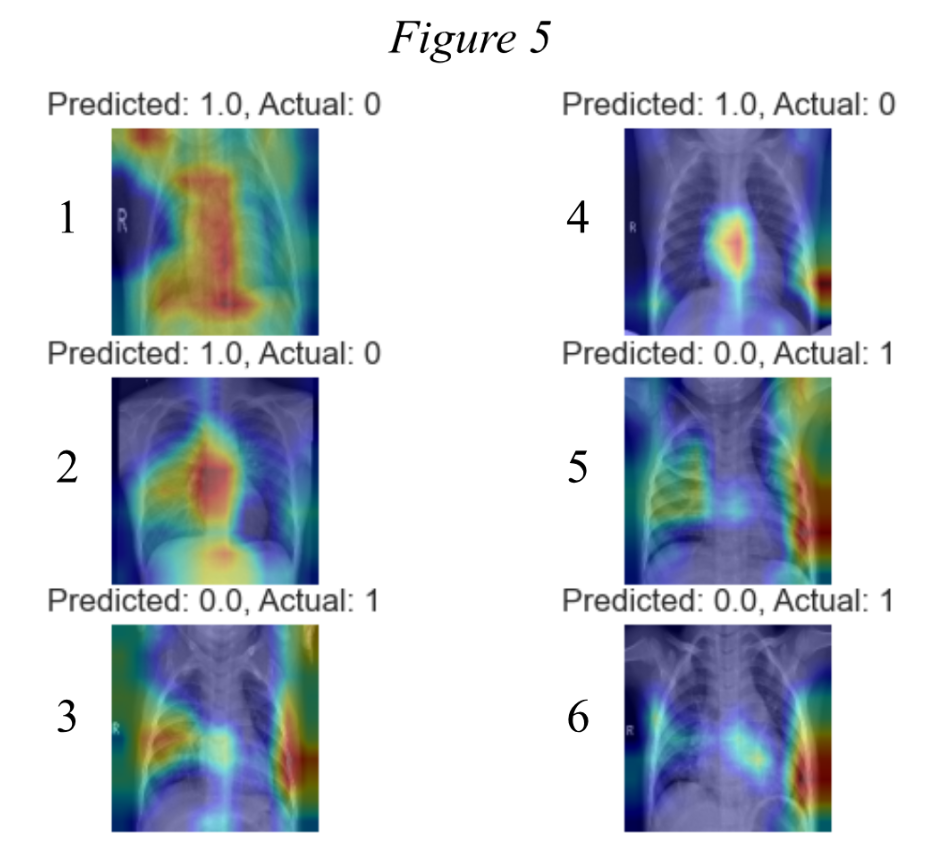
**Hyperparameter Optimization** 

The selected method for the hyperparameter optimization was Grid Search, which consists in an exhaustive search over a series of values. This was meant to optimize the number of epochs and batch size for the CNN model. After testing, epochs came out as the primary hyperparameters since it had a higher impact than both batch size and learn rate. The number of epochs was optimized to 8. As shown in the accuracy graph in *Figure 3*, there is a noticeable spike in accuracy at 8 epochs. This choice balances model performance and computational efficiency by avoiding overtraining while still achieving high accuracy.

The final model was trained using 8 epochs and a batch size of 32 and evaluated on both the training and testing sets. The precision and recall of the model can be seen in *Table 2*, with the final model having an accuracy of 91%.

The results of the confusion matrix in *Figure 4* highlights the trade-off between recall and precision for “*Pneumonia*”. Even though precision is high, the model struggles more with false positives. This situation inherently calls for further techniques of case balancing to improve the recall power. The decision to optimize the epochs was driven by the need to maximise accuracy. This was done in a way to avoid overfitting the model. The results indicate that the final model achieved balanced performance in prediction. Ultimately, the accuracy is less by about 1.6% (rounded to 2% in *Table 2*) compared to the baseline model, but the optimized model classifies less cases of “*Pneumonia*” as “*Normal*”. This case is preferable as it is better to identify and treat potential pneumonia cases than risk missing a diagnosis, which might have fatal consequences.

**Explainable AI Analysis**

Explainable AI (XAI) refers to processes applied to machine learning outputs to try to uncover why models produce a given outcome. For the purposes of this project, Grad-CAM– an XAI technique used to understand convolutional neural network models– was applied to inspect misclassification errors. Applying Grad-CAM to an image produces a heatmap where the closer the color is to red, the greater the focus the model has on that area in the prediction. The images in *Figure 5* are the misclassification errors that were investigated with Grad-CAM. For image 1 the model predicted "*Normal*, but the actual class was “*Pneumonia*”. The heatmap focuses heavily on the spinal region, also showing a large activation in the top left corner near the shoulder. Overall the model seems to focus on a relatively large and irrelevant area in the image, features which are not crucial for pneumonia detection. While this activation does include the lungs, the overall quality of the prediction was likely reduced by the noise. The image in the middle left was predicted “*Normal*” when the class was “*Pneumonia*”. There is less activation in the overall image, but an unjustified focus on the mid-spine/ heart region is still problematic. This same misclassification trend is observed in image 4, which could mean that the model may have found a spurious correlation between the mid-spinal/ heart region and images labeled as “*Pneumonia*”. Misclassifications of “*Normal*” predicted as “*Pneumonia*” shared the issue of the model focusing on irrelevant background features rather than the lungs. As presented in image 3, 5 and 6, the heatmap lights up red predominantly in areas outside of the body’s outline. This activation is not at all reasonable, because there is nothing in the periphery of the body scan that should serve as a health indicator. In this instance the background noise within the model is the main focus when classifying the image as “*Normal*”, which could indicate a systematic bias within the model. These systematic features among misclassified samples suggest a need for further refinement to address these biases and improve the model’s focus on clinically relevant areas for more accurate predictions.

**Results**

*Table 1: Baseline Model Results*

| Class | Precision | Recall | F1-Score |
| --- | --- | --- | --- |
| Pneumonia (0) | 0.92 | 0.88 | 0.90 |
| Normal (1) | 0.93 | 0.96 | 0.94 |
| Accuracy |  |  | 0.93 |

*Table 2: Hyperparameter Optimization*

| Class | Precision | Recall | F1-Score |
| --- | --- | --- | --- |
| Pneumonia (0) | 0.93 | 0.83 | 0.88 |
| Normal (1) | 0.90 | 0.96 | 0.93 |
| Accuracy |  |  | 0.91 |

### 

*Table 3: Results of Grad-CAM Analysis*

| Image | Predicted  Class | Actual Class | Heatmap Focus | Misclassification Insight |
| --- | --- | --- | --- | --- |
| 1 | Normal | Pneumonia | Peripheral and spine | Fixation on spine and large irrelevant area |
| 2 | Normal | Pneumonia | Central chest area | Incorrectly focuses on non-lung regions. |
| 3 | Pneumonia | Normal | Peripheral of the lungs | Misses critical lung opacity features. |
| 4 | Normal | Pneumonia | Heart area/ Spine | Non-relevant anatomical regions drive prediction. |
| 5 | Pneumonia | Normal | Edges and ribs | Fixation on ribs and  irrelevant background area |
| 6 | Pneumonia | Normal | Bottom-Right | Barely focuses on the lungs |

**Conclusion**

This project aimed to classify chest X-rays into “*Normal*” or “*Pneumonia*” using a Convolutional Neural Network and evaluate its decision-making process with explainable AI techniques. The baseline model achieved strong performance metrics, and a relatively balanced model with an accuracy of 93%. Despite its success, the model exhibited a tendency to misclassify “*Pneumonia*” as “*Normal*”, as observed in the confusion matrix. These false negatives are potentially harmful, therefore conducting hyperparameter optimization was done to further refine the model. The final model optimized runtime by improving efficiency by reducing the number of epochs to 8, while improving the precision in the “*Pneumonia*” class. The tradeoff in accuracy was justified by the notion that it is better safe than sorry– hence flagging the illness when it is not actually there rather than risk it going undetected.

To delve deeper into the model’s decision-making, the XAI method Grad-CAM was applied to analyze misclassified cases, uncovering potential issues in the model's attention mechanisms. The heat maps revealed a tendency to focus on irrelevant areas such as the spine, heart, and peripheral regions of the images when incorrectly classifying “*Pneumonia*” as “*Normal*” , often overlooking critical lung features. Similarly, “*Normal*” cases misclassified as “*Pneumonia*” often highlighted the peripheral regions. These findings emphasize possible systematic biases from within the model, and indicate the need to refine the model's ability to prioritize clinically relevant areas– in this case the lungs.

To improve performance, class imbalance should be minimized. Techniques such as weighted loss functions or oversampling “*Pneumonia*” cases can help the model handle underrepresented classes more effectively. Additionally, expanding the training dataset with more diverse and balanced samples may aid in reducing biases and enhance generalization. Further applications of XAI techniques such as Layer-wise Relevance Propagation or exploring architectures like EfficientNetB0 might help optimize the model further. Combining these strategies has the potential to significantly enhance the model's diagnostic accuracy and reliability, increasing its value as a tool for medical image analysis.