# Detection of Pneumonia from Chest X-ray Images

Georgios Karagounis-MTN2310

Angeliki Kostopoulou-MTN2316

# Table of contents

MSc in ARTIFICIAL INTELLIGENCE

# 1 Abstract

Pneumonia has been directly responsible for a huge number of deaths globally.Visually, it becomes difficult to identify pneumonia, since it can be confused with other respiratory diseases, such as tuberculosis. Moreover, there is significant variability in the way chest X-ray images are acquired and processed, which can impact the quality and consistency of the images. This can make it challenging to develop robust algorithms that can accurately identify pneumonia in all types of images.

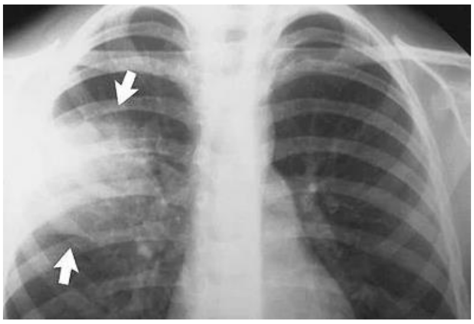
# 2 Introduction

Pneumonia is a respiratory disease that causes inflammation in each or both lungs, resulting in symptoms such as cough, fever, and difficulty breathing. Early detection of pneumonia is essential for effective treatment and improved patient's recovery. Unfortunately, pneumonia is just one of several lung diseases, thus radiographic results do not always confirm a pneumonia diagnosis. Therefore, with current technology, it is impossible to distinguish pneumonia from other lung diseases with certainty using radiological criteria .

Developing accurate,pneumonia detection, algorithms requires large amounts of high-quality labeled data, which can be difficult to obtain. This is particularly challenging in the case of pneumonia, where expert radiologists are required to label the data, and the number of available labeled images is limited. Machine learning, a subset of artificial intelligence, has emerged as a powerful tool for detecting and diagnosing pneumonia from medical images such as chest X-rays.


**Figure 1: Types of pneumonia.**

The above images show three types of pneumonia (Figure 1). We will deal with a binary classification problem, so we will only be interested in whether the image describes pneumonia or not (Figure 2 shows how pneumonia is recognized).



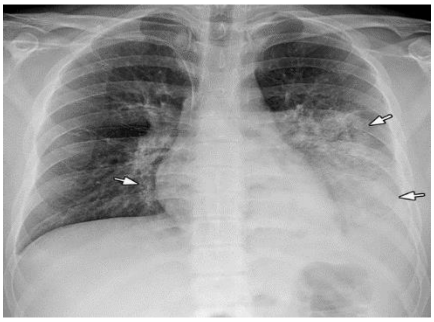


**Figure 2: X rays images showing pneumonia.**

# 3 Proposed Methodology

## 3.1 Input Dataset

Here, the pneumonia-chest X-ray dataset is utilized to gather pneumonia X-ray pictures that consider images from different open sources and which has been overhauled routinely. Here, two datasets are used to train the models for diagnosing pneumonia. The first dataset consists of 5856 images of chest X-rays of which 4273 are pneumonia images and 1583 are normal chest X-ray images. A total of 80% of the data are used for training, producing 4642 images (3418 images of pneumonia and 1224 normal images), 15% of the data are used for testing, producing 919 images (641 cases of pneumonia and 278 normal images), and the final 5% of the data are used for validation (214 cases of pneumonia and 81 non-pneumonia images). The figure 3 and 4 shows the chest X-ray distribution between the train and test set.

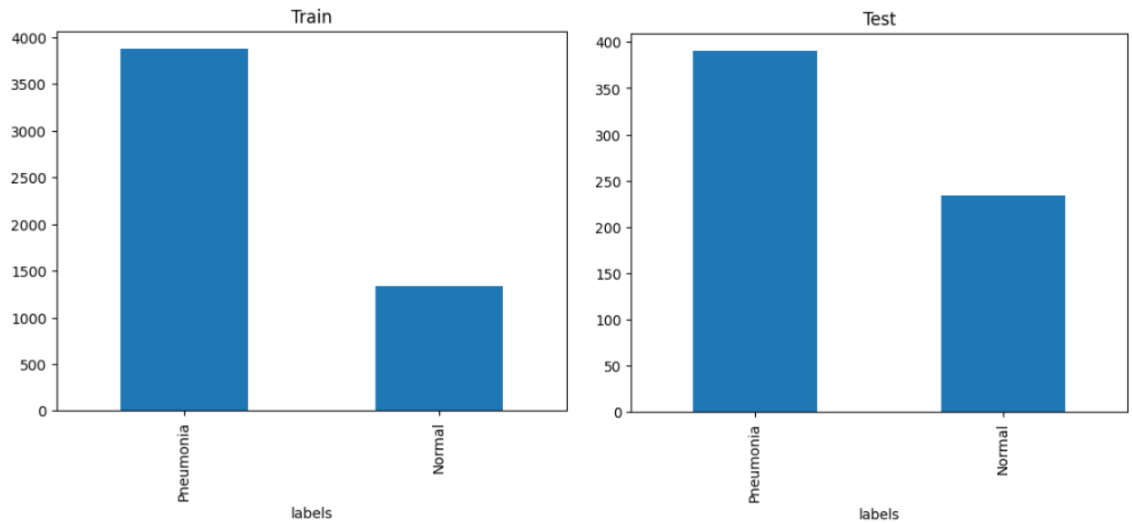



**Figure 3: Bar chart of the distribution of images within the train and test set.**

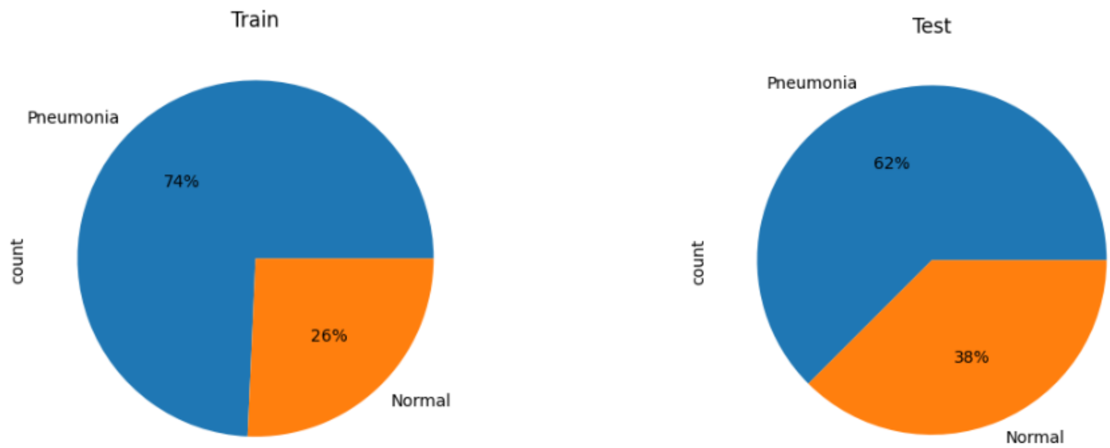



**Figure 4: Bar chart of the distribution of images within the train and test set.**

## 3.2 SIFT Algorithm

While humans can easily identify objects in images despite variations in angle or scale, machines struggle to achieve it. However, through machine learning, we can train machines to identify images at an almost human level, making computer vision an exciting field to work in. In this section, we will use SIFT – an image-matching algorithm in data science that uses machine learning to identify key features in images and match these features to a new image of the same object.

As we can see in the images below, the key features seem to be closer to the lungs.

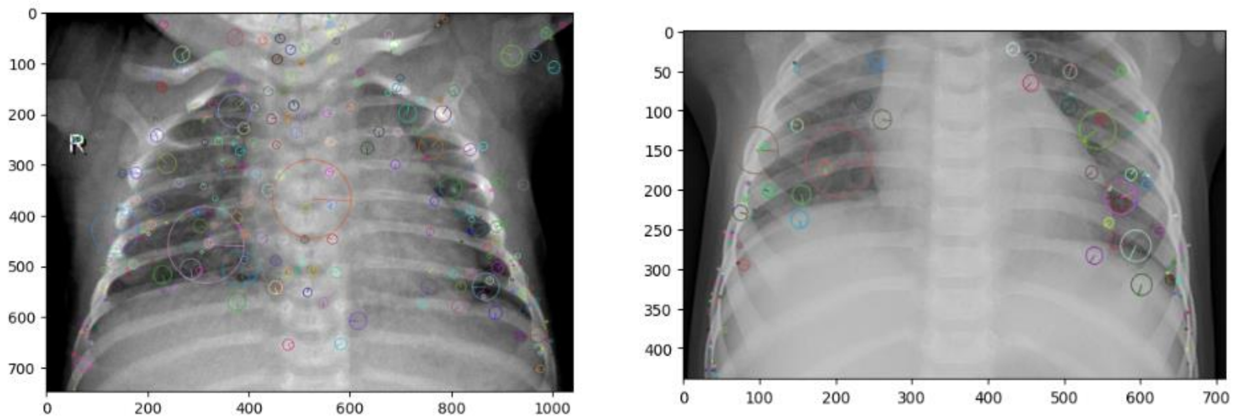


**Figure 5: Key points of x-ray images.**

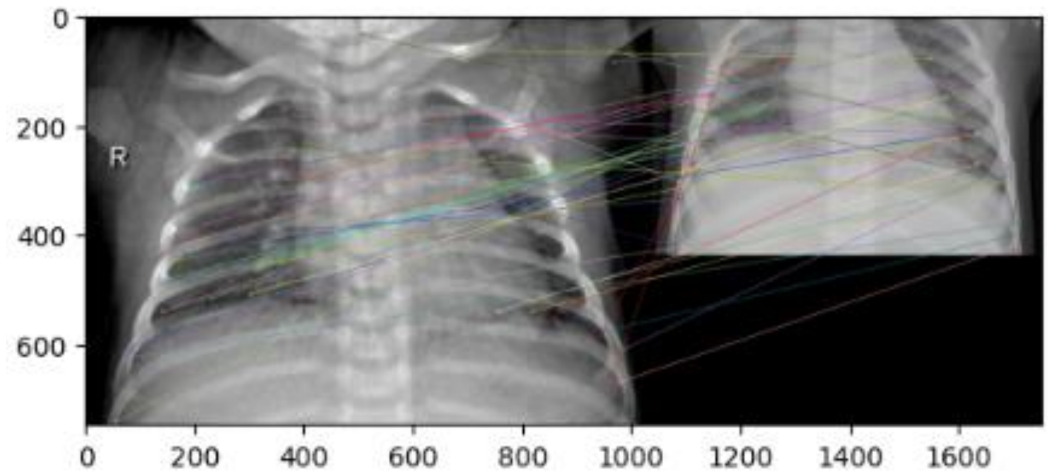


**Figure 6: Key points of x-ray images matching.**

## 3.3 Feature Extraction

### 3.3.1 Data preprocessing

Firstly, we need to unzip the dataset. Secondly, in order to be able to compare the x-rays, we should make the images comparable on same sizes (the selected size is 64x64). By inserting the photo we get three channels (r,g,b), from these three channels we take the average (figure 8). The pixel intensity we get from an x-ray is shown in figure 7.

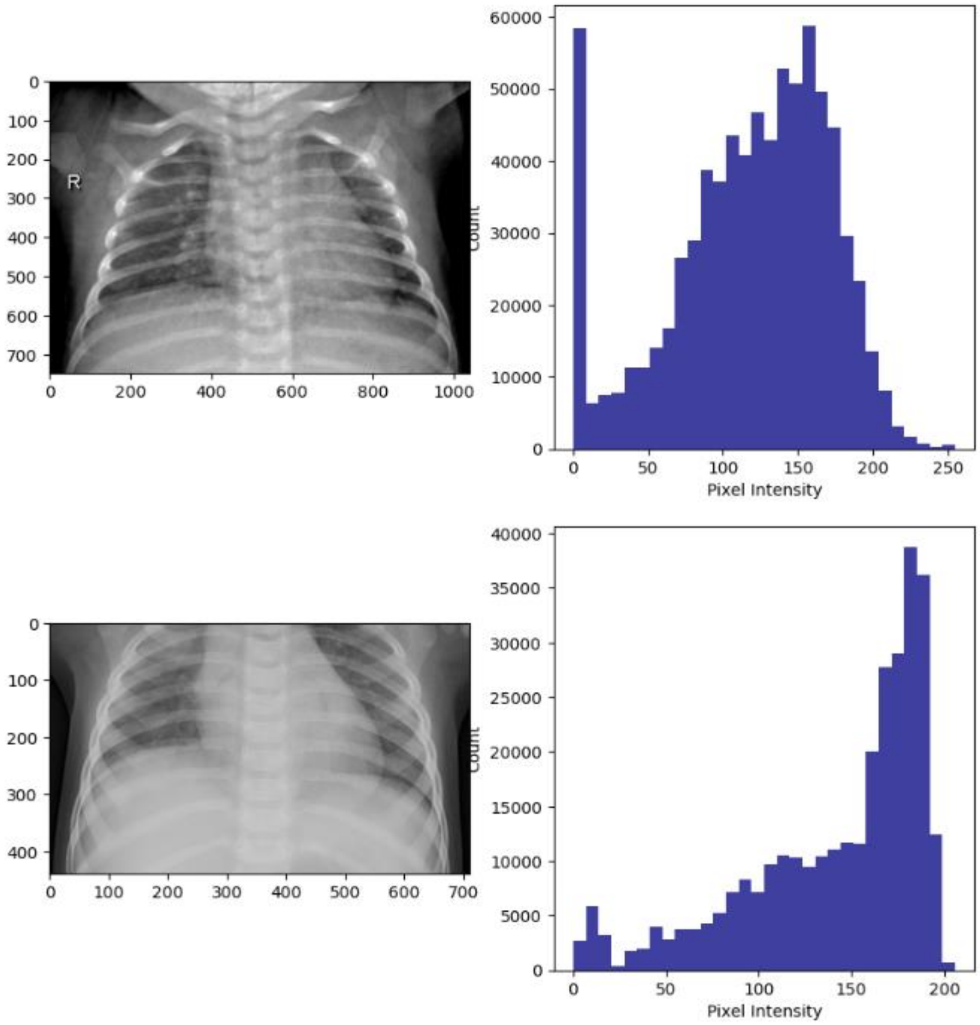



**Figure 7: pixel intensity of x ray image.**

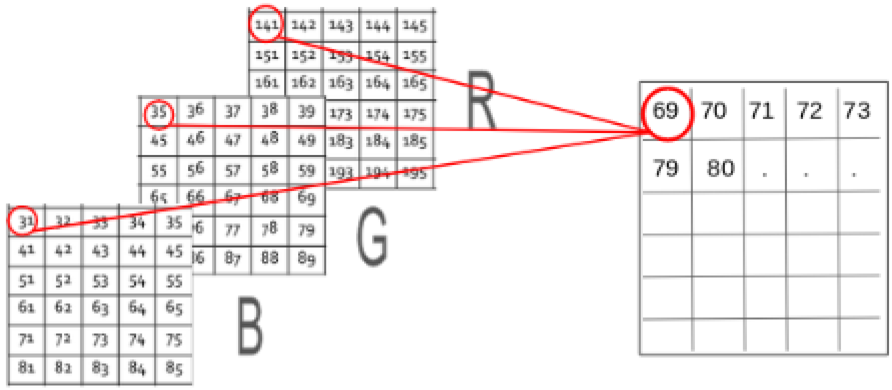



**Figure 8: Feature extraction from data set with Mean Pixel Value of Channels.**

### 3.3.2 Features

The features that we extracted are captured in a 64 x 64 table.

## 3.4 Pneumonia Prediction

To start training the algorithms first we need to normalize data to have zero mean and unit variance.

### 3.4.1 Hyperparameter Tuning

For the definition of the hyperparameters, we take the recall as the first criterion, as we are primarily interested in finding if he has pneumonia and secondly if he does not. So as a second criterion we take Precision.

#### *3.4.1.1 Hyperparameter Tuning for svm with rbf kernel*

The hyperparameters we want to adjust in this case are C and gamma. As we can see from figure 9 and 10 the best option is C=100 and gamma=0.0001.

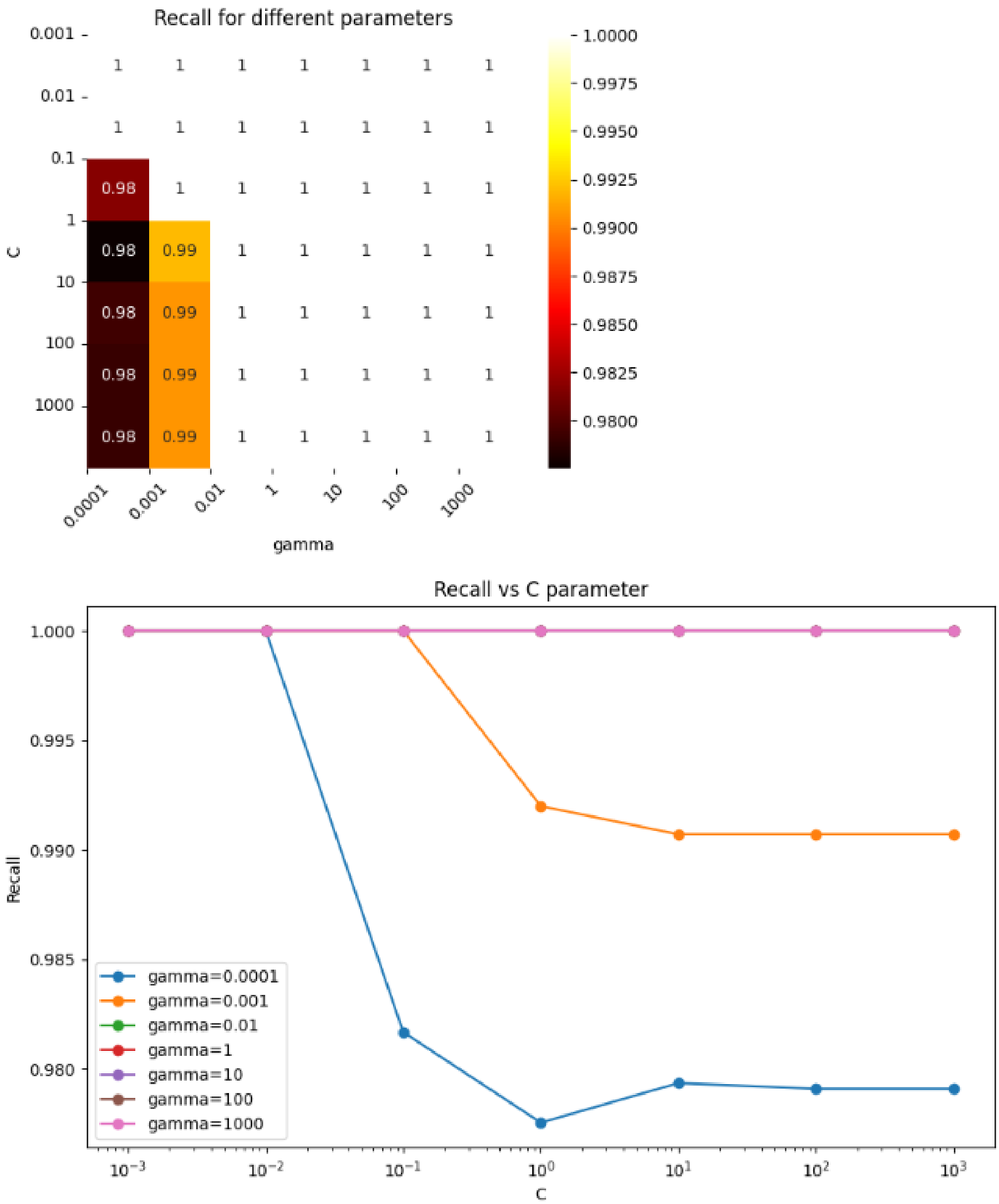
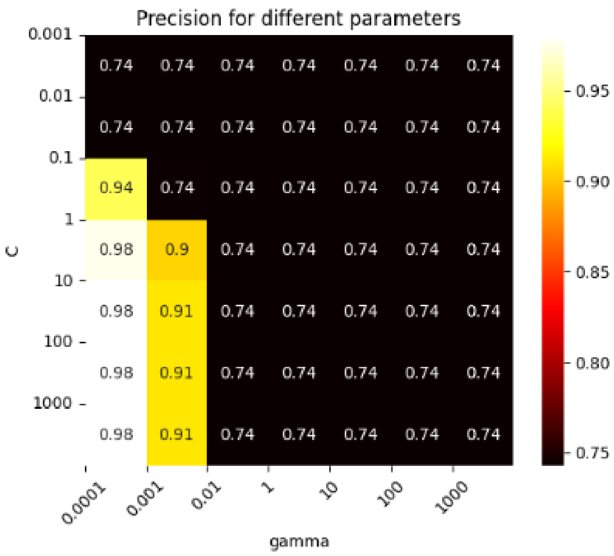



**Figure 9: Hyperparameter Tuning with recall.**



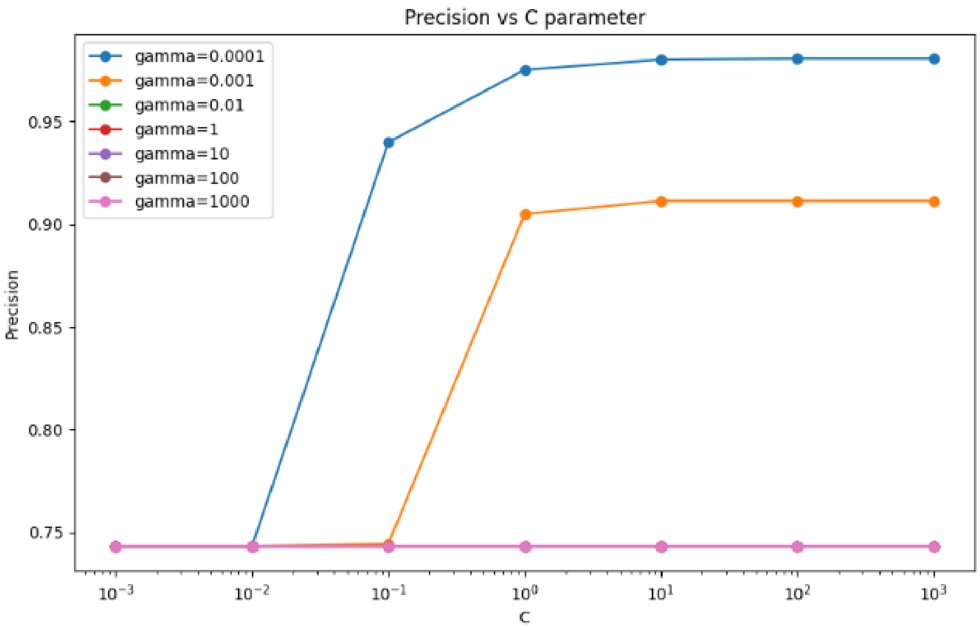




**Figure 10: Hyperparameter Tuning with precision.**

### *3.4.1.2 Hyperparameter Tuning for SVM with polynomial kernel*

The hyperparameters we want to adjust in this case are C and gamma. As we can see from figure 11 and 12 the best option is C=1 and gamma=0.001.

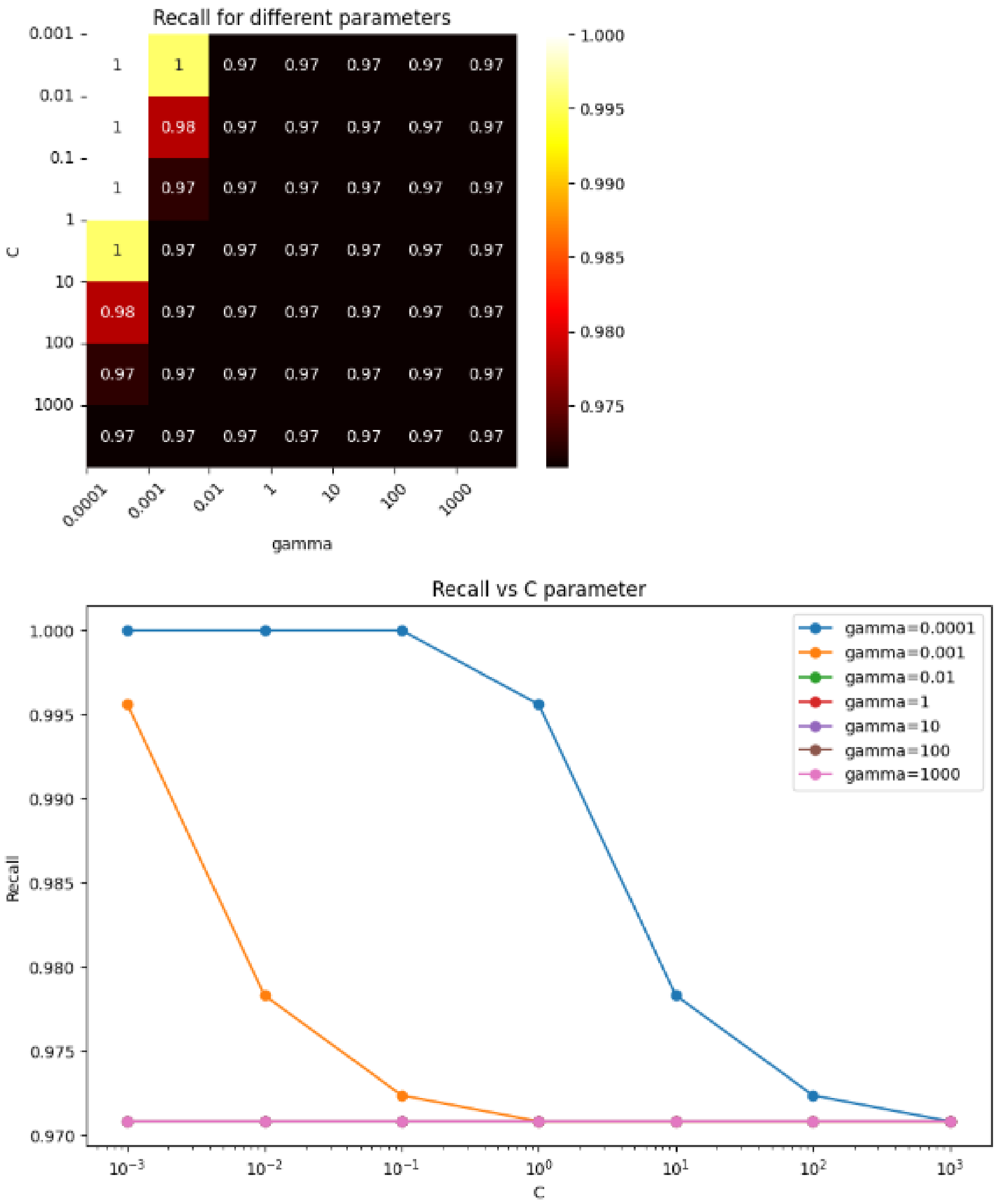
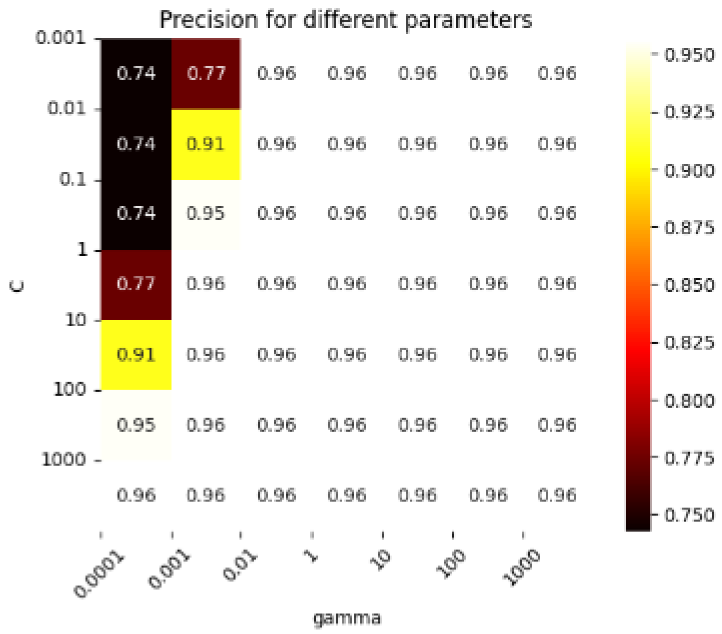



**Figure 11: Hyperparameter Tuning with recall.**



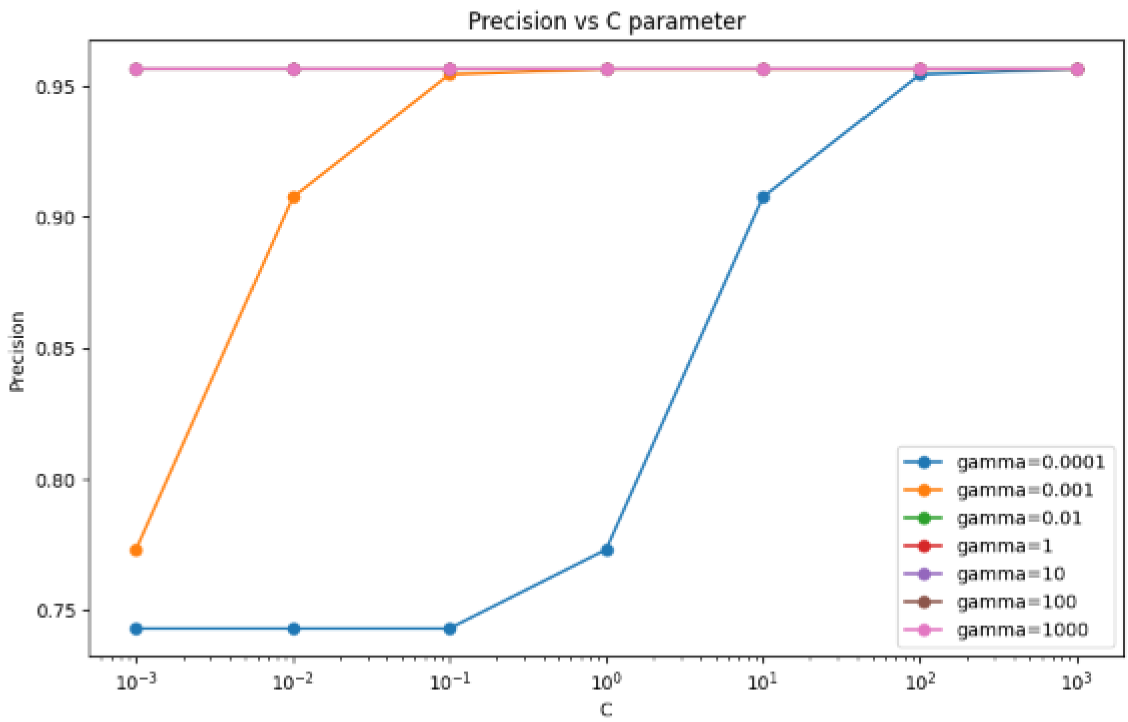




**Figure 12: Hyperparameter Tuning with precision.**

### *3.4.1.3 Hyperparameter Tuning for random forest*

The hyperparameters we want to set in this case are max_features and max_depth. As we can see from figure 13 and 14 the best option is max_features = log2 and max_depth=50.

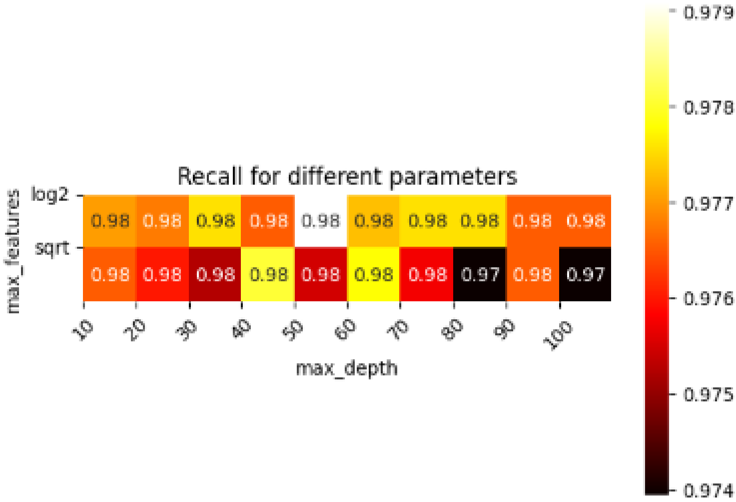



**Figure 13: Hyperparameter Tuning with recall.**

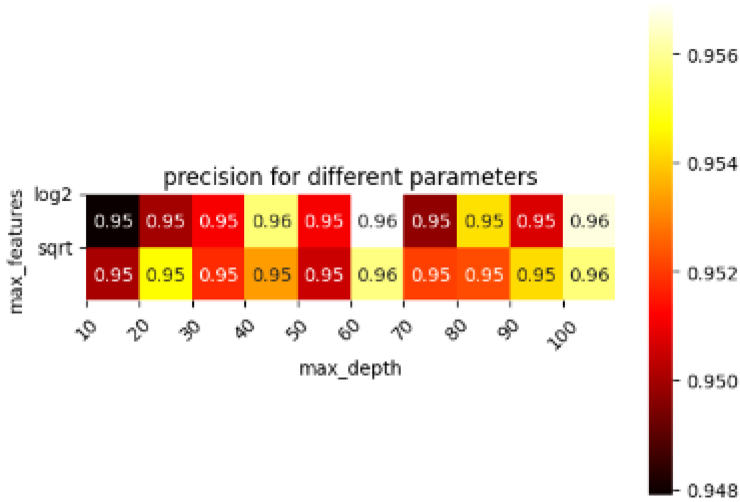



**Figure 14: Hyperparameter Tuning with precision.**

### 3.4.1.4 Hyperparameter Tuning for decision-tree

The hyperparameters we want to set in this case are max_features and max_depth. As we can see from figure 15 and 16 the best option is max_features = log2 and max_depth=20.

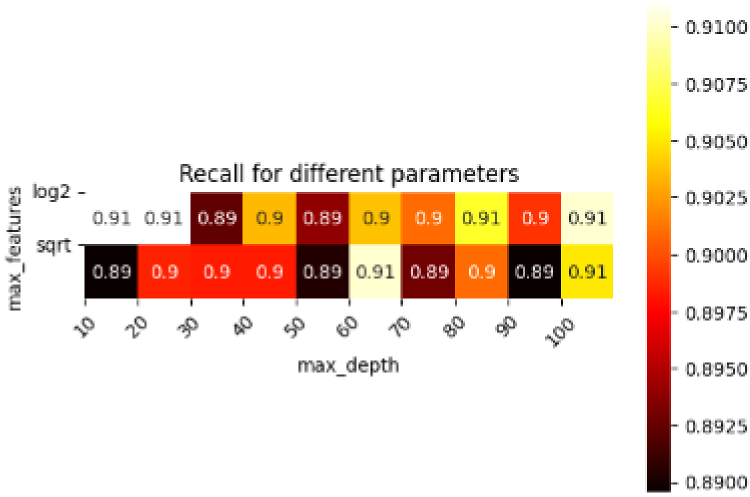



**Figure 15: Hyperparameter Tuning with recall.**

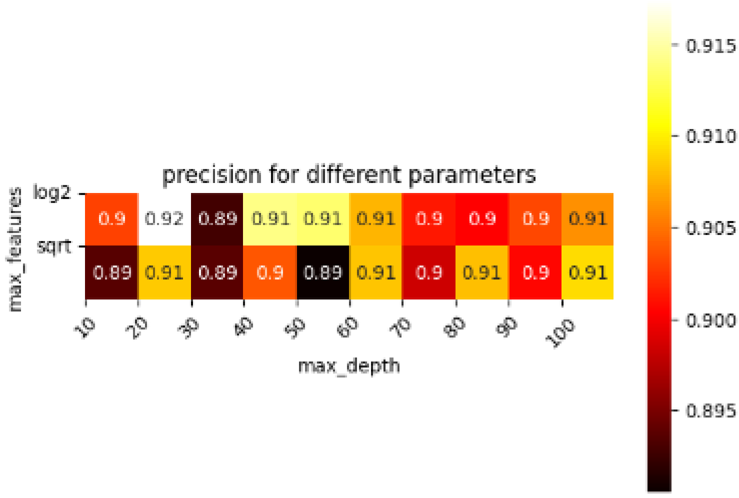



**Figure 16: Hyperparameter Tuning with precision.**

### *3.4.1.5 Hyperparameter Tuning for KNN*

The hyperparameters we want to set in this case are n_neighbors and weights. As we can see from figure 17 and 18 the best option is n_neighbors=7 and weights= uniform.

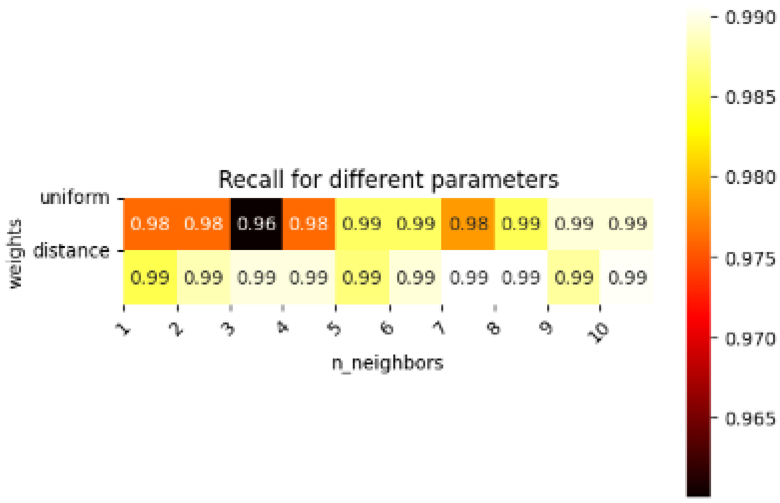



**Figure 17: Hyperparameter Tuning with recall.**

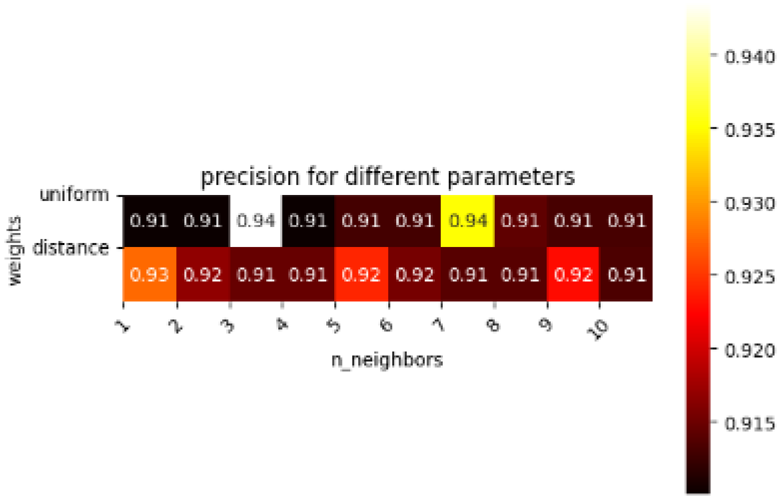



**Figure 18: Hyperparameter Tuning with precision.**

### 3.4.2 Model Selection

The model we choose is Svm with rbf kernel, although it is not the fastest (figure 19) it has the highest f1 score(figure 20,21). F1 is the metric we are interested in as it is the harmonic mean of the precision and recall. Which are the most important metrics when we process medical data.

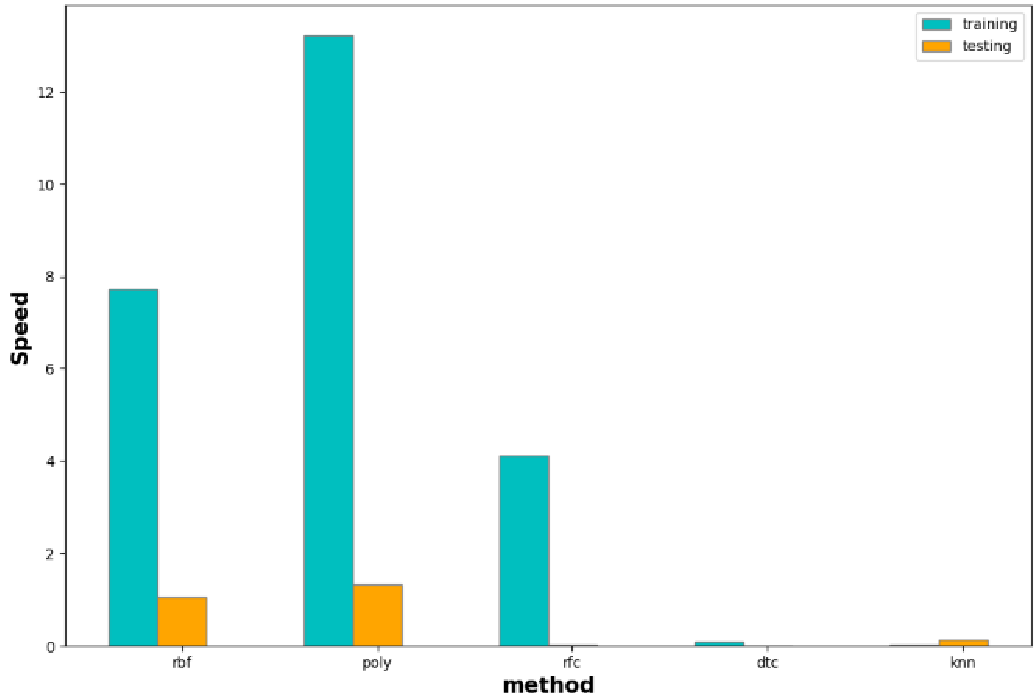



**Figure 19: Speed test.**

```
Support Vector Machine Classifier with rbf kernel
accuracy score is:0.8044871794871795
recall score is:0.9743589743589743
precision score is:0.7723577235772358
f1 score is:0.8616780045351473
Support Vector Machine Classifier with polynomial kernell
accuracy score is:0.7980641025641025
recall score is:0.9743589743589743
precision score is:0.7584830339321357
f1 score is:0.8529741863075196
Random Forest
accuracy score is:0.7916666666666666
recall score is:0.9897435897435898
precision score is:0.75390625
f1 score is:0.8558758314855877
Decision Tree
accuracy score is:0.6842948717948718
recall score is:0.8948717948717949
precision score is:0.691089108910891
f1 score is:0.7798882681564245
KNN
accuracy score is:0.7307692307692307
recall score is:0.9948717948717949
precision score is:0.7003610108303249
f1 score is:0.8220338983050848
```

**Figure 20: Metrics.**

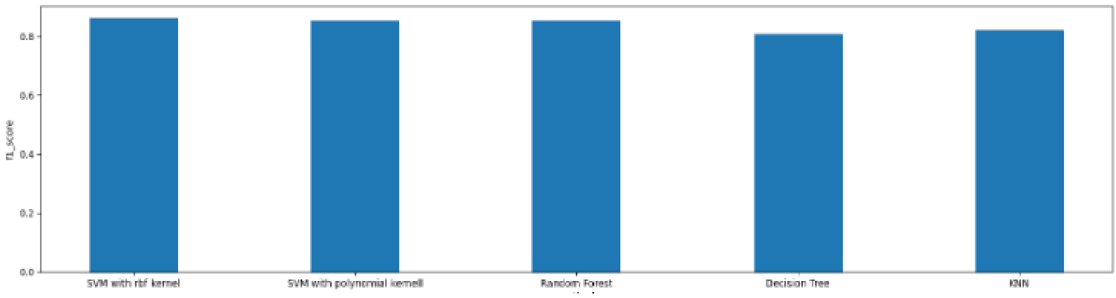



**Figure 21: f1 score bar chart.**

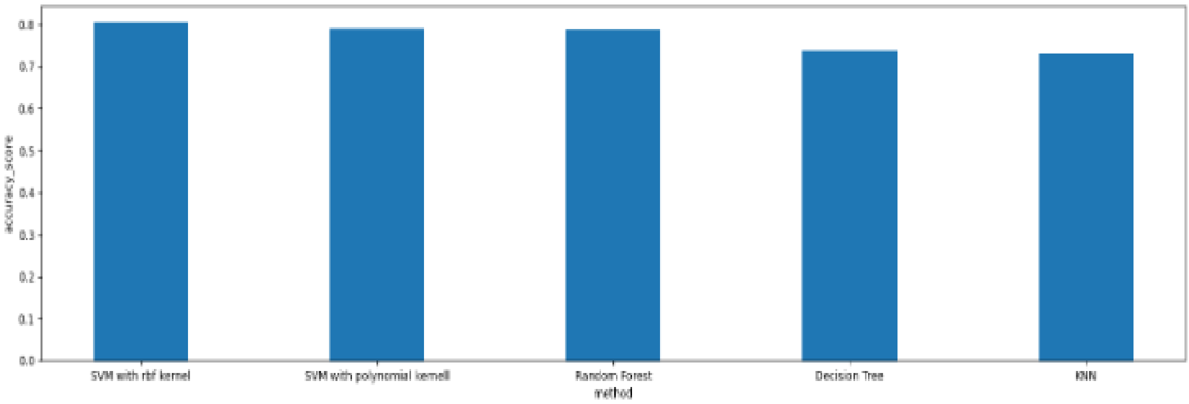



**Figure 21: Accuracy score bar chart.**

In the following confusion matrixes (figure 22,23,24,25,26) it appears that the recognition of an image to not be pneumonia is at best just over 50%. This is because we chose to do hyper parameter tuning according to the recall.

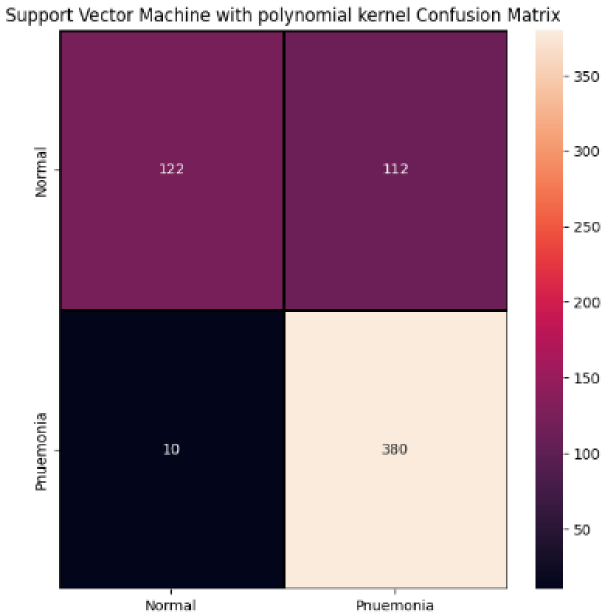



**Figure 22.**

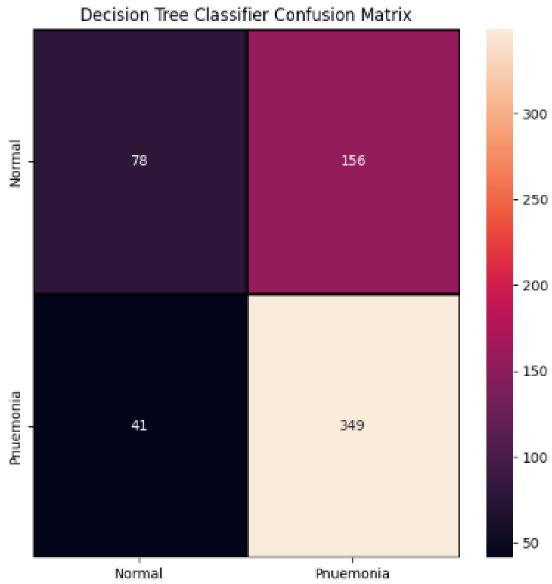



**Figure 23.**

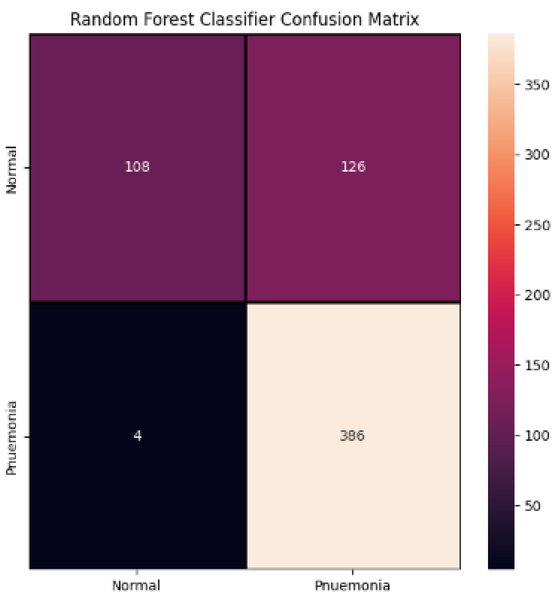



**Figure 24.**

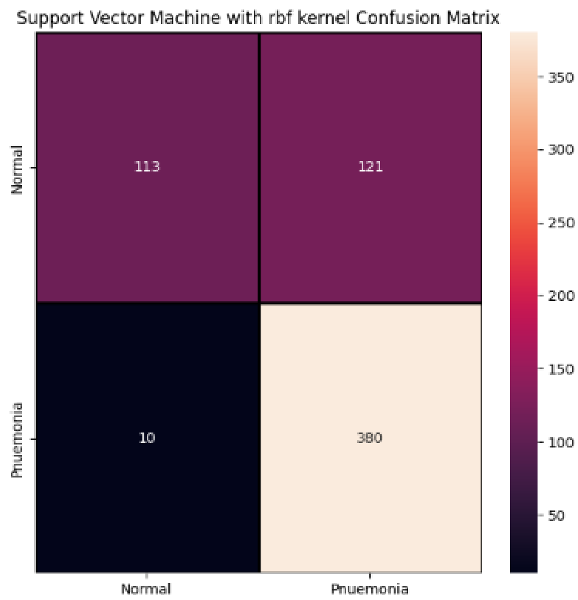



**Figure 25.**

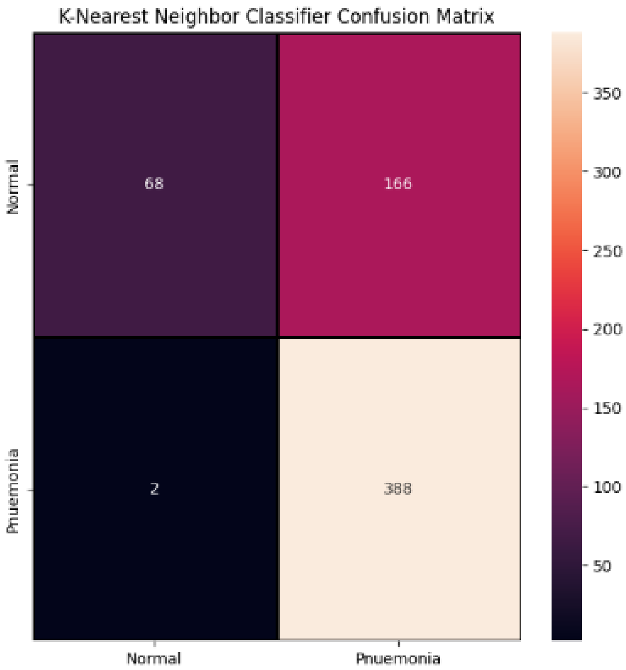



**Figure 26.**

# 4 User interface

In this section, the user interface made for the best algorithm (svm with rbf kernel). The algorithm is trained with all the data.

The first image (figure 27) shows the Interface where we can import the x ray image from the Browse button.

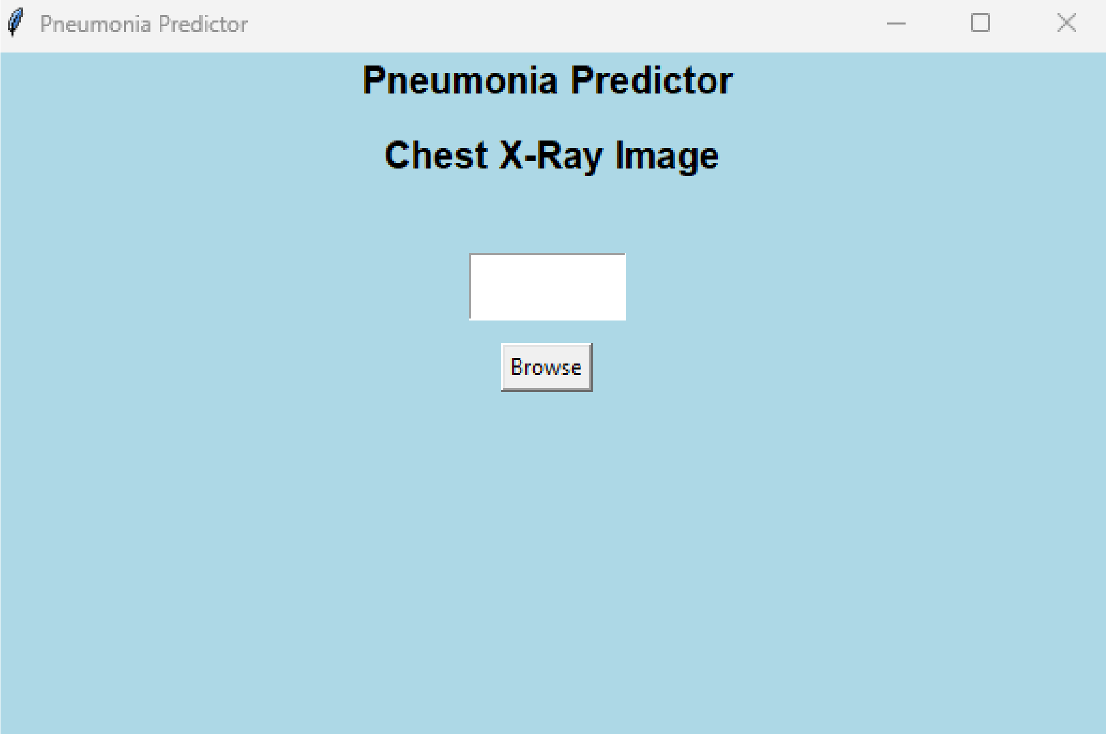



**Figure 27.**

The second image (figure 28) shows how  the results appear.

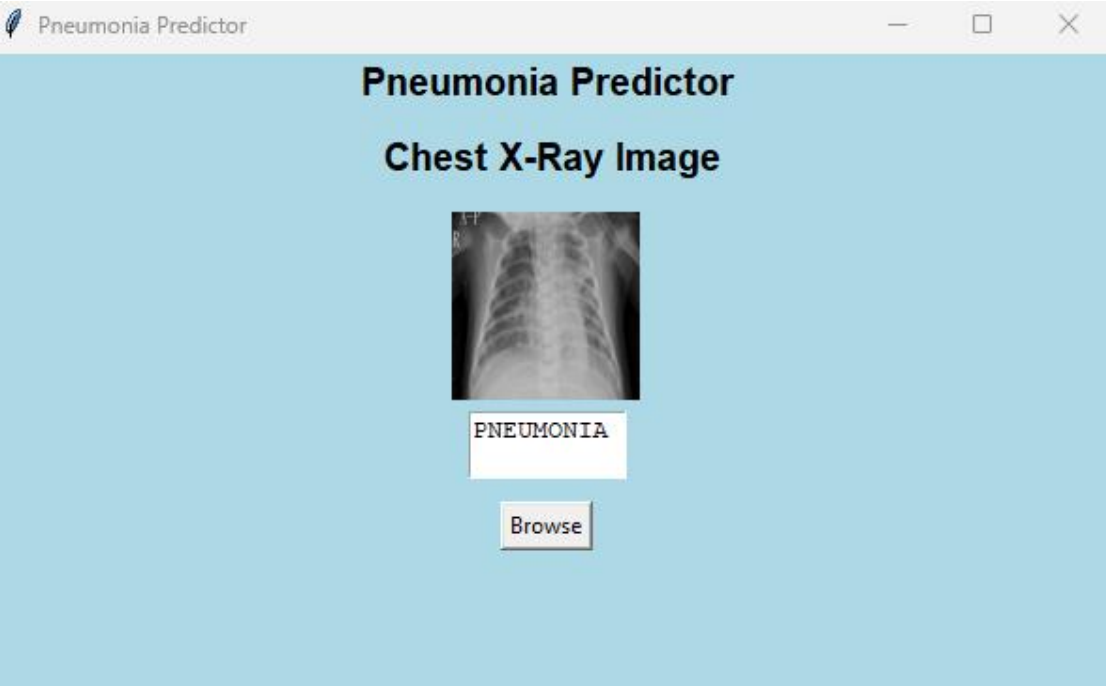



**Figure 28.**

# 5 References

[1] Kumar, A. Man. “C and Gamma in SVM.” *Medium*, 17 Dec. 2018, medium.com/@myselfaman12345/c-and-gamma-in-svm-e6cee48626be.

[2] Saleh, Mana, et al. “Detection of Pneumonia from Chest X-Ray Images Utilizing MobileNet Model.” *Healthcare*, vol. 11, no. 11, 26 May 2023, pp. 1561–1561, https://doi.org/10.3390/healthcare11111561. Accessed 27 Jan. 2024.

[3] Singh, Aishwarya. “How to Detect the Same Object in Different Images Using SIFT?” *Analytics Vidhya*, 9 Oct. 2019, www.analyticsvidhya.com/blog/2019/10/detailed-guide-powerful-sift-technique-image-matching-python/