TABLE OF CONTENTS

TASK 3

3.1 Report Findings From Data Analysis Plan.

3.1.1 Introduction

The purpose of this report is to conclude the data that have been analyzed. How we use various sources of data either in tabular or raw data been analyse using variety of software. The performance of the model with percentage of accuracy also will be included. From here we can see either this project can be brought further in order to help doctors or scientists to gain fast result while examining people with disease or non-disease, or either it is chronic or non-chronic diseases. In this report also we are going to include the information about the data collected, the method that we used to gain specific result either it is suitable or not to be implemented in the future industry project.

3.1.2 Data

Data that been used mostly in image format. There are also from data of image that we got with some details of patient with some important attributes such age, gender, smoker or non-smoker and others. But in order to implement faster model on classification, we use COVID-19 patient's brain MRI images. Below are some of the data images that been use for training the model.

Below shows some of the raw data images gathered from other resources and we convert it into testing data images. The figure 1 shows the testing result of the possibilities for patients to have positive COVID19 and negative COVID19.

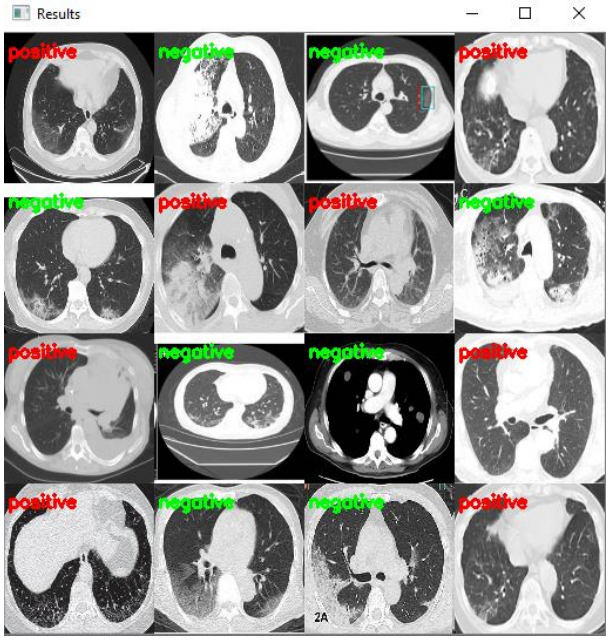


Figure 1: Testing Results on Brain MRI images

3.1.3 Method

Method that we used to gather the data is by browse to any website that contain patient's data. The data comes from any format, since we are work on image processing, so the data images collected are more focused on chest x-ray and brain MRI. Image data also included in qualitative research where we must search, observe and record them as a big dataset. We also do online tracking such as social media tracking to gather data since everyday there will be new disease with new inspection and from there we get to used it as raw data and turn them into valuable insights that we can use to enhance strategies, products and business decisions. Method to do analyse the data is by performing 4 analysis method which are K-nearest Neighbor (KNN), Standard Vector Machine (SVM), Naïve Bayes with Gaussian, and Neural Network (Deep Learning – ResNet).

3.1.4 Algorithm Used

Image that we are going to select as model is chest x-ray and brain MRI image. In order to process these model. Feature extraction, image segmentation, context navigation, isolation filtering on images will be used as for implementing the classification algorithm.

- **Context navigation**

  Suspicious spot found after make some filter and using volume viewer in Matlab. From here, we can question ourselves either the image contain a tumor or cavity or spots. Then the classification algorithm takes place where we can separate into positive case or negative case.

- **Isolation filtering**

  Isolation filtering using volume viewer tools in Matlab software. From the original image of the x-ray or brain MRI, they will be isolated using isosurface where all the bones will be excluded since we would like to focus on what is inside the organ (lungs) and brain.

- **K-nearest Neighbor (KNN)**

  KNN were implemented with number of neighbors of 5. Tuning this value can lead to a better model performance

- **Standard Vector Machine (SVM)**

  The decision boundary maximizes the distance from the nearest data points.

- **Naïve Bayes with Gaussian**

  Naïve Bayes can be used in binary or multiclass problems. This were implemented as binary class problem in this project

- **Neural Network (Deep Learning – ResNet)**

  A model called Residual Neural Network is used under Neural Network algorithm. This is one of the famous yet most effective method in Machine Learning.

3.1.5 Performance Metrics

The performance metrics describe the performance of the model. An important aspect of evaluation metrics is their ability to differentiate between model outcomes or final results. We have seen many aspiring data analysts and scientists not even concerning to study how strong their models are. However, not all performance metrics are suite to be implemented for all type of models. In this case, to evaluate the performance of the model and study on how strong our model is by focusing on the accuracy. The higher accuracy of performance can lead to better decision or result in predicting the next new model. As for selecting the suitable performance metrics to be implemented on the model, we can see through the types of data that we used, what are the attributes that suite to be used, and what performance that we are looking for.

Based on our project, the model will be evaluated by using 4 types of performance metrics that suitable for image processing. Most of them are based on the confusion matrix formula and it is suitable for classification algorithm. The performance metrics are Precision, Recall, Accuracy and $F_1\ score$. Below in Table 1 shows the performance metrics and their respective formula.

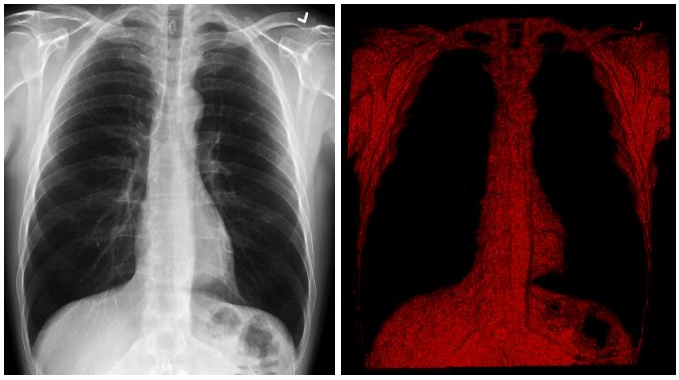
Performance metrics	Formula	Explanation
Precision	$\frac{tp}{tp + fp}$	➢ tp - true positive (real positive) ➢ tn - true negative (it's not a negative) ➢ fp - false positive (detected negative, while it's positive) ➢ fn - false negative (detected positive, while it's a negative)
Recall	$\frac{tp}{tp + fn}$	
Accuracy	$\frac{tp + tn}{tp + tn + fp + fn}$	
$F_1\ score$	$2 \times \frac{precision\ \times recall}{precision + recall}$	

3.1.6 Model and Evaluation

3.1.6.1 Using Matlab software:

3.1.6.1.1 Isolation Filtering and Context Navigation

Below shows 3 types of picture, 2 chest x-ray and 1 brain MRI image. Chest x-ray 1 shows the normal lungs which this patients has no disease detected in his lungs. We can see that after isolation filtering about 45% (red-black colour), the lungs is clear, no spots or any bubble appeared in his lungs.

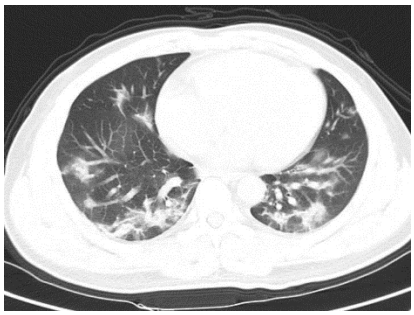


Chest X-ray 1: Normal chest x-ray

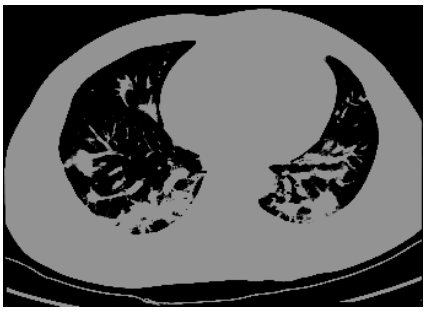
(Radiopaedia 2019) Pulmonary cavities are thick-walled abnormal gas-filled spaces within the lung. The first picture is original x-ray for pulmonary cavitation. Second picture shows result after done linear filtering. In order to reduce noise in the picture such as the bone structure, isolation filtering been done and the results is as shown in third x-ray. Now we can start context navigation where the thick-walled abnormal gas-filled spaces is located. As end result, this patient has a well-defined cystic lesion in the right upper zone. It has a thick wall and an air-fluid level.

Chest X-ray 2: Pulmonary cavitation disease

Below shows the brain MRI image which has been chose to implement our model. The first image shows the original image, we can see there is noise in the image where all nerves can be specified as Coivd-19 virus too. In order to avoid the noise, maximum intensity projection has been used to reduce noise. Therefore, the second picture is the result and we can see more clear where real spot of Coivd-19 virus is located and it shows result like there was high volume gas or bubble inside the brain.



Original Brain MRI 1: Positive Covid-19



Brain MRI 1(noise reduced): Positive Covid-19

3.1.6.2 Using Python software:

Inferences and Interpretations of Results

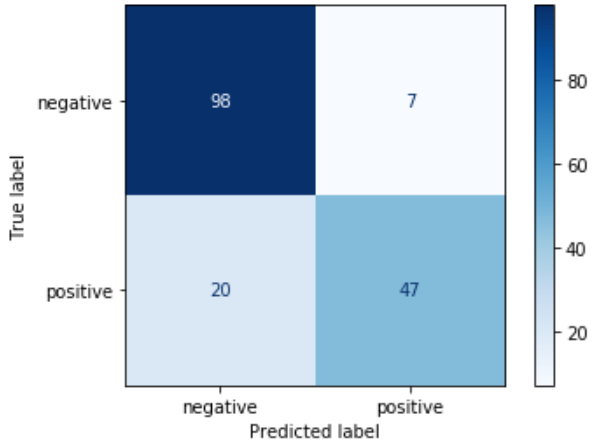
Below is the classification of performance report for all the 4 methods:

3.1.6.2.1 SVM

Accuracy SVM:
0.8430232558139535

	precision	recall	f1-score	support
0	0.93	0.83	0.88	118
1	0.70	0.87	0.78	54
accuracy			0.84	172
macro avg	0.82	0.85	0.83	172
weighted avg	0.86	0.84	0.85	172

Performance 1



Confusion Matrix 1: Standard Vector Machine

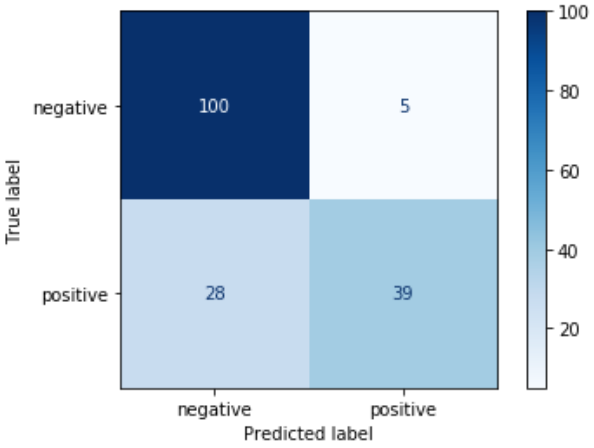
Performance 1 above shows the value that we obtained after implementing performance metrics for precision, recall, and f1-score on our SVM model. For precision of the model are 70% positive and 93% negative. For recall of the model are 87% positive and 83% negative. For f-score of the model are 78% positive and 88% negative.

3.1.6.2.2 K-Nearest Neighbor

Accuracy KNN:
0.8081395348837209

	precision	recall	f1-score	support
0	0.95	0.78	0.86	128
1	0.58	0.89	0.70	44
accuracy			0.81	172
macro avg	0.77	0.83	0.78	172
weighted avg	0.86	0.81	0.82	172

Performance 2



Confusion Matrix 2: K-Nearest Neighbor

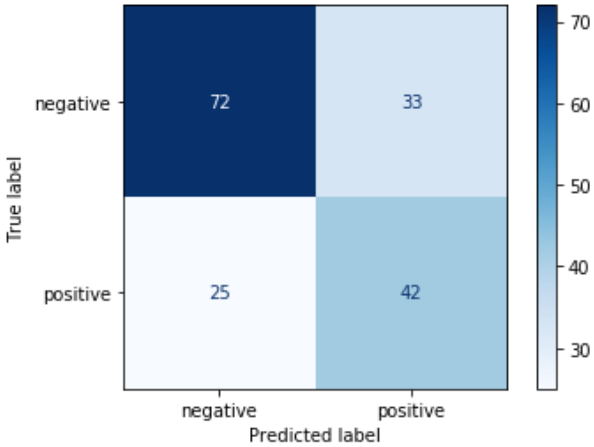
Performance 2 above shows the value that we obtained after implementing performance metrics for precision, recall, and f1-score on our KNN model. For precision of the model is 95% positive and 58% negative. For recall of the model is 89% positive and 78% negative. For f-score of the model is 70% positive and 86% negative.

3.1.6.2.3 Naïve Bayes

Accuracy NAB:
0.6627906976744186

	precision	recall	f1-score	support
0	0.69	0.74	0.71	97
1	0.63	0.56	0.59	75
accuracy			0.66	172
macro avg	0.66	0.65	0.65	172
weighted avg	0.66	0.66	0.66	172

Performance 3



Confusion Matrix – Naïve Bayes

Performance 3 above shows the value that we obtained after implementing performance metrics for precision, recall, and f1-score on our NAB model. For precision of the model is

63% positive and 69% negative. For recall of the model is 56% positive and 74% negative. For f-score of the model is 59% positive and 71% negative.

```
[INFO] evaluating network...
              precision    recall  f1-score   support

    negative       0.89      0.84      0.87        81
    positive       0.79      0.86      0.82        56

   micro avg       0.85      0.85      0.85       137
   macro avg       0.84      0.85      0.84       137
weighted avg       0.85      0.85      0.85       137
```

Performance 4: Evaluation Network

3.2 Evaluation of the Best Once Model According to Fit and Significance.

Based on the result analysis, out of the 4 methods implemented, 3 of them has the accuracy. As for accuracy of SVM model obtained 0.8430232558139535 which is 84%, accuracy of KNN model obtained 0.8081395348837209 which 81%, accuracy of NAB model obtained 0.6627906976744186 which is 66%. From all of the above accuracy results, SVM shows the highest accuracy with 84%. By using confusion matrix also shown that SVM model has the highest value for true positive and predict positive which is 47% instead of Naïve Bayes model's value obtained is 42%, and K-NN model's value obtained is 39%. We can conclude that the best fit and significance model is SVM model.

3.3 Reference

1. Stack overflow, 2020, Classification report in scikit learn, retrieved on 8 July 2020 https://stackoverflow.com/questions/58428533/classification-report-in-scikit-learn
2. Andy Thé, MathWorks, 2020, Introduction to MATLAB with Image Processing Toolbox, online, retrieved on 5 July 2020, at https://www.mathworks.com/videos/introduction-to-matlab-with-image-processing-toolbox-90409.html
3.