Diagnosing Pneumonia with Chest Radiographs through Convolutional Neural Networks

OPIM 5509 Final Project

—

Arshiya Anjum
Yongxin Cui
Swati Dhami
Qi Zhou

04/17/2020

Contents

Executive Summary

The objective of this project is to predict whether pneumonia exists in a given image. We do so by predicting bounding boxes around areas of the lung. As reading is increasingly important for diagnosing the disease, we believe our project is tremendously helpful in terms of applying deep learning to reading medical images. To get started, we obtained two sets of data: one is unstructured and graphic and the other is a structured data frame. Both of the two datasets comprise 26,684 observations. Before we train the model, we split the model into training, validation and test datasets in a proportion of 7:2:1. We also set a cutoff point of 0.4 with an optimal function. Following that, we built four models with 2 in gray scale and 2 in an automatic color channel. Among the four models, Model 4 turns out the best results with a precision value of 96% and a F1 score of 89%.

By comparing the model architectures, we find that a few parameters play critical roles. We decreased from 128 to 32, and the accuracy increased by 1-2%. Second, we recommend the kernel sized at 3*3 pixels given the reduced pixels of each image after being processed. Third, the max pooling is better to be maintained at 2*2 pixels and could be enlarged to 3*3 in the layer before being flattened. Fourth, we shouldn't set too high a dropout because of underfitting. We decreased the dropout from 0.5 to 0.2, and the accuracy increased by around 3-4%. At last, an optimal batch size should be set around 30-50. Either exceeding or being underneath this range would affect the performance of the established architecture.

As per the constraint of the project, we would attribute it to the limit of computational capacity of our laptops. If we could maintain the full size for each image and achieve a higher clarity, we believe that the accuracy will also be significantly improved with better defined layers. Overall, we hold that we have accomplished the objective we set when we started the project. The selected model could be immediately put into use and is expected to reduce reading 60% less of the chest radiographs.

Background Information

Importance and Benefits

Around the globe, 15% of deaths of children under 5 years old are claimed by pneumonia. In the US, pneumonia accounts for over 500,000 visits to emergency departments and over 50,000 deaths in 2018, keeping the ailment on the list of top 10 causes of death in the country (Aronson, 2010).

The diagnosis of pneumonia requires review of a chest radiograph by highly trained specialists and confirmation through clinical history, vital signs and laboratory exams. Pneumonia usually manifests as an area or areas of increased opacity on the radiographs (see Exhibit 1). However, reading radiographs is complicated with the presence of a number of other conditions in the lungs such as fluid overload, bleeding, volume loss, and lung cancer. In this case, comparison of chest radiographs of the patient taken at different time points are helpful in making the diagnosis.

This project aims to improve the efficiency of diagnostic services and would significantly reduce the workloads of specialists. We expect the volume of images that need to be reviewed will drop by 60% and thus expedite the delivery of diagnostic results to the patients and reduce the cost as well.

Objective and Methodology

The objective of this project is to build an algorithm to detect a visual signal for pneumonia in medical images. Specifically, the algorithm needs to automatically locate lung opacities on chest radiographs. We associate the diagnosis by relating it to the presence of bounding boxes around areas of the lung. Samples without bounding boxes are negative and contain no definitive evidence of pneumonia while samples with bounding boxes indicate evidence of pneumonia. When making predictions, a model is expected to detect bounding boxes as much as possible, and we will choose the best model with the highest capacity in doing that through a careful assessment of each of them.

Data Description

Data Retrieval

We obtained both the image data and the structured data frame file from the Radiological Society of North America, which has a database of radiological images archived by disease. The images are in DICOM format with an average size of around 12 MB. The image sample is composed of 26,684 chest radiographs that were taken from 2007 to 2012 at its member hospitals and clinics in the US.

Data Dictionary

There are five variables in the input dataset and one target variable. Image ID is the primary identification. X and Y are the coordinates that locate a lesion on the image, while width and height together denote the size of the suspicious area. Target is the diagnosis result, with "1" denoting positive and "0" negative. The input variables are continuous data and the target variable is categorical. The original data of one single image is in a shape of (1024, 1024, 1) with 1024*1024 pixels on a grayscale channel. Due to the limit of computational capacity, we have to resize to 340*340 each for converting images to arrays.

Exploratory Data Analysis

Among the 26,684 observations, 6,012 are diagnosed, accounting for 22.53% (Diagnosis Rate). Then among the 6,012 observations, we conducted an exploratory data analysis with the four input variables. According to the boxplots of width and height (see Exhibit 2), a most typical lesion is 216 pixel wide and 314 pixel high, with a standard deviation of 59.5 pixels and 154.7 pixels respectively, leading to a median size of 6 squared pixels (see Exhibit 3). Based on the distribution of coordinates, the lesions are almost evenly located between two thoraces (as seen in Exhibit 4 and Exhibit 5), with the right thorax slightly less than the left one, while they are located between 200 and 600 pixels on the vertical axis (see Exhibit 6).

Data Preprocessing

Split the Dataset

In this project, the dataset is split into 3 parts, training, validation and test. 70% of data was attributed to training, 20% was for validation and 10% was for test dataset. In order to ensure the performance of the model, the target ratio in each part is similar. 3.3 in training, 3.6 in validation and 3.4 in test. The X-rays are also split into 3 parts accordingly. The value used to match the dataset and the X-rays is the name of the patient and the name of x-days.

Load X-Ray Files

The file containing the information of the image is a dcm format, a special file format used to save medical images. As the figure shows, the information of patients could be found in this file. All of the information may be useful, but in this project, only the X-ray information would be used.

Resize and Transform X-Rays

There are more than 26,000 images in this project and the size of each image is 1024 *1024, which is hard for the personal computer with limited memory to process. Therefore, our team resize each image to 340 *340. X-Rays are white and Black style, so the channel for the array is supposed to be 1. However, the challenge here is that the CV2 method will transform the x-rays into 3 channels. And there would be no channel information if we set it to be white flag. As a result, the method used to load x-rays is TensorFlow image loading, which can transform arrays into the desired format.

Model Description

Architecture

The best model architecture of our project contains 10 layers with 3 convolution layers, 3 max pooling layers and 4 dense layers. After each convolution layer and max pooling layer there is a dropout of 0.2. Even after every dense layer there is a dropout of 0.2. Adding different dropouts varying from 0.2 to 0.9 values between layers and the dropout of 0.2 gives the best learning rate. The convolution and dense layers have relu activation. The three convolution layers have 320, 8256, 36928 trainable params, kernel size of 3, 2, 3 and no. of channels of 32, 64, 64 respectively. The four dense layers have 6422656, 16512, 8256, 65 trainable params and kernel size of 2,2,3 respectively. The output layer has sigmoid activation. The model has total trainable params of 6,492,993. Convolution layer(filters =32, kernel size =3), Maxpooling layer(size =2), Dropout =0.2, Convolution layer(filters =64, kernel size =2), Maxpooling layer(size =2), Dropout =0.2, Convolution layer(filters =64, kernel size =3), Maxpooling layer(size =3), Dropout =0.2, 2 Hidden layers (128 nodes), Dropout =0.2, 1 Hidden layer (256 nodes), Dropout= 0.2, Output Node (Sigmoid activation). The image of the model architecture is put in the model appendix (see Exhibit 7).

Model Fitting

Model is compiled with a loss of binary cross entropy and accuracy metrics. Model was fit on training data and evaluated on validation data for 30 epochs and a batch size of 100. Early stopping was applied with a patience of 5 from minimum validation accuracy because of which model was stopped on after sixth epoch. The validation accuracy of this model was 80.96%. Using an optimizer ADAM enables us to use adaptive learning rate for different parameters which helps the model to learn different patterns.

Results

Initially the image of size (340,340,1) was taken and local, global features were extracted for the classification problem. We have observed that the relu activation used in the different layers helped our model to learn non linear patterns in the image data. Before sending into the dense layers there are 50176 nodes. Adding different dropouts varying from 0.2 to 0.9 values between layers and the dropout of 0.2 gives the best learning rate. Using an optimizer ADAM enables us to use adaptive learning rate for different parameters which helps the model to learn different patterns.

Different models were built with varying layers and dropouts out of which the best model was described in the above section. The accuracy of our model for training, validation and testing are 81.26%, 80.96%, 79.61% respectively. The proportion of false negatives, false positives, true positives and true negatives is constant over all partitions of the data. The validation loss and training loss is the same in 4, 5, 6th epoch.

The validation accuracy and training accuracy is the best in 3rd epoch. The precision values are 96%, 94.77%, 95.11% for training, validation and testing respectively. The F1 scores are 88.76%, 88.66%, 87.85% for training, validation and testing respectively.

The cost of telling people with pneumonia that they don’t have is more than telling people that they tested positive even if they won’t. The sensitivity values are around 82.53%, 83.29% and 81.61% for training, validation and testing respectively and the specificity values are around 70.01%, 61.44%, 61% for training, validation and testing respectively. The False Negative Rate is 17.47%, 16.17%, 18.39% for training, validation and testing respectively. This shows that our model is detecting patients with pneumonia more accurately which is an advantage because we are not telling people with pneumonia that they don't have.

The values of all measures shown in the table below and the graphs of validation loss vs training loss and training accuracy vs validation accuracy (see Exhibit 8) show the model was not overfitted since the measures are almost equal for all the three sets of data and the losses are also almost the same. The measures generated in the classification report (see Exhibit 9) and confusion matrices (see Exhibit 10) for all the three datasets are all showing that the model achieves a good classification prediction.

Discussion

Our goal in this project was to detect the images with pneumonia, usually air or light denser does not absorb X rays hence they are black in the images. If a person has bacterial or virus pneumonia the lung cells will be filled with liquids because of which they appear grey in the CXR. But it is not definite to identify the lesions in the lungs because they can be at any part of the lung. These images are produced with great clarity 1024 pixels but due to limitation of computational capacity of our laptops we reduced the clarity. Later these images were converted to numpy arrays on the local computer and then uploaded on to the drive for further analysis. If we would have used computers with larger computational capability then we would have achieved better accuracy. This is a limitation which needs to be taken care of in future in order to achieve good accuracy.

The images initially had three channels i.e. an RGB image and models were fit on these images with which we achieved a highest accuracy of 78%. It was thought that the problem was more about detecting grey areas inside the lung area which was supposed to be black. Hence the images were converted to grayscale and models were built which has increased our accuracy by 2%. Therefore, we achieved a final accuracy of 80.96% due to this conversion. This accuracy was achieved by using a simple neural network without using any pre trained convolutional neural network layers and complex neural networks with hundreds of layers. If there is an access to more advanced computational ability more sophisticated neural networks can be used to increase the accuracy. We can also enhance the existing image quality if there is access to better facilities.

Various models were built for classifying images by setting different architectures by changing the number of layers from 8 to 10 , by changing the filter size from 32 to 64 i.e. no of channels , by varying kernel sizes among 2 x 2 and 3 x 3 various models to extract different features, by changing the number of nodes in layers (256,128,64,32) and by changing the batch sizes of images from 50 to 100. After running the models with various changes appropriate batch size, hidden units, kernel sizes, layers were selected such that there was no overfitting and there is a good learning rate in the model. The best model is chosen based on accuracy and sensitivity because of the high cost associated with sensitivity as discussed in the previous section.

The percentage of people who have pneumonia in this dataset was only 22.53% and that is clearly observed in the confusion matrices for training, validation and testing data. While evaluating the model different cutoffs such 0.3, 0.4, 0.5, 0.6, 0.7 were tried to predict the positive pneumonia patients to achieve better classification accuracy and 0.4 was decided as the ideal cutoff with best accuracy among other cutoffs tried. Ultimately, we were able to identify pneumonia from the lung X-ray images, which is contributing to the huge number of patients every year with an accuracy of 80.96 %.

Conclusion and Reflection

Accomplished Achievements

Overall, we hold that we have accomplished the objective we set when we started the project. The selected model could be immediately put into use and is expected to reduce reading 60% less of the chest radiographs. Specifically, the best accuracy of 80.96%, F1 Score of 88.66% occurs at a cutoff of 0.4. Looking at the validation and training curve loss it is evident that there is no overfitting. Able to classify the CXR with decent accuracy with images of reduced pixel and simple neural networks. Best model was identified based on the accuracy and sensitivity values and the architecture was decided by trying different options of batch sizes, epochs and hidden units. The sensitivity values are around 82% and the specificity values are around 61%. Converting images to grayscale and modelling has increased the accuracy by 2%

Next Steps

Because of the limit of computational capacity, we didn't have the full size for each image. If we could maintain a higher clarity of the images, we believe that there will be a lift in the accuracy as well. With the use of pre trained layers of neural networks and deeper neural networks with hundreds of layers there might be a chance where the neural network can learn more and classify more accurately. Enhancing image quality to more than what is given i.e. 1024 pixels by modern image quality improvement techniques will definitely contribute for better classification. Types of lesions for different conditions other than pneumonia can also be classified, broadening the range of this challenge and making it a multi class classification problem which can be used on an expanded horizon.

References

1. RSNA Database: https://www.rsna.org/en/practice-tools/data-tools-and-standards/image-share-validation-program
2. A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3):229–236, may 2010.
3. J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In ICCV, 2015.
4. 2.S. Bird, E. Klein, and E. Loper. Natural language processing with Python." O'Reilly Media, Inc.", 2009.

Appendix

Exhibit 1: Normal and Abnormal Chest X-Ray

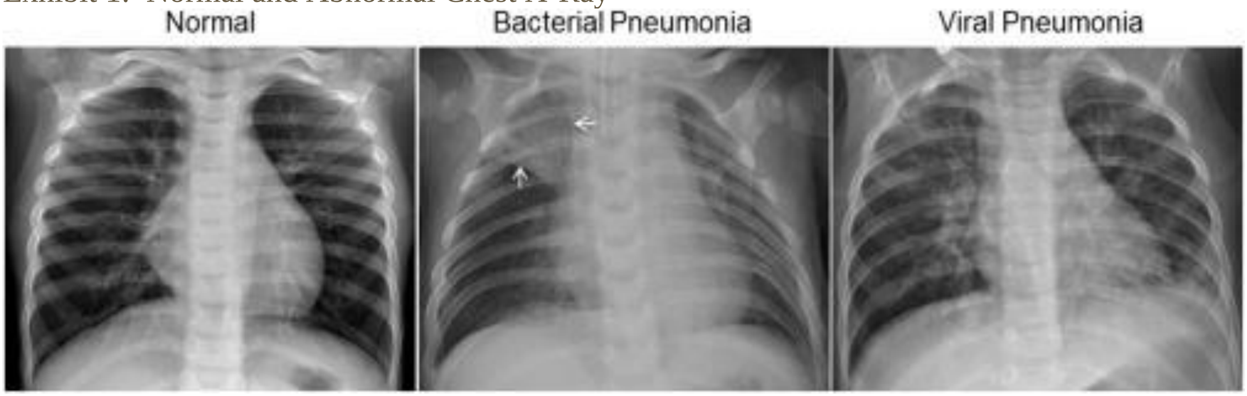


Exhibit 2: Distribution of width, height, x, and y

```
[30]: <matplotlib.axes._subplots.AxesSubplot at 0x1c6e0e76c50>
```

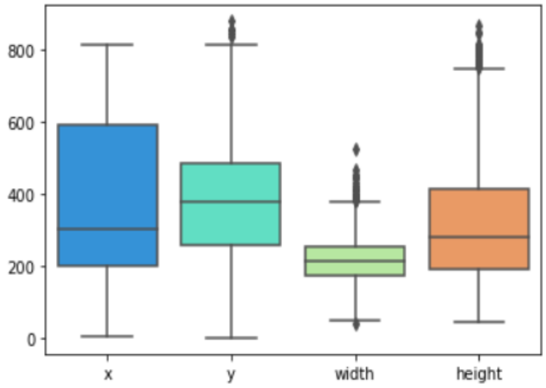


Exhibit 3: Area of Bounding Boxes

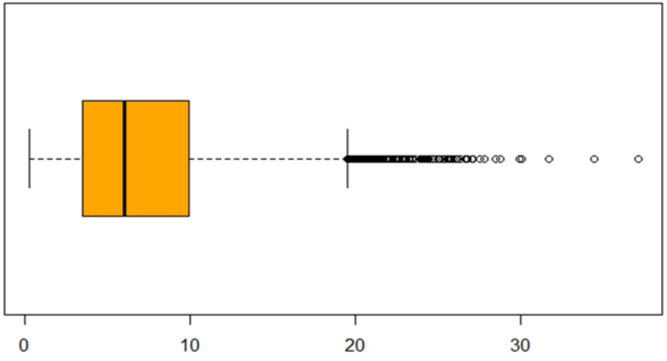


Exhibit 4: Locations of Diagnosed Area

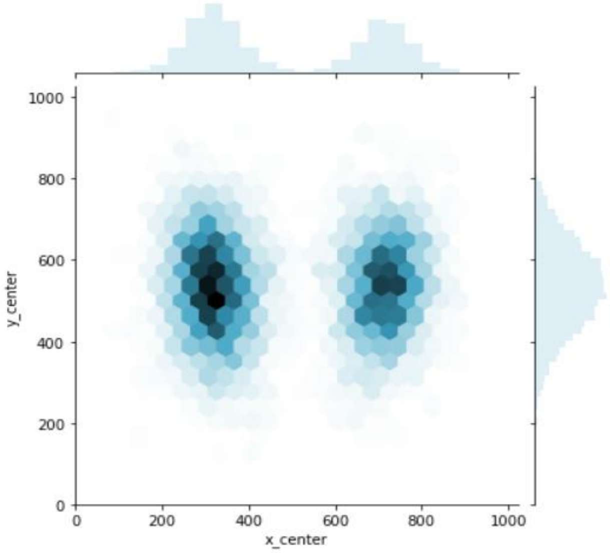


Exhibit 5: Occurrences at the X Axis

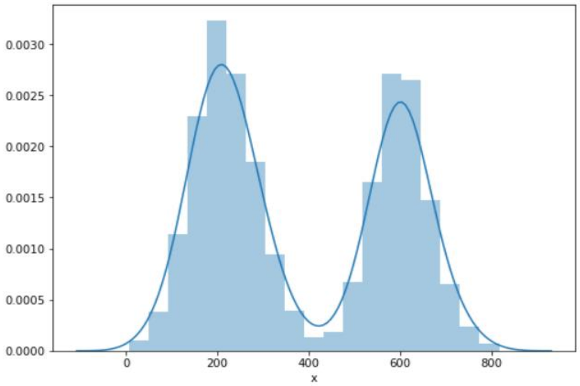


Exhibit 6: Occurrences at the Y Axis

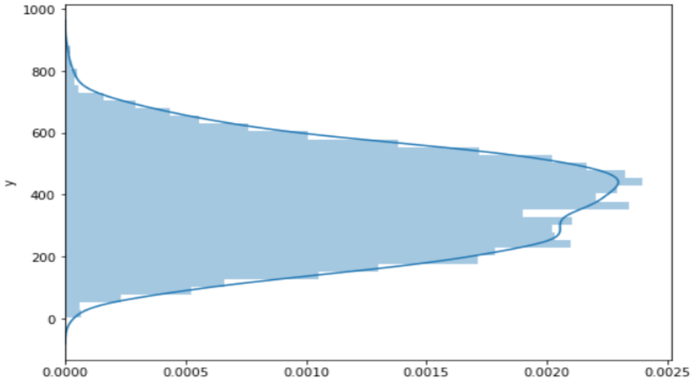


Exhibit 7: Model Architecture Display

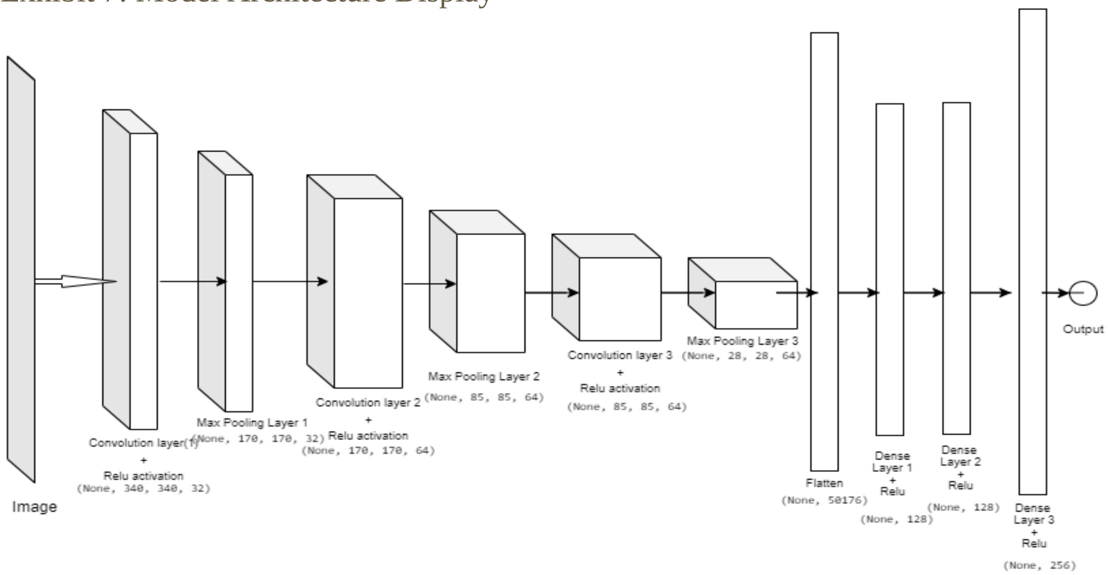


Exhibit 8: Accuracy and Loss Plots

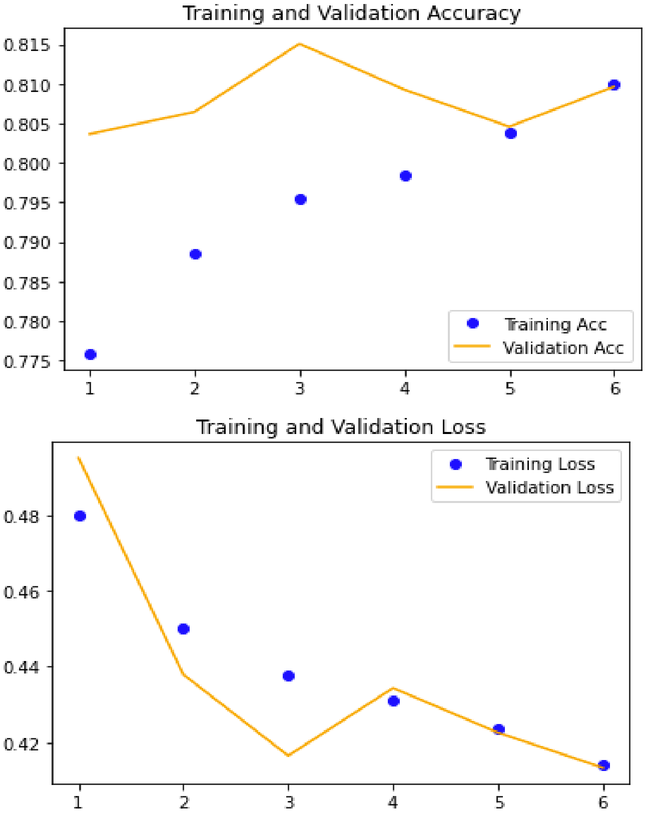
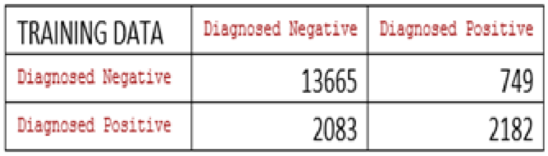
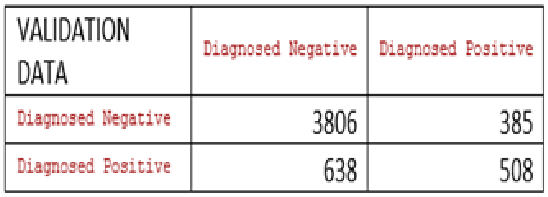


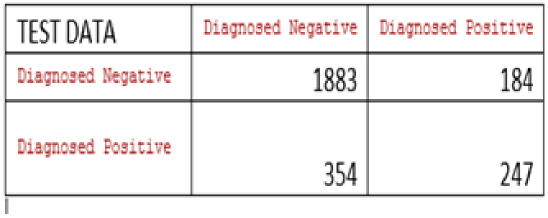
Exhibit 8: Confusion Matrix



TRAINING DATA	Diagnosed Negative	Diagnosed Positive
Diagnosed Negative	13665	749
Diagnosed Positive	2083	2182

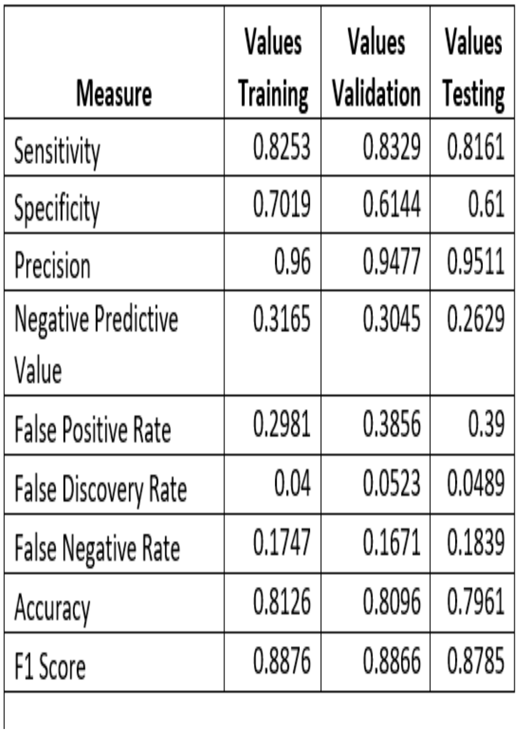


VALIDATION DATA	Diagnosed Negative	Diagnosed Positive
Diagnosed Negative	3806	385
Diagnosed Positive	638	508



TEST DATA	Diagnosed Negative	Diagnosed Positive
Diagnosed Negative	1883	184
Diagnosed Positive	354	247

Exhibit 10: Classification Report



Measure	Values Training	Values Validation	Values Testing
Sensitivity	0.8253	0.8329	0.8161
Specificity	0.7019	0.6144	0.61
Precision	0.96	0.9477	0.9511
Negative Predictive Value	0.3165	0.3045	0.2629
False Positive Rate	0.2981	0.3856	0.39
False Discovery Rate	0.04	0.0523	0.0489
False Negative Rate	0.1747	0.1671	0.1839
Accuracy	0.8126	0.8096	0.7961
F1 Score	0.8876	0.8866	0.8785

- End-