# [PREDICTIVE METHODS FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS]

Final Report

Information Technology Capstone Project

COMP5703

Group Members

1. Fullong Chen (450526408)
2. Ryan Kang Jia Yi (430417630)
3. Yeseul Yoon (490241543)
4. Kuo Yuan (480416889)

# ABSTRACT

*Breast cancer is one of the most common cancers for women in the world and especially in Australia. It is vital to have a system that allows the healthcare industry to swiftly and accurately detect such cancer. The goal of this capstone project is to build a predictive model that has the capability to predict if a breast cancer is benign or malignant during the stage of diagnosis, and if the cancer is recurrent or non-recurrent during prognosis stage.*

*In order to do so, the project aims to build  and experiment a predictive model that streamlines different combination of machine learning algorithms, namely, Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Naïve Bayes (NB) as an input to be trained and fitted onto an Artificial Neural Network (ANN). The models will be evaluated and compared to identify the best performing model. To further improve the model's accuracy and overall performance, different techniques will be applied such as GridSearchCV and Upsampling. The dataset used for this project will be obtained from the Wisconsin Breast Cancer dataset.*

*It is hoped that this project could develop a predictive model with high and reliable accuracy to assist healthcare workers in the fight against breast cancer.*

# TABLE OF CONTENTS

# 1. INTRODUCTION

Breast cancer is a cancer that develops from breast tissue (Institue, 2014)[1]. In Australia, the overall five-year survival rate for breast cancer shows about 90% (Council, n.d.)[2]. If the tumour is limited to the breast, 96% of patients will be alive five years after diagnosis; this figure excludes those who die from other diseases (Council, n.d.). Once the diagnosis is made, additional examinations should be conducted to determine which treatments are most likely to be effective (Institue, 2014). Usually, doctors give patients a prognosis, the likely outcome of the disease, based on the type of breast cancer, the test results, the rate of tumour growth, as well as your age, fitness, and medical history (Council, n.d.). The most common types of breast cancer tend to have an excellent long-term prognosis, especially if the cancer is detected early (Council, n.d.).

Classification and data mining methods are an effective way to classify data. Especially in the medical industry, where those methods are widely used in diagnosis and analysis to make decisions (Hiba et al., 2016)[3]. This study underscores the implementation and development of the novel prediction model for breast cancer diagnosis and prognosis. The performances of the models are expected to be analysed to inspect which model outperforms on breast cancer diagnosis and prognosis. Meanwhile, we will discuss skills for the performance improvement of models.

# 2. Related Literature

## 2.1. Introduction

Breast cancer is considered one of the most leading cancers and a major killer among women globally. Although it is very nearly impossible to prevent breast cancer, the early diagnosis of breast cancer during the early stages has been positively connected to a decrease in the mortality and morbidity of the illness[4]. Subsequently, the prognosis is also important for patients who already got the surgery to be cured and cared accordingly.

As our project aims to predict if breast cancer is benign or malignant and also predict if breast cancer is recurrent or non-recurrent, we reviewed diverse literature in terms of diagnosis and prognosis of breast cancer. Moreover, in order to acquire a deeper understanding of breast cancer, we went through not only technical literature about how to build outperforming machine learning algorithms, but also the literature which contains general information of breast cancer such as epidemiology, etiology, and risk factors.

## 2.2. Body

Multiple studies found the magnitude of breast cancer. First of all, according to WHO(World Health Organisation)'s paper, breast cancer is the most commonly diagnosed cancer in women, accounting for about one in four of newly diagnosed cases of cancer[5]. Specifically, WCRF(World Cancer Research Fund International) observed that 1.7 million new cases diagnosed only in 2012[6]. Although its relatively high incidence and even lack of its early symptoms[7], the earlier detection of breast cancer can significantly improve the chance of survival. WCRF found the evidence supporting this theory, the five-year survival rate for patients diagnosed with Stage I/ II breast cancer is 80-90%, however for stages III/IV, the survival rate falls to 24%[8]. The bar chart below depicts the survival rates according to the stages.

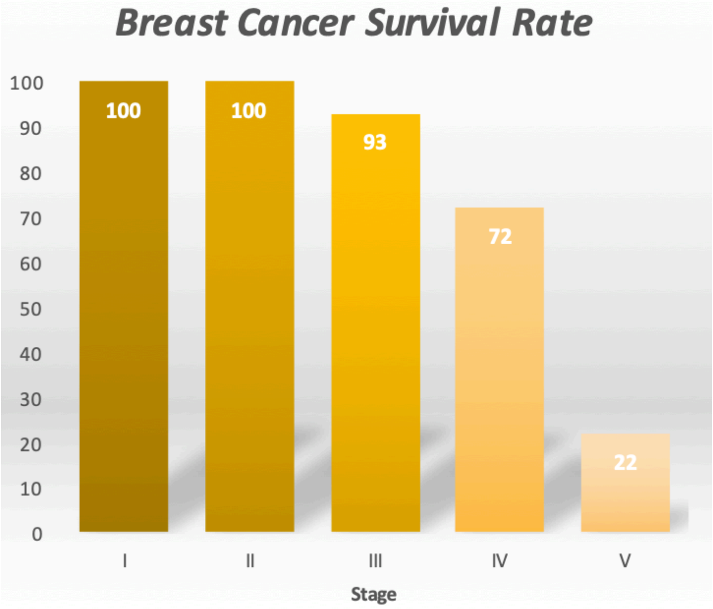


*Figure1 : Breast Cancer Survival Rates*[9]

Hence, it is very clear that accurate classification of benign tumours is a key in order to encourage patients to undergo appropriate treatment and lead the greater outcome. Thus, the diagnosis of breast cancer by correct classification of patients into malignant or benign is the subject of much research (Goel et al., 2018)[10].

Various machine learning algorithms and neural network methods have been implemented on Breast Cancer Wisconsin diagnosis and prognosis dataset. More details can be found below.

- **Decision Tree**

Decision Tree is one of the most widely used classification methods. Decision trees are very simple to understand, interpret and also improve human readability. A decision tree consists of nodes that have exactly one incoming edge, except the root node that has no incoming edges. A node with outgoing edges is an internal node, while the other nodes are called leaves or terminal nodes or decision nodes(Kharya, 2012)[11]. The author applied the best combination of parameters on 10-fold cross-validation, and this model achieved an accuracy of 93.62% with a specificity of 90.66% on diagnosis dataset. The final decision tree model is shown below. On the other hand, the model only performed 76.26 of accuracy on prognosis dataset (AlamKhan et al., 2013)[12].

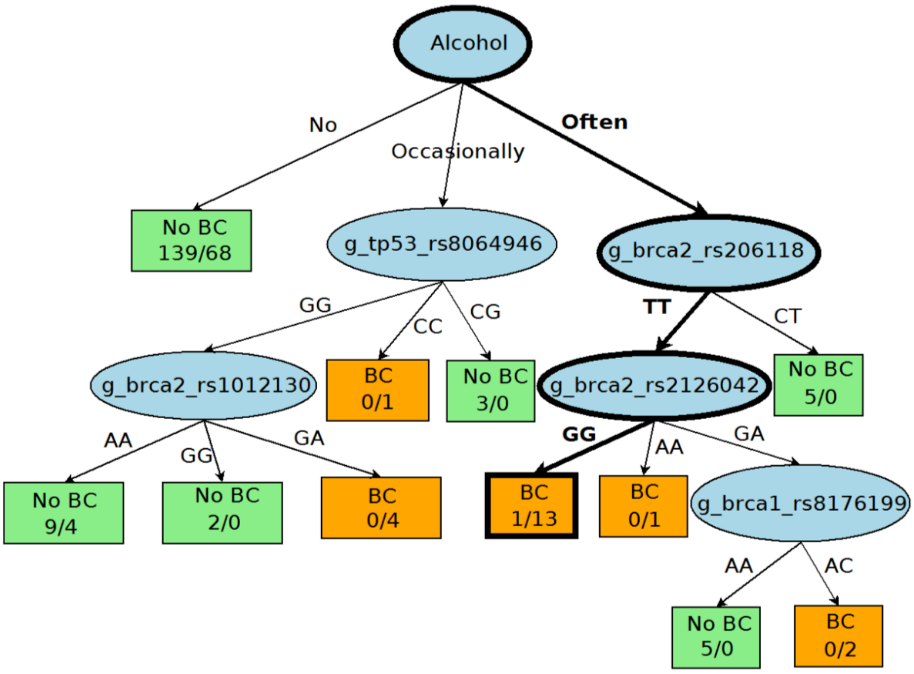


*Figure2 : Decision Tree Model (Kharya, 2012)*

- **Nearest Neighbour**

Nearest Neighbour algorithms classify the data by finding its closest neighbours in a multidimensional feature space populated by known examples from a training data set. The better combinations of dimensions (data features) for nearest neighbours tend to lead a stronger predictive performance (Nisbet et al., 2009)[13]. The results of this algorithm could vary depends on how to measure the distance between the data, hence the two methods, Manhattan distance, and Euclidian distance were reviewed this time. According to previous study on diagnosis dataset[14], Nearest neighbour using Manhattan distance and Euclidian distance resulted in 93.567252% and 94.736844% of accuracy, respectively (Agarap's et al., 2018). AlamKhan's Nearest neighbour model, using k set to 2, performed 76.77% of accuracy on prognosis dataset (AlamKhan et al., 2013).

- **Support Vector Machine**

Support Vector Machines(SVMs) are a set of supervised learning methods used for classification, regression and outliers detection[15]. They are effective especially on high dimensional spaces and efficient, since they use a subset of training points in support vectors. SVM reached 98% of accuracy (Entezari, 2018)[16] on diagnosis data and reached 78.35% of accuracy using polynomial kernel (Maglogiannis et al., 2007)[17].

- **Linear Regression**

Linear Regression finds a linear relationship between the target (a dependent variable) and predictors (independent or explanatory variables). Linear Regression is used in a wide range, for example, forecasting, prediction, and error reduction. While the main goal of our project is prediction, in this study, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables[18]. From previous work, Linear Regression achieved an accuracy of 96.09375% for diagnosis dataset (Agarap's et al., 2018) and 79.29% for prognosis dataset (AlamKhan et al., 2013).

- **Logistic Regression**

Logistic Regression classifier is commonly used when the outcome (target) is binary or categorical. This model is very much like Linear Regression, but a key disparity from Linear Regression is that the target value is a binary (0 or 1) rather than a numeric value. On diagnosis dataset, Logistic Regression showed 95.7% (Khairunnahar et al., 2019)[19] and 75.4% (Sun et al., 2019)[20] on prognosis dataset.

- **Naive Bayes**

A Naive Bayes classifier is a probabilistic machine learning model and frequently used for a classification task. This model works simply although it is powerful for predictive modeling. Naive Bayes reached 84.5% of accuracy on diagnosis data(Kharya et al., 2012) and 70.71% on prognosis data(AlamKhan et al., 2013), which is the lowest among all the model applied on prognosis data.

- **Artificial Neural Networks**

Artificial Neural Networks(ANNs) are biologically inspired computational networks and the most commonly used based on a supervised procedure and comprise three layers: input, hidden, and output[21]. Artificial neural network showed the accuracy of 92.9%(Yana & Stoyan, 2019)[22] in a previous study and 90.22% (Maglogiannis et al., 2007), on diagnosis and prognosis dataset, respectively.

- **Deep Neural Networks**

Deep Neural Networks(DNNs) are distinguished from the single hidden layer neural network in terms of their depth. In deep learning networks, each layer trains on a distinct set of features based on the previous layer's output. The further toward into the network, the more complex the features the nodes can recognise and learn. Deep neural network reached 95% of accuracy(Zhang, Y. (2019)[23] on diagnosis data and on the other hand, the model performed 74.9%(Sun et al., 2019) of accuracy on prognosis data.

- **Model Limitations**

It is compelling to take account of each model's limitation in order to fully assimilate them and suggest appropriate models. The table below summarises the limitation of each model.

| Model | Limitation |
|---|---|
| **Decision Tree** | • Often longer time to train the model involved<br>• Relatively expensive as the complexity<br>• Less adequate for dealing with continuous values |
| **Nearest Neighbour** | • The number of neighbours needs to be determined<br>• Higher computation time required<br>• Result fluctuation depends on the type of distance measurement |
| **Support Vector Machine** | • Less suitable for the large dataset<br>• Underperformed when the number of features exceeds the number of the training data sample<br>• Not easy to find out a good kernel function |
| **Linear Regression** | • Sensitive to outliers of given data<br>• Limitation to linear relationships |
| **Naive Bayes** | • Sensitive to skewed data<br>• Calculation of the prior possibility needed<br>• Underperformed when the attributes are related |
| **Artificial Neural Networks** | • Longer training time required<br>• Large amount of data needed<br>• Difficult to visualise and present the model |

*Table 1 : Literature review - the limitation of six models*

### 2.3. Conclusion

| Model | Benchmark - Diagnosis | Benchmark - Prognosis |
|---|---|---|
| **Random Forest** | 93% | 76.26% |
| **Deep Neural Network** | 95% | 74.9% |
| **Artificial Neural Networks** | 92.9% | 90.22% |
| **Decision Tree** | 93.62% | 76.26% |
| **Nearest Neighbour** | 93.567252% (Manhattan distance)<br>94.736844% (Euclidian distance) | 76.77% |
| **Support Vector Machine** | 98% | 78.35% (Polynomial)<br>90.21% (Gaussian RBF) |
| **Linear Regression** | 96.09% | 79.29% |
| **Naive Bayes** | 84.5% | 70.71% |
| **Logistic Regression** | 95.7% | 75.4% |

*Table 2 : Benchmark accuracy for breast cancer diagnosis and prognosis dataset*

This literature review aims to examine the benefits of early diagnosis and accurate prognosis of breast cancer in terms of higher survival rate which was proven on multiple studies and to understand various machine learning algorithms and neural network methods in order to carry out the outcome with higher accuracy. Some of the methods reviewed above were used in our own model implementation along with our distinct approach in order to come up with a final novel model.

# 3. RESEARCH/PROJECT PROBLEMS

## 3.1. Research/Project Aims & Objectives

To detect breast cancer earlier is the most effective way to cut down breast cancer deaths (Kharya, 2012). Early diagnosis needs an accurate and reliable diagnosis procedure that can be used by physicians to distinguish benign breast tumours from malignant ones (Kharya, 2012).

Since large volumes of medical data are being collected and made available to the medical research groups, diverse machine learning models have been developed for both diagnosis and prognosis dataset. Comparative analysis between previous results and our experiments will be also delivered.

This project focuses on the use of machine learning and deep learning models for both breast cancer diagnosis and prognosis classification problems and come up with a novel prediction model. Although we assume that predicting the outcome of a disease could be a challenging task where to develop data science applications, we also anticipate our model to make a valuable contribution in the medical field.

## 3.2. Research/Project Questions

What exist algorithm methods could be used or combined in order to build the outperforming novel prediction model so the model is able to assist patients against breast cancer and escalate the survival rate?

## 3.3. Research/Project Scope

This study gives the novel prediction model for the breast cancer diagnosis and prognosis handling a binary classification problem. Candidate models involve more than one foundation algorithms from machine learning and deep learning. By model result comparison, we may select an optimal model to address breast cancer diagnosis and prognosis problems.

# 4. METHODOLOGIES

## 4.1. Methods

To solve the problem for our project, following machine learning procedures, training literature review models and performing a novel machine learning method is crucial. Since this is a machine learning project, the machine learning process used in many industries will be followed.  From regular feedback and insight from the client and tutor, we are made sure to stay on track. The diagram below shows a summary of our machine learning method.

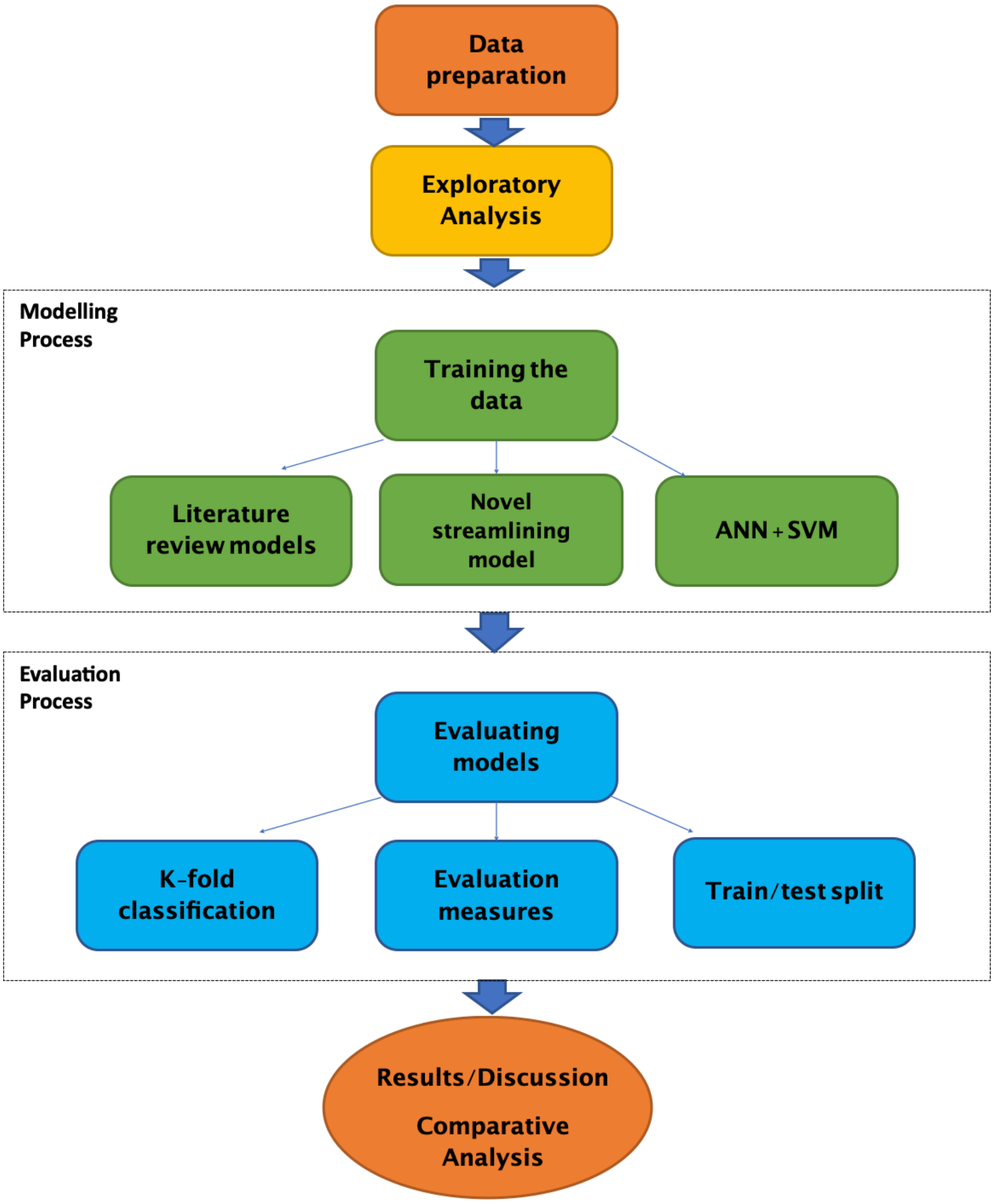


*Figure 3 : Our project method and machine learning process*

The steps of our machine learning problem of breast cancer can be summarised as followed:

1. Choosing datasets and identifying the project problem

2. Diagnosis and Prognosis breast cancer datasets were chosen then loaded and prepared. Cleaning up the datasets, recoding variables, scaling variables and handling missing variables were done.

3. Following that data exploratory analysis was done to determine if there are any outliers or anything of interest that may impact our modelling process. Plots of distribution and other plots were done that can help our prediction of diagnosis and prognosis of breast cancer.

4. Next was the training process. Firstly the dataset was trained and modelled following the literature review models where possible. The details of the literature review models trained are outlined in the data analysis section.

5. The literature review models were trained closely to the literatures methods where possible but were otherwise done with a basic default machine learning process. Training/test split of 70% training and 30% test were used because most literature reviews did not mention their split ratios at all. Testing error were mainly recorded to compare with the literature review results and to check if the literature reviews were valid.

6. Next evaluation measures were applied to the literature review methods such as confusion matrix and k-fold classification. From the k-fold classification it was discovered that we can apply a new method of combining models.

7. Therefore a streamlined machine learning method was developed. The diagram in the data analysis section of our report outlines our streamlined method. In our streamlined method the datasets was split into 70% training and 30% test and then different combinations were streamlined, optimised and results were output.

8. An ANN+SVM model was also trained to compare with our streamlined machine learning method.

9. Finally different evaluation measures were calculated and optimisation of parameters of the streamlined machine learning method were performed. Results were compiled and comparative analysis was done.

## 4.2. Data Collection

The datasets used in this project are the Breast cancer Wisconsin (Diagnosis) and the Breast cancer Wisconsin (Prognosis), retrieved from the links below.

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(prognostic)

The Breast cancer Wisconsin (Diagnosis) has 569 of instances and 32 of attributes (including ID and a target variable), the Breast cancer Wisconsin (Prognosis) has 198 of instances and 34 attributes (including ID and a target variable). Moreover, he prognosis dataset has 4 missing attribute values and they were simply deleted. The variables and their description can be seen in the table below.

| Variable/Features | Meaning |
|---|---|
| **Id** | The Id number of the patients |
| **Outcome** | Whether the disease has recurred (R) or non-recurred (N) |
| **Tumor size** | Diameter of the removed tumour in centimetres |
| **Lymph node status**<br>*(non-existent for diagnosis dataset)* | Number of positive axillary lymph nodes |

| The 10 Real-valued features (means, standard error, worst) | |
|---|---|
| **Variable/Features** | **Meaning** |
| **Radius** | Distance from centre to points on the perimeter |
| **Texture** | Standard deviation of the gray-scale values |
| **Perimeter** | Perimeter of the cancer area |
| **Area** | Area of the cancer area |
| **Smoothness** | Local variation in radius lengths,<br>how round the cancer area is |
| **Compactness** | (Perimeter squared divided by area)-1 |
| **Concavity** | Severity of concave portions of the contour |
| **Concave points** | Number of concave portions of the contour |
| **Symmetry** | Value of symmetry |
| **Fractal dimension** | Coastline approximation – 1,<br>measure of how complicated the cancer figure is |

*Table 3, 4 : Data descriptions*

For each real-valued feature, the mean, standard error and worst/largest of these features was computed for each image.

## 4.3. Data Analysis

### 4.3.1 Pre-processing and Exploratory Data Analysis

Pre-processing and Exploratory Data Analysis(EDA) are top priorities for every data science projects. They helped us to deeply understand the datasets and decide how to manipulate the variables that always impact on the outcome.

- Data standardisation

For making sure that data is internally consistent; that is, each data type has the same content and format[24]. Data standardisation was implemented followed by data analysis and visualisation.

Rather than pasting all the graphs that the group managed to draw, a few graphs that are considered as essential or more informative will be shown as follows(Figure 5 to Figure 8).

- Violin Plot

Drawing the distribution of the data across several levels of categorical variables and can be an effective way to compare one or more features' distributions. Violin plots are similar to box plots, but they also show the probability density of the data at different values, usually smoothed by a kernel density estimator[25]. Moreover, the indubitable advantage of the violin plot over the box plot is that the violin plot is showing the shows the entire distribution of the data.

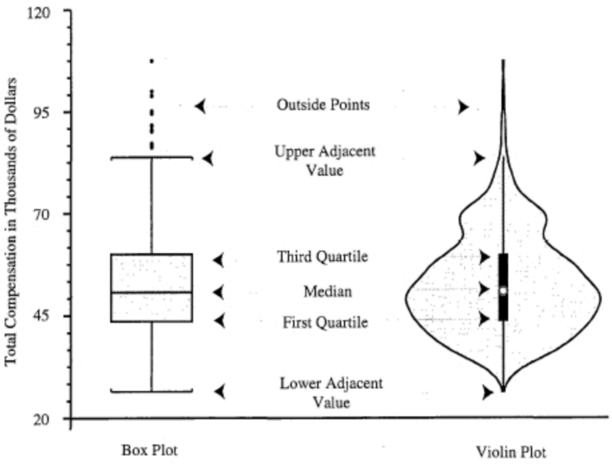


*Figure 4 : Component comparison between Box plot and Violin Plot (Hintze & Nelson, 1998)[26]*

Looking at the violin plot of Breast Cancer Wisconsin data - Diagnosis (Figure 5) below, which is the first attachment showing EDA progress, texture_mean feature shows parted median value of malignant class(M, coloured in green) and benign class(B, coloured in orange). What it is meaning is that this feature may be helpful for classify the data into two different classes. This finding is also applicable for concave points_worst and texture_worst in terms of their distribution. However, smoothness_se, for instance, its median value of both classes does not look like parted so it does not give good information for prediction. In contrast to diagnosis dataset, in Breast Cancer Wisconsin data - Prognosis, not many features show separated data distribution, as the second violin plot (Figure 7) illustrates.

- Heat Map

To comprehend the correlation between the variables. The heat map is a data visualisation technique that shows the magnitude of a phenomenon as colour in two dimensions. The heat map provides an excellent visualisation skill when comparing multiple variables and the relationships between them. Here, the variation in colour may be by hue or intensity27.

According to the heat map of diagnosis dataset (Figure 6) below, concave points_mean feature has very strong correlation with perimeter_mean, concavity _mean/worst and concavity points_worst, representing 0.9. However, the correlations between symmetry_se and other variables are relatively low. The majority of the values on the heat map shows 0.1 and not even one variable has over 0.5 correlation with symmetry_se.

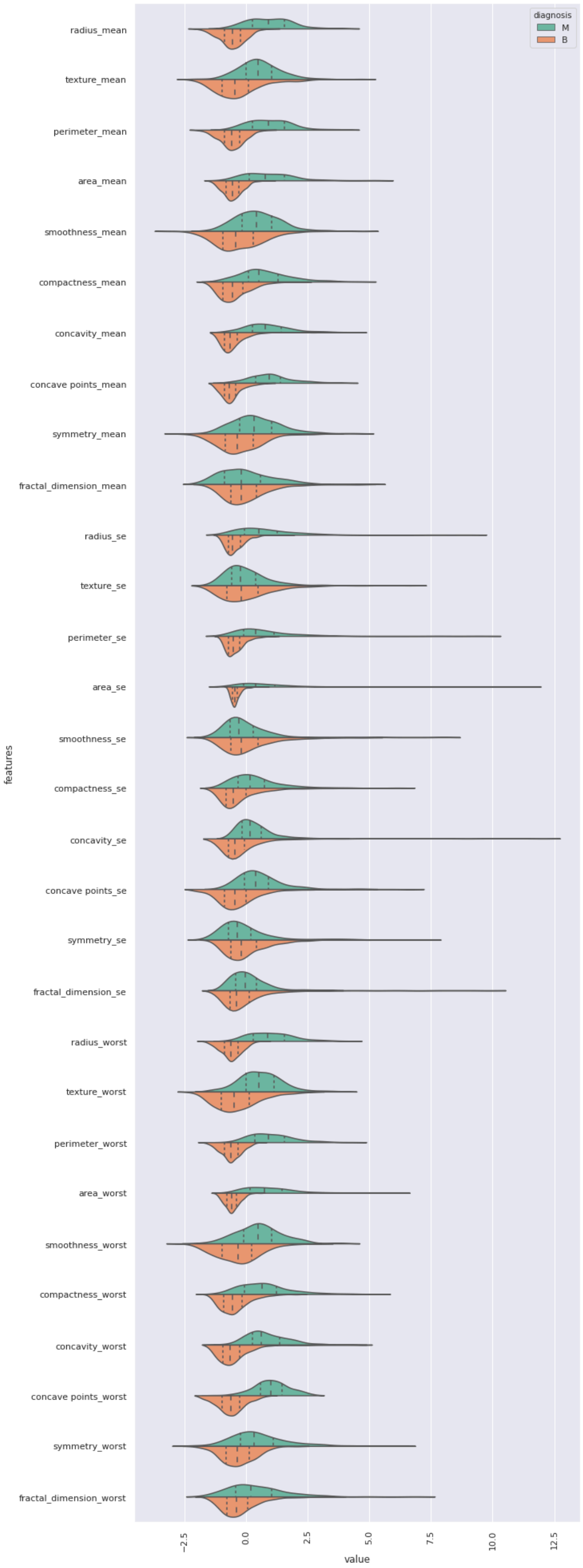


*Figure 5 : Diagnosis dataset - violin plot*

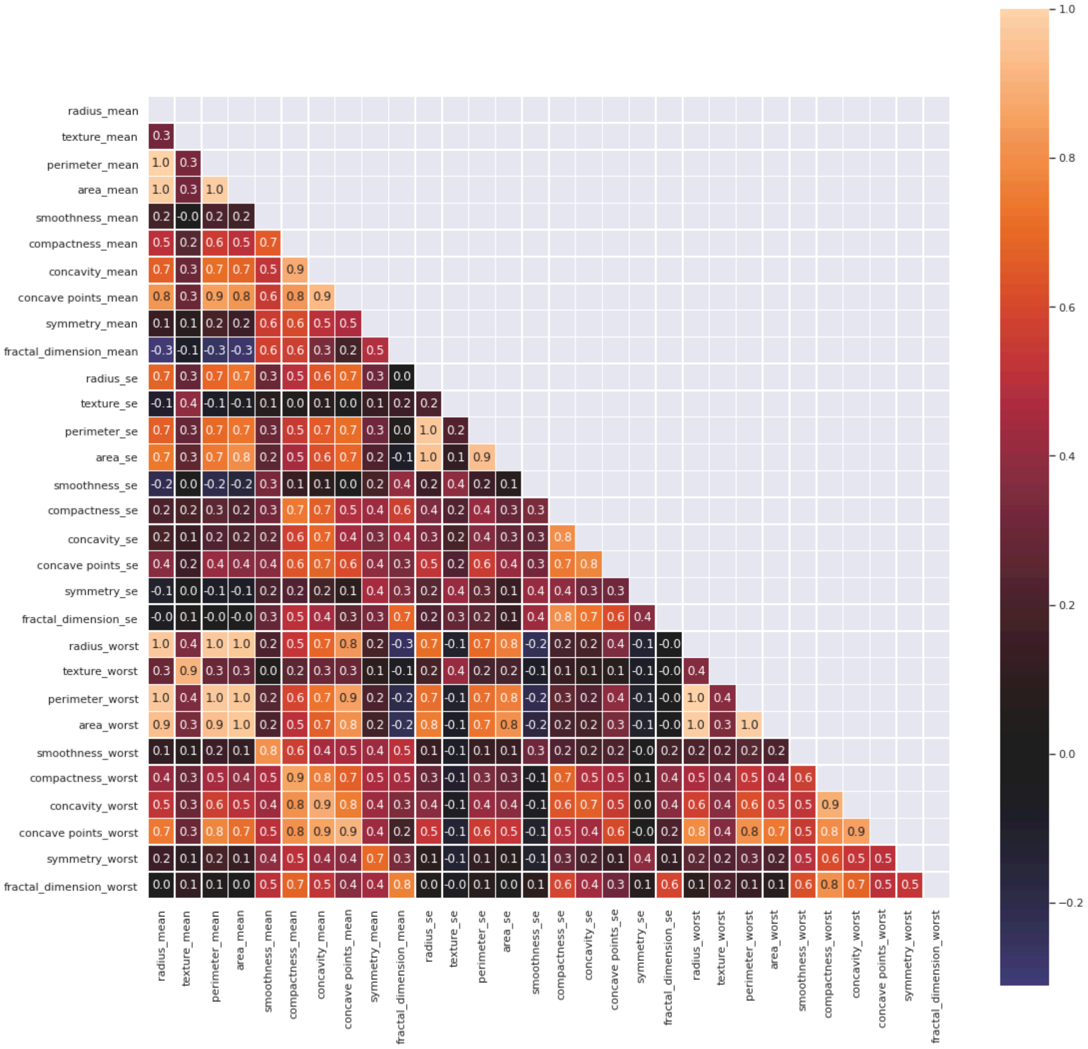


*Figure 6 : Diagnosis dataset - heat map*

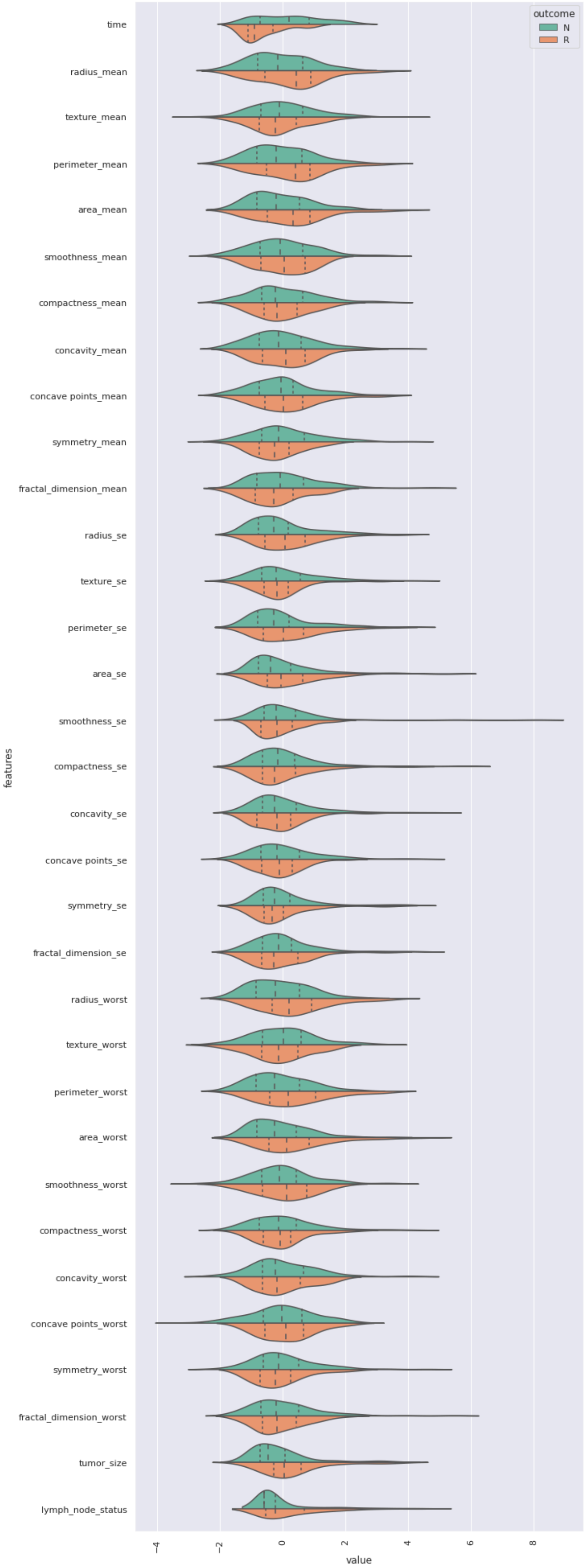


*Figure 7 : Prognosis dataset - violin plot*

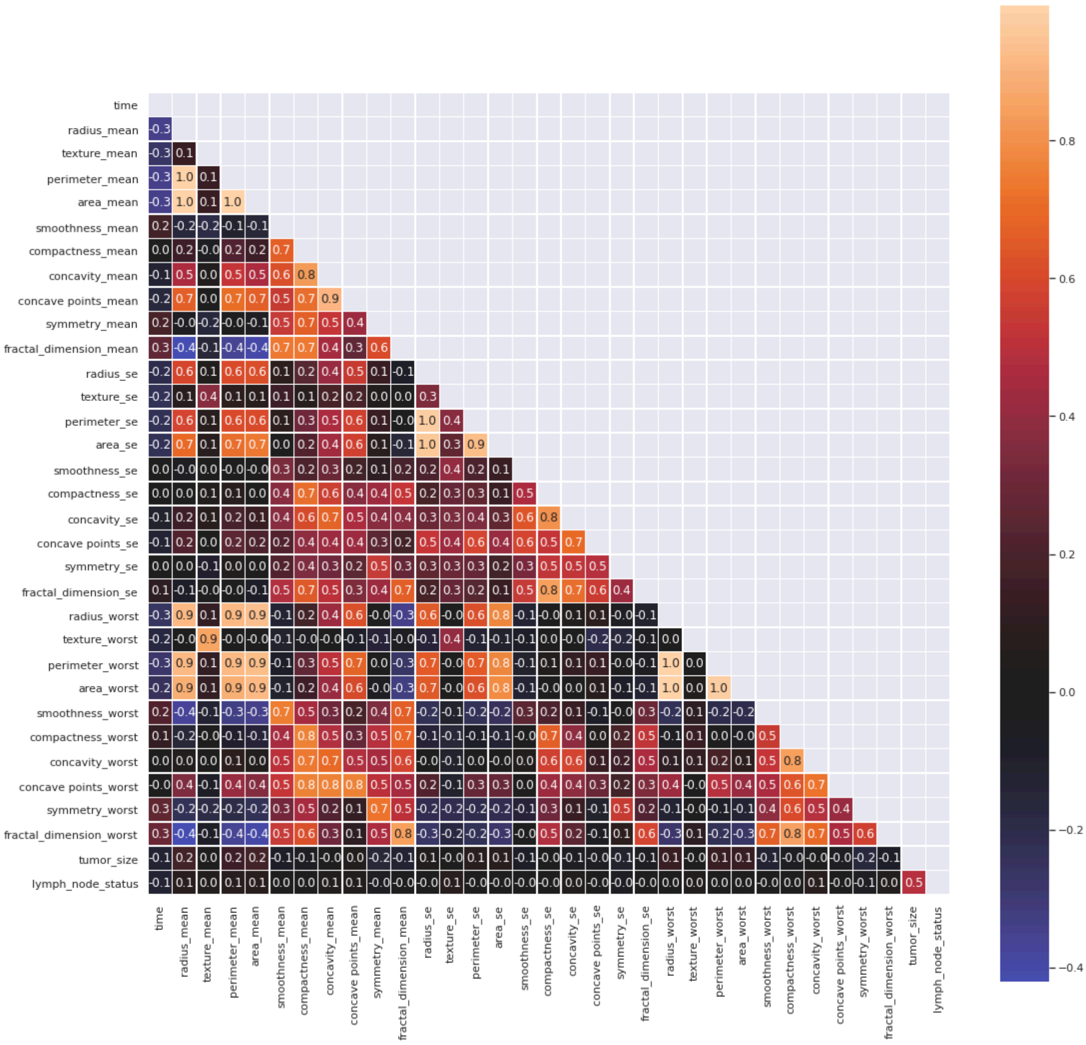


*Figure 8 : Prognosis dataset - heat map*

### 4.3.2 Imbalanced data

Many of real-world classification problems has imbalanced sample distribution, such as spam/fraud detection and medical data. One of the dataset in our project, that is Breast Cancer Wisconsin Prognosis dataset, is highly skewed - it has 151 non-recur outcomes and 47 recur outcome. The figure below shows how the outcome is distributed in Prognosis data.

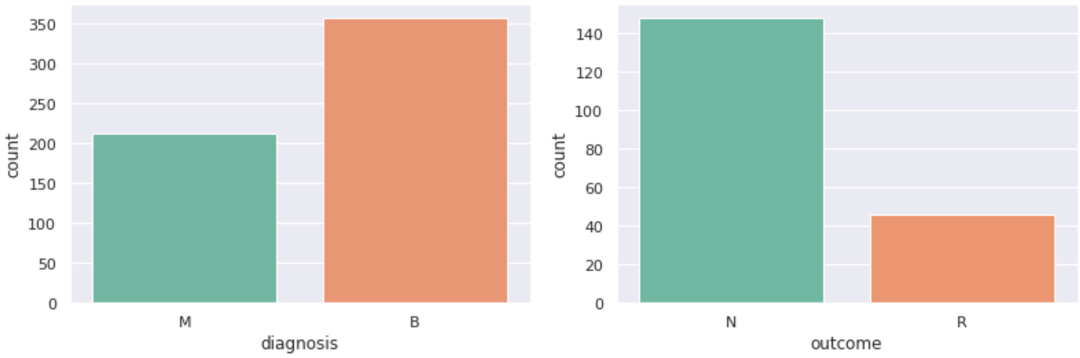


*Figure 9, 10 : Diagnosis and prognosis dataset class distribution*

When the distribution of given data is skewed or biased, an imbalanced classification problem could occur. To curb this problem, two additional techniques - algorithm approach and data approach - have been implemented on both Breast Cancer Wisconsin Diagnosis and Prognosis dataset. First skill is to modify machine learning algorithm to boost the performance on minority class by cost-sensitive learning or to use the costs as a penalty for misclassification while the model is being trained. To do this, the cost for misclassification is added to the error or used to weight the error. The scikit-learn library in Python kindly provides cost-sensitive extensions via the 'class_weight' argument. Hence, this function is applied on Support Vector Classifier (Support Vector Machine), Decision Tree Classifier, and Logistic Regression Classifier in this project. Secondly, the dataset can be re-sampled and this approach is considered more flexible. For re-sampling, up-sampling and under-sampling are most commonly used. For our dataset, up-sampling is implemented to increase the number of minority class.

The table below shows the class distribution of two dataset and the techniques applied.

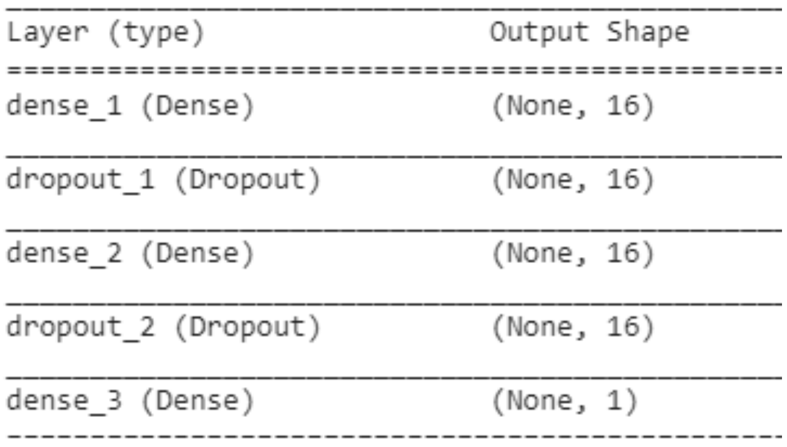
| Dataset | Class Distribution | Applied technique |
|---|---|---|
| **Diagnosis** | {212 malignant,<br>357 benign} | Up-sampling |
| **Prognosis** | {151 non-recur,<br>47 recur} | Up-sampling<br>Cost-sensitive extensions |

*Table 5 : Class distribution and applied techniques*

## 4.4. Basic Machine Learning Models

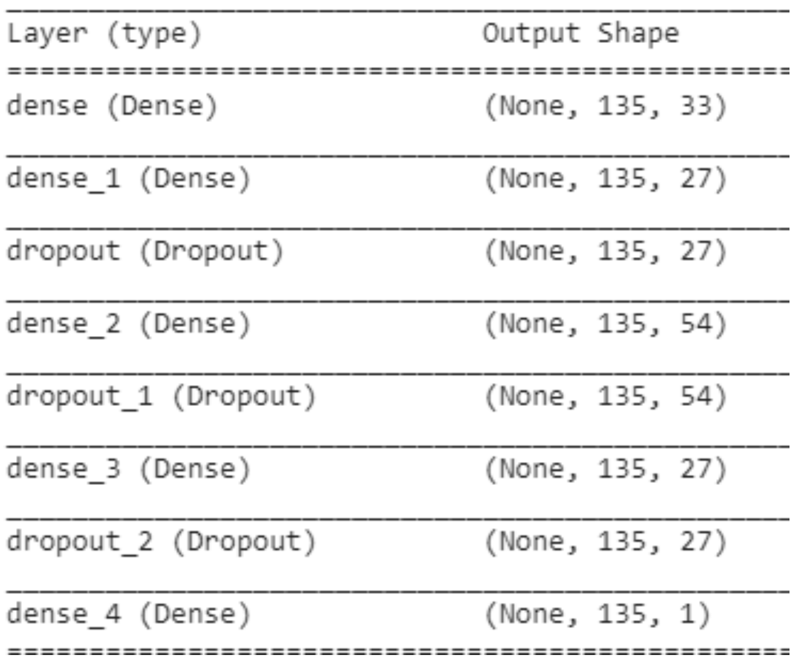
Each model was implemented on the diagnosis and prognosis breast cancer dataset to check the validity of the literature review models. Some literature reviews had no description on parameters or their method, so the default parameters were used just to check their results.

The general structure of the Artificial Neural Network and Deep Neural Network can be seen in the following figures below.



```
_________________________________________________________________
Layer (type)                 Output Shape
=================================================================
dense_1 (Dense)              (None, 16)
_________________________________________________________________
dropout_1 (Dropout)          (None, 16)
_________________________________________________________________
dense_2 (Dense)              (None, 16)
_________________________________________________________________
dropout_2 (Dropout)          (None, 16)
_________________________________________________________________
dense_3 (Dense)              (None, 1)
-----------------------------------------------------------------
```

*Figure 11: General ANN Structure*



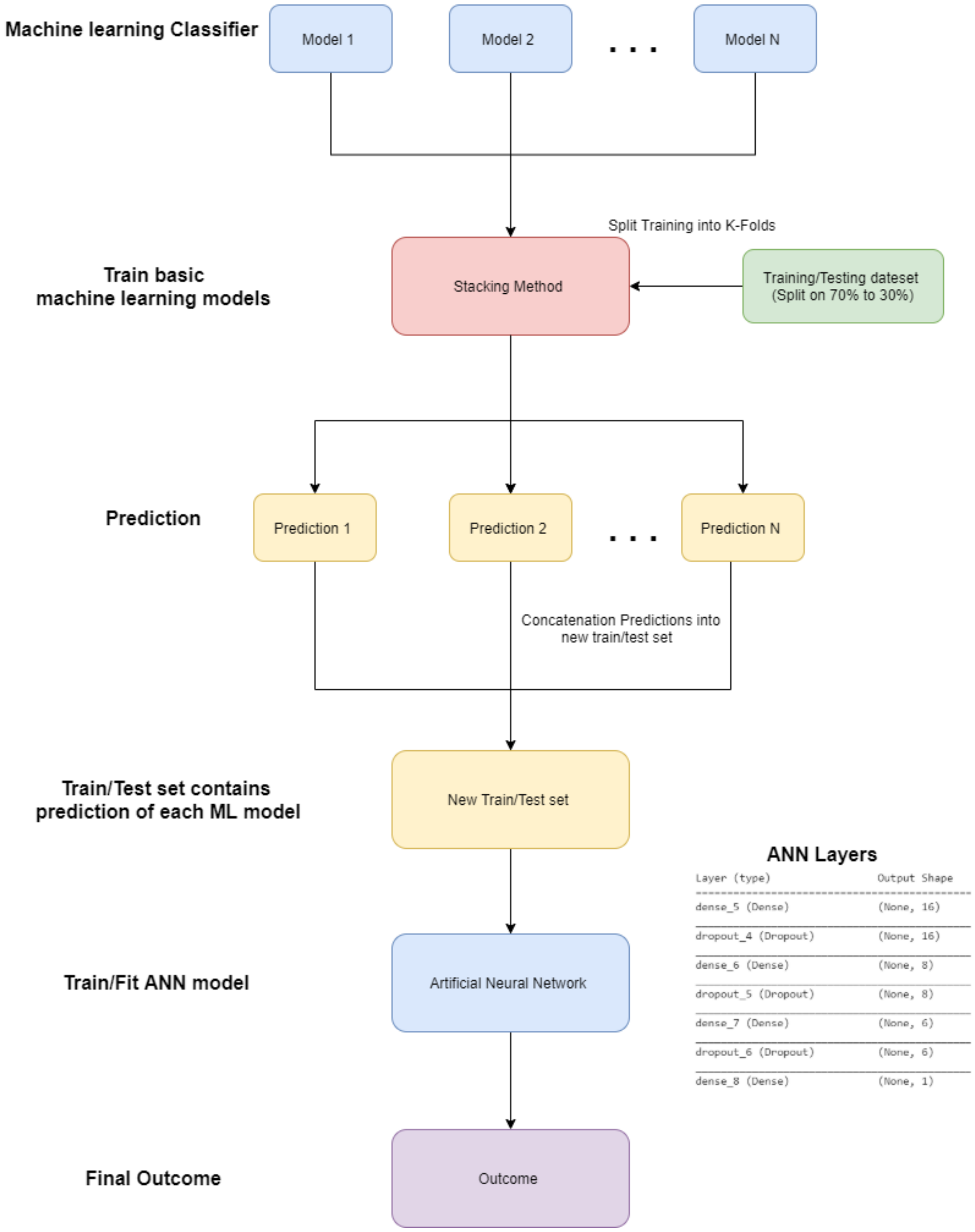
```
_________________________________________________________________
Layer (type)                 Output Shape
=================================================================
dense (Dense)                (None, 135, 33)
_________________________________________________________________
dense_1 (Dense)              (None, 135, 27)
_________________________________________________________________
dropout (Dropout)            (None, 135, 27)
_________________________________________________________________
dense_2 (Dense)              (None, 135, 54)
_________________________________________________________________
dropout_1 (Dropout)          (None, 135, 54)
_________________________________________________________________
dense_3 (Dense)              (None, 135, 27)
_________________________________________________________________
dropout_2 (Dropout)          (None, 135, 27)
_________________________________________________________________
dense_4 (Dense)              (None, 135, 1)
=================================================================
```

*Figure 12: General DNN Structure*

The ANN structure has a total of 3 dense layers and 2 dropout layers. Whereas for the DNN, it has structure of 5 dense layers and 3 dropout layers.

## 4.5. Proposed Novel Streamlined Model

The streamlined method can be summarised in the diagram below. Different combinations were trained.


Machine learning Classifier
Model 1
Model 2
. . .
Model N
Split Training into K-Folds
Train basic machine learning models
Stacking Method
Training/Testing dateset (Split on 70% to 30%)
Prediction
Prediction 1
Prediction 2
. . .
Prediction N
Concatenation Predictions into new train/test set
Train/Test set contains prediction of each ML model
New Train/Test set
ANN Layers
Train/Fit ANN model
Artificial Neural Network
Final Outcome
Outcome

```
Layer (type)                 Output Shape
=========================================
dense_5 (Dense)              (None, 16)
_________________________________________
dropout_4 (Dropout)          (None, 16)
_________________________________________
dense_6 (Dense)              (None, 8)
_________________________________________
dropout_5 (Dropout)          (None, 8)
_________________________________________
dense_7 (Dense)              (None, 6)
_________________________________________
dropout_6 (Dropout)          (None, 6)
_________________________________________
dense_8 (Dense)              (None, 1)
```

*Figure 13 : Streamlined model diagram*

The steps of the streamlined method can be summarised as below:

1. Machine learning classifier/s on training dataset
2. K-fold applied to these classifiers retrieving most common outcome
3. Outcome from machine learning classifier/s concatenated
4. New training dataset streamlined as a result
5. New dataset input into default ANN
6. Outputs final outcome and evaluated
7. A pseudo code is also shown below of the streamlined method.

**Streamline algorithm**

```
Initialise training dataset
Initialise test dataset
Initialise models list = [Model 1....Model n]
Initialise empty stacking training dataset
Initialise empty stacking test dataset

for i, model in enumerate(models list):
        for train fold, test fold in k-fold(training dataset):
                stacking training dataset[train fold, i] =
                model(train fold).predict(train fold)
                stacking test dataset[test fold, i] =
                mean(model(train fold).predict(test fold))
        Concatenate stacking training dataset[:,i] (i.e. each model is a column)
        Concatenate stacking test dataset[:,i] (i.e. each model is a column)

Run concatenated stacking training dataset and initial training set outcomes
through ANN
Evaluate and test result using stacking test results
```

An ANN+SVM as a final layer model was also proposed and trained. This was used for comparison and is a model that has not been used on the diagnosis or prognosis breast cancer dataset before.

# 5. RESOURCES

The project resource plays a key part to determine the success of the project. Each resource is carefully allocated to ensure proper and efficient usage of it. The sections below outline the different types of resources required to successfully complete this project.

## 5.1. Hardware

Google's cloud computer specification (To execute codes on Google Colab)

- CPU : Xeon Processors @2.3Ghz
- GPU : Tesla K80
- RAM : 12GB GDDR5 VRAM
- Disk : 33GB

## 5.2. Software

The goal of our project is to build multiple predictive models that will be used on medical-related datasets. To ensure project success, we have employed the following software tools:

**Python Programming language**: Python is one of the most common programming languages, and it offers many libraries that can handle data science tasks such as import datasets, data exploration, data pre-processing, and most importantly, building of models. It is also the most familiar and experienced language among the group members. The following are the Python libraries that are needed for this project.

- Pandas
- NumPy
- Scikit-Learn
- SciPy
- Keras
- TensorFlow
- Matplotlib
- Seaborn

**Google Drive**: Used to store and share project-related information and documents among the group.

**Google Colab**: Allow the group to perform coding tasks collectively and utilise Google's cloud computer which has more processing capability. Used to conduct all coding and visualisation related tasks such as EDA, Pre-processing ,implementation of models and analysis of model's results.

**Microsoft Words/Pages**: Used for official documentation of project progress, project proposal, and project report.

**Microsoft Excel**: Used for early stage data exploration and analysis.

**Slack**: The project official communication channel, and it is used to facilitate communication among the group members and the project supervisor.

**Facebook Messenger**: The project unofficial communication channel used to facilitate communication among the group members.

**Zoom**: An online platform used to hold weekly meetings with the project supervisor and project client to update project progress and obtain feedback.

## 5.3. Materials

To support the project, online materials are used to develop and understand key concepts of machine learning algorithms and medical-related datasets. These materials are also used to set the benchmarks results obtained from literature reviews.

Articles are used to have a better understanding of methods, performance evaluation, and previous case studies done by researchers. The following are some of the sources used to search for articles:

- Paper with Code: https://paperswithcode.com/
- Towards Data Science: https://towardsdatascience.com/
- ScienceDirect: https://www.sciencedirect.com/
- ResearchGate: https://www.researchgate.net/
- Google Scholar: https://scholar.google.com/

The following materials were used to explore and research on different types of medical-related dataset that could be used for the project:

- Kaggle: https://www.kaggle.com/
- UCI: https://archive.ics.uci.edu/ml/index.php

### 5.4. Roles & Responsibilities

Throughout the duration of the entire project, each member is assigned a primary and a secondary role as a programmer. The responsibility of a programmer is to ensure good quality and efficient codes are written and rectify any bugs and issues that may occur. The table below outlines the detailed responsibility of each member's respective primary role.

| Group Member | Primary Role | Secondary Role | Responsibility |
|---|---|---|---|
| **Ryan Kang** | Project Manager | Programmar | • Facilitate communication between the group, client, supervisor and among the group members.<br>• Ensure group is on schedule with milestone and deliverables |
| **Fullong Chen** | Data Exploratory Expert | Programmar | • Responsible for providing detailed and informative information from datasets.<br>• Ensure the dataset is suitable and compatible with proposed models. |
| **Yeseul Yoon** | Visualisation Expert | Programmar | • Responsible for translating raw data and predictive results into informative visualisation. |
| **Kuo Yuan** | Performance Analyst | Programmar | • Responsible for analysing and presenting performance of models built and compared against benchmarks models. |

*Table 6 : Roles & Responsibilities*

# 6. Milestones / Schedule

To help monitor the group's progress and status, a project plan and Gantt chart were created. The project can be broken down into 4 phases consisting of research phase, data preparation phase, modelling phase and results phase.

## 6.1. Project Plan

| Weeks | Tasks | Reporting | Date |
|---|---|---|---|
| **Week 1** | - | - | 24-02-2020 |
| **Week 2** | • Establish communication channel<br>• Brainstorm project ideas | None | 04-03-2020 |
| **Week 3** | • Research datasets, problem/ idea and potential predictive models | None | 09-03-2020 |
| **Week 4** | • Literature reviews on previous case studies of methods deployed and model performance | Client meeting to review datasets, idea and proposed models | 16-03-2020 |
| **Week 5** | Proposal Report Due | - | 23-03-2020 |
| **Week 6** | • Exploratory data analysis (with visualisation) | None | 30-03-2020 |
| **Week 7** | • Pre-processing datasets (Handle missing value, Normalisation, changing outcome to binary value)<br>• Basic machine learning algorithm implementation (SVM, ANN & DNN) | None | 06-04-2020 |

| Weeks | Tasks | Reporting | Date |
|---|---|---|---|
| **Week 8** | • Machine learning models optimisation (GridSearchCV, hyper-parameter tuning, feature selection)<br>• Cross check literature review's model performance and results | None | 13-04-2020 |
| **Week 9** | Progress Report Due<br>• Research on Novel Model<br>• Implementation of more basic machine learning models (RF, LR,DT & NB) | - | 27-04-2020 |
| **Week 10** | • Implementation of proposed novel model (Streamline different ML model + ANN model)<br>• Analysis and visualisation of model's result | Client meeting review project progress and obtain feedback on activities and deliverables | 04-05-2020 |
| **Week 11** | • Implementation of ANN+SVM model<br>• Implementation of different combination of ML model +ANN<br>• Proposed novel model performance optimisation<br>• Analysis and visualisation of model's result | None | 11-05-2020 |
| **Week 12** | • Documentation | - | 18-05-2020 |
| **Week 13** | Final Report/Presentation | Client meeting to review final product of project | 25-05-2020 |

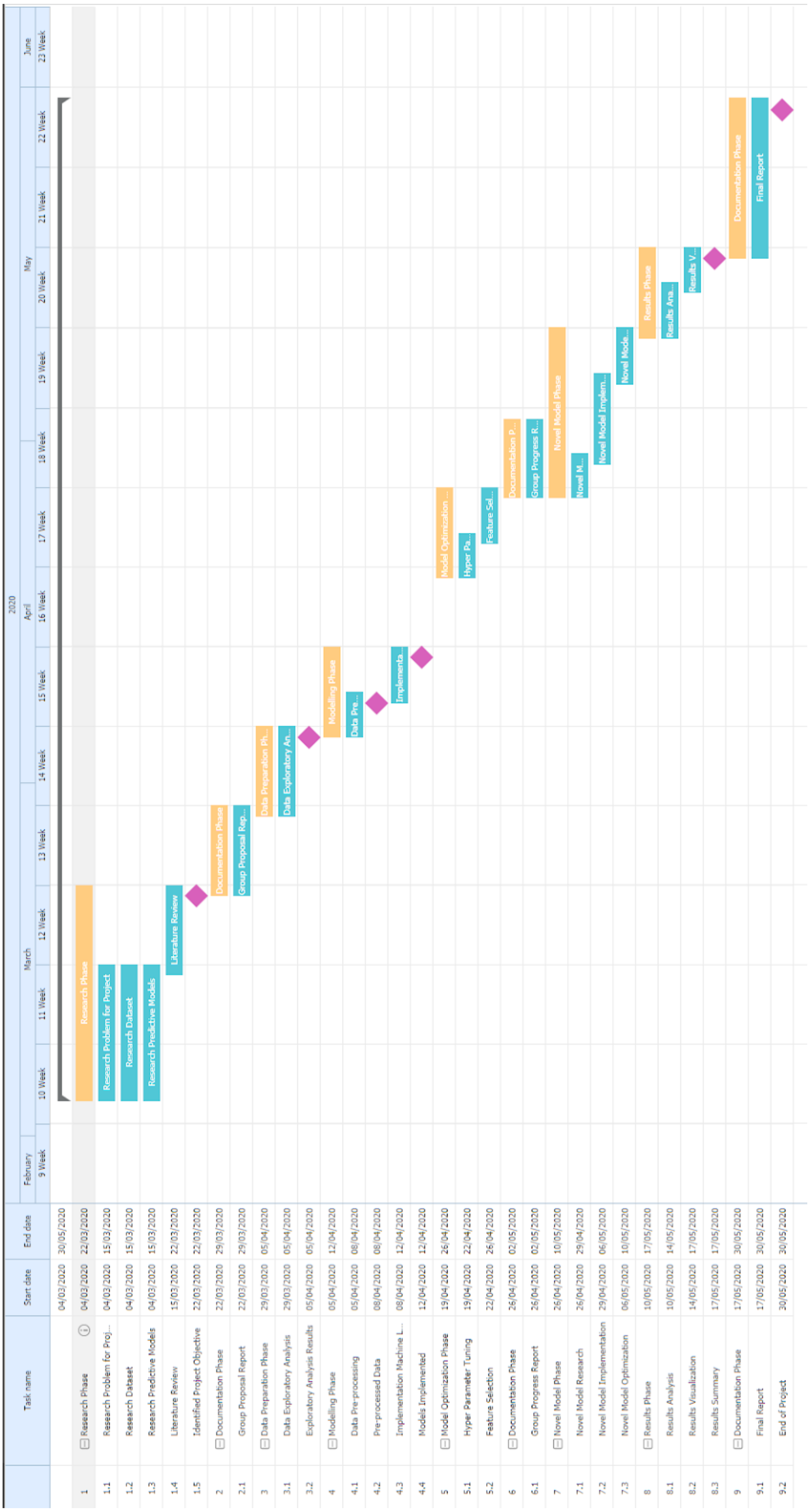
*Table 7 : Project Plan*

## 6.2. Milestone Summary

| No. | Milestones | Dates |
|---|---|---|
| **1** | Identified Project Objective | 22/03/2020 |
| **2** | Proposal Report | 29/03/2020 |
| **3** | ExploratoryAnalysis Results | 05/04/20 |
| **4** | Pre-processed data | 08/04/2020 |
| **5** | Basic ML models implemented | 12/04/2020 |
| **6** | Optimized ML models | 26/04/2020 |
| **7** | Progress Report | 03/05/2020 |
| **8** | Implemented Novel Model | 13/05/2020 |
| **9** | Optimized Novel Model | 17/05/2020 |
| **10** | Result Summary | 17/05/2020 |
| **11** | Final Report | 30/05/2020 |

*Table 8 : Milestone summary*

## 6.3. Project Gantt chart



| | Task name | Start date | End date |
|---|---|---|---|
| | | 04/03/2020 | 30/05/2020 |
| 1 | ⊟ Research Phase | 04/03/2020 | 22/03/2020 |
| 1.1 | Research Problem for Proj... | 04/03/2020 | 15/03/2020 |
| 1.2 | Research Dataset | 04/03/2020 | 15/03/2020 |
| 1.3 | Research Predictive Models | 04/03/2020 | 15/03/2020 |
| 1.4 | Literature Review | 15/03/2020 | 22/03/2020 |
| 1.5 | Identified Project Objective | 22/03/2020 | 22/03/2020 |
| 2 | ⊟ Documentation Phase | 22/03/2020 | 29/03/2020 |
| 2.1 | Group Proposal Report | 22/03/2020 | 29/03/2020 |
| 3 | ⊟ Data Preparation Phase | 29/03/2020 | 05/04/2020 |
| 3.1 | Data Exploratory Analysis | 29/03/2020 | 05/04/2020 |
| 3.2 | Exploratory Analysis Results | 05/04/2020 | 05/04/2020 |
| 4 | ⊟ Modelling Phase | 05/04/2020 | 12/04/2020 |
| 4.1 | Data Pre-processing | 05/04/2020 | 08/04/2020 |
| 4.2 | Pre-processed Data | 08/04/2020 | 08/04/2020 |
| 4.3 | Implementation Machine L... | 08/04/2020 | 12/04/2020 |
| 4.4 | Models Implemented | 12/04/2020 | 12/04/2020 |
| 5 | ⊟ Model Optimization Phase | 19/04/2020 | 26/04/2020 |
| 5.1 | Hyper Parameter Tuning | 19/04/2020 | 22/04/2020 |
| 5.2 | Feature Selection | 22/04/2020 | 26/04/2020 |
| 6 | ⊟ Documentation Phase | 26/04/2020 | 02/05/2020 |
| 6.1 | Group Progress Report | 26/04/2020 | 02/05/2020 |
| 7 | ⊟ Novel Model Phase | 26/04/2020 | 10/05/2020 |
| 7.1 | Novel Model Research | 26/04/2020 | 29/04/2020 |
| 7.2 | Novel Model Implementation | 29/04/2020 | 06/05/2020 |
| 7.3 | Novel Model Optimization | 06/05/2020 | 10/05/2020 |
| 8 | ⊟ Results Phase | 10/05/2020 | 17/05/2020 |
| 8.1 | Results Analysis | 10/05/2020 | 14/05/2020 |
| 8.2 | Results Visualization | 14/05/2020 | 17/05/2020 |
| 8.3 | Results Summary | 17/05/2020 | 17/05/2020 |
| 9 | ⊟ Documentation Phase | 17/05/2020 | 30/05/2020 |
| 9.1 | Final Report | 17/05/2020 | 30/05/2020 |
| 9.2 | End of Project | 30/05/2020 | 30/05/2020 |

*Figure 14 : Project Gantt chart*

# 7. Results

## 7.1. Basic Machine Learning Models Performance

The results for our implementation of the literature review models can be found in the table below. The table consists of the results of our implementation of the models on the diagnosis and prognosis dataset and the literature review model results. To ensure consistency with the literature reviews models, all models implemented in this project used a training/test split of 70% to 30% ratio.

| Basic Machine Learning Models | Diagnosis Dataset | | Prognosis Dataset | |
|---|---|---|---|---|
| | Experimental Accuracy | Benchmark Accuracy | Experimental Accuracy | Benchmark Accuracy |
| **Decision Tree** | 91.22% | 93.62% | 72.88% | 76.26% |
| **Random Forest** | 97.07% | 93.00% | 76.27% | 76.26% |
| **Logistic Regression** | 98.00% | 95.70% | 78.91% | 75.4% |
| **Support Vector Machine** | 98.83% | 98.00% | 75.21% | 78.35% |
| **Naïve Bayes** | 93.56% | 84.50% | 66.1% | 70.71% |
| **Deep Neural Network** | 95.36% | 95.00% | 82.79% | 87.5% |
| **Artificial Neural Network** | 98.24% | 92.90% | 84.74% | 90.22% |

*Table 9 : Basic machine learning models results comparison*

- Breast Cancer Wisconsin - Diagnosis data

From the literature review models, the best performing model for the diagnosis breast cancer dataset is the support vector machine (SVM) with an accuracy of 98%. This is followed by logistic regression 95.7%, deep neural network 95% and artificial neural network 92.9%.

Similarly when we performed the model ourselves with default parameters we also achieved the support vector machine as our best model followed by logistic regression, artificial neural network and deep neural network.

Since not all parameters were stated in the literature review models, most of our own performance accuracies were applied with default parameters. Also due to the difference in pre-processing and training/test split the results may differ slightly.

- Breast Cancer Wisconsin - Prognosis data

From the literature review models the best performing model for the prognosis breast cancer dataset is the artificial neural network with an accuracy of 90.22%. This is followed by the deep neural network 87.5% and support vector machine 78.35%.

Similarly when we performed the models ourselves with default parameters we also achieved the artificial neural network as the best performing model followed by the deep neural network, logistic regression and support vector machine.

Since not all parameters were stated in the literature review models, most of our own performance accuracies were applied with default parameters. Also due to the difference in pre-processing and training/test split the results may differ slightly. However results are fairly close to each other, if not, our results are lower showing that the literature review benchmarks are fairly valid since their models would have been performed better with optimised parameters.

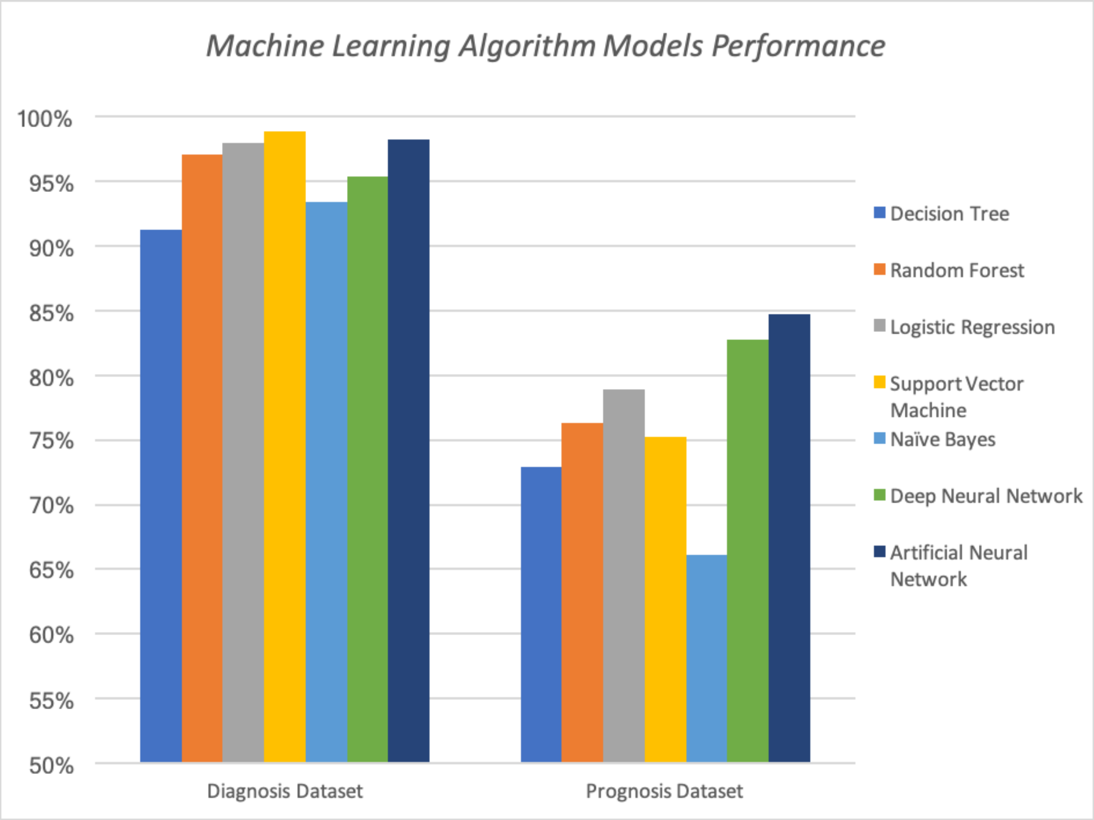


*Figure 15 : Basic machine learning models performance*

From the graph of our own performance results, the results can be seen more clearly with support vector machines performing the best for the diagnosis breast cancer dataset and the artificial neural network performing the best for the prognosis breast cancer dataset. Therefore a combination of these two methods may be suitable to hopefully improve our results.

Overall from the literature review models, support vector machine, deep neural network, logistic regression and artificial neural network seem to perform the best. The other models also have a fairly high accuracy and therefore we have decided to build a streamlined model involving neural networks and support vector machines which have the main high accuracy results along with different combinations involving logistic regression, naive bayes, random forest and decision tree.

## 7.2. Streamlined Model Performance

As mentioned in the methodologies section, our proposed model contains a structure which streamlines a set of basic machine learning models together that act as input for an Artificial Neural Network (ANN) model. To obtain the best performing model, different sets of machine learning models are streamlined together. Moreover, in order to curb the issue raised due to imbalanced data distribution and small sample size of data, upsampling was implemented on each combination model.

Figure 16 and Figure 17 illustrate the results of our streamlined method on the diagnosis and prognosis dataset, respectively.

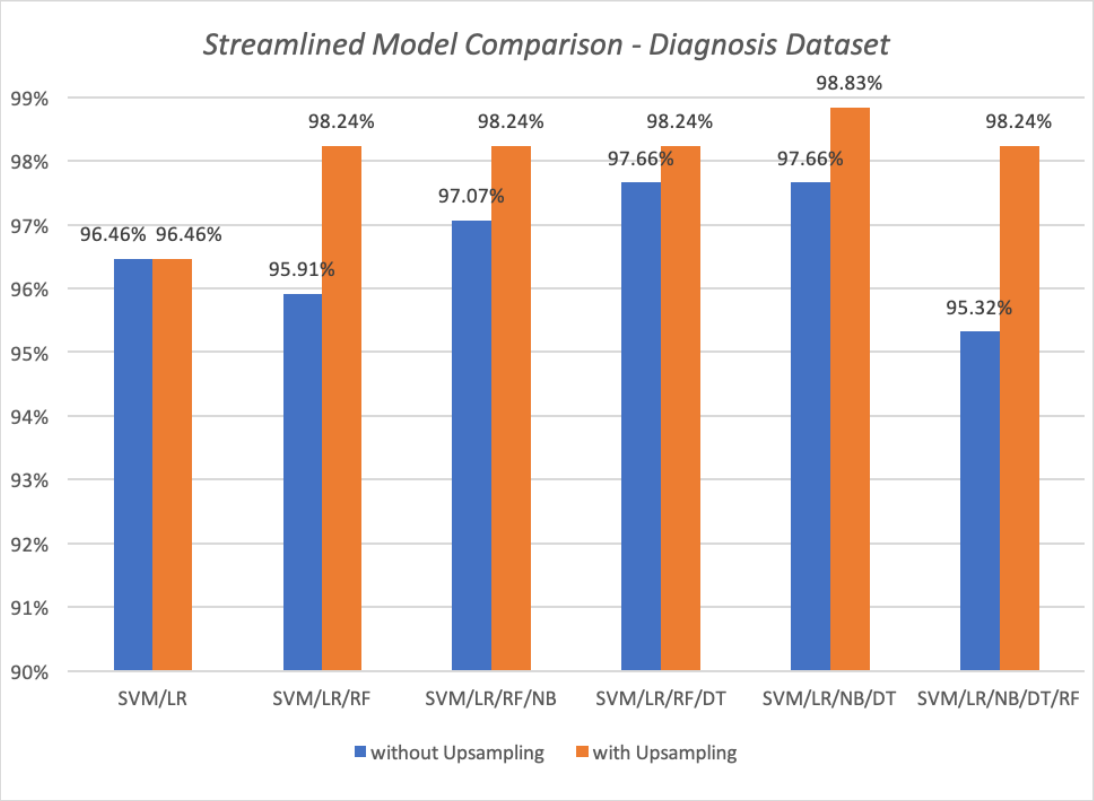


*Figure 16 : Streamlined model comparison - Diagnosis dataset*

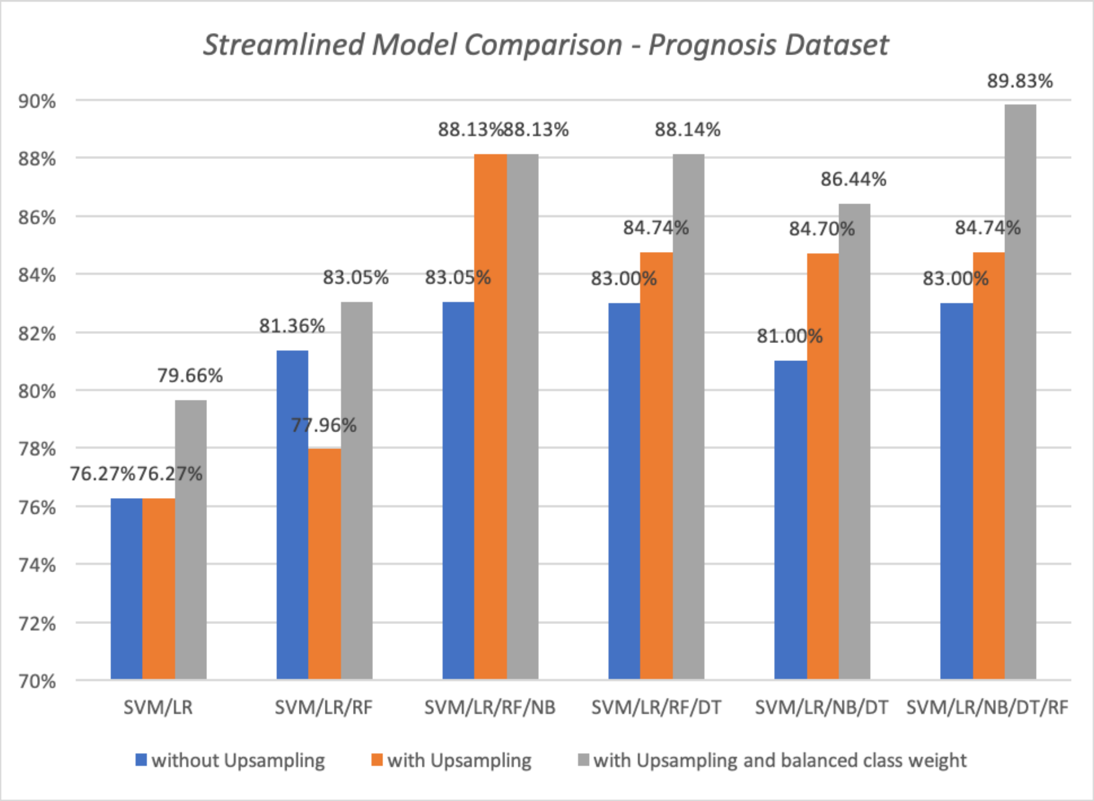
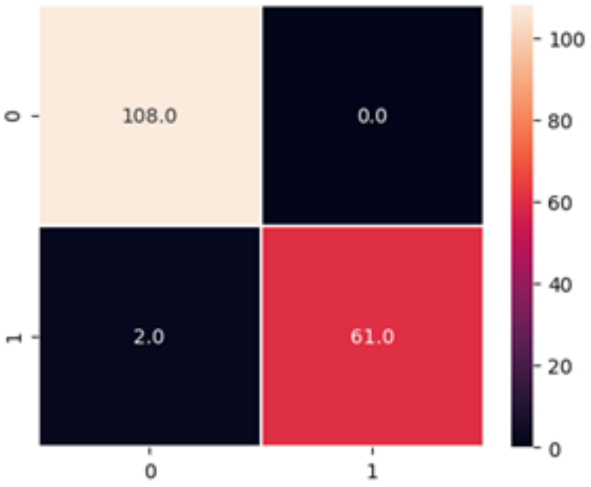


*Figure 17 : Streamlined model comparison - Prognosis dataset*

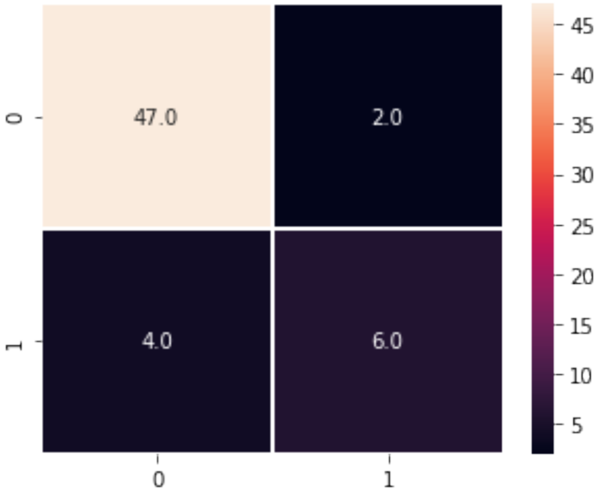
By comparing all the different streamlined models, we have identified the best performing model for each dataset. For the diagnosis dataset, the best model has a configuration of streamlining SVM + LR + NB + DT together with upsampling applied to the dataset. The performances and the confusion matrix of the model can be outlined in the following table and figure.



| | Performance |
|---|---|
| **Accuracy** | 0.988 |
| **Precision** | 1.0 |
| **Recall** | 0.968 |
| **F1** | 0.983 |

*Figure 18 : Streamlined model Confusion Matrix - Diagnosis dataset*
*Table 10 : Streamlined model Performance Summary - Diagnosis dataset*

The best streamlined model for prognosis dataset has a model configuration of combining SVM + LR + NB + DT + RF together with upsampling and balanced class weight. The following table and figure below highlight the performance and confusion matrix of the model.



| | Performance |
|---|---|
| **Accuracy** | 0.89 |
| **Precision** | 0.75 |
| **Recall** | 0.60 |
| **F1** | 0.66 |

*Figure 19 : Streamlined model Confusion Matrix - Prognosis dataset*
*Table 11 : Streamlined model Performance Summary - Prognosis dataset*

### 7.3. Further Model Performance

To further evaluate our streamlined model, we compared it with another predictive model such as an Artificial Neural Network that incorporates a Support Vector Machine as the neural network final layer. We first implemented the ANN + SVM model on both with and without upsampling technique applied. The results of the models can be seen in the following table.

| **Proposed Model** | **Upsampling** | **Diagnosis Dataset** | **Prognosis Dataset** |
|---|---|---|---|
| **ANN + SVM (Final Layer)** | - | 94.73% | 86.44% |
| **ANN + SVM (Final Layer)** | ✓ | 96.49% | 76.21% |

*Table 12 : Further model performance comparison*

The results obtained from the table indicates that without upsampling techniques applied, the ANN + SVM model trained on the diagnosis dataset showed an accuracy of 94.73% and the ANN + SVM model trained on the prognosis dataset showed an accuracy of 86.44%.

We proceeded to compare the results of the ANN + SVM models with the best streamlined model identified in the previous stage. The comparison can be seen represented in a bar plot below.

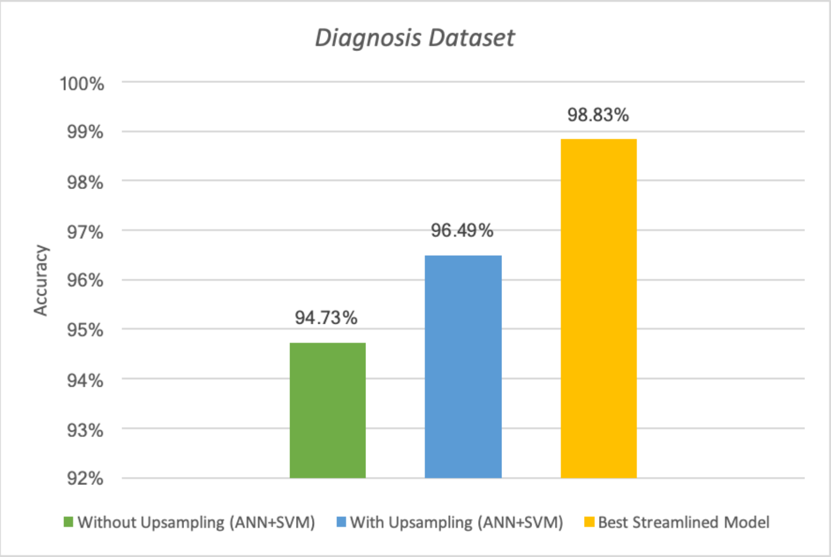


*Figure 20 : Further model performance comparison - Diagnosis dataset*

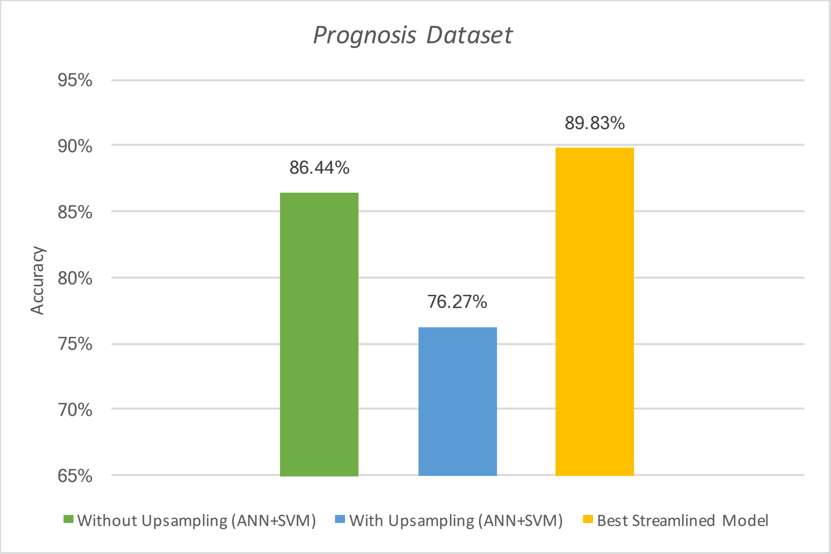


*Figure 21 : Further model performance comparison - Prognosis dataset*

By comparison, we can clearly see that our streamlined model showed a significant higher accuracy compared to the ANN + SVM models on both diagnosis and prognosis dataset.

# 8. DISCUSSION

- Basic Single Model Performance Comparison

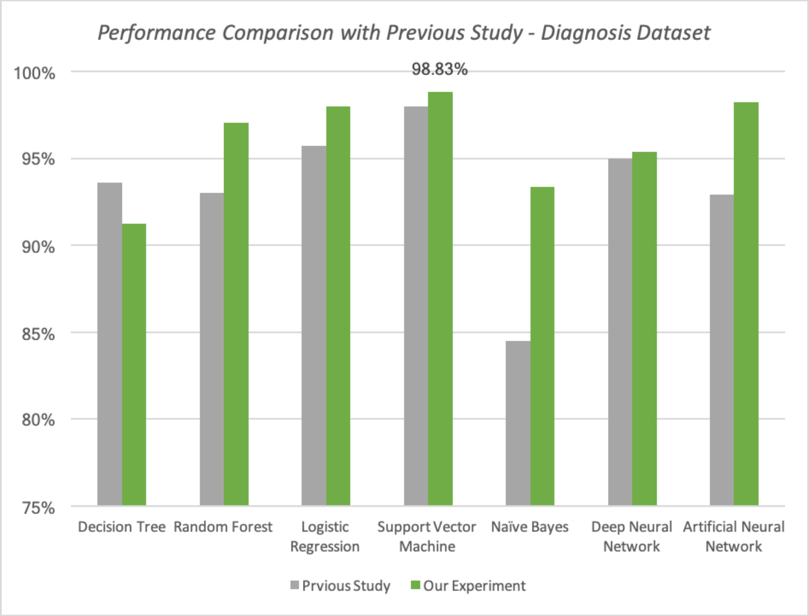


*Figure 22 : Peer model performance comparison - Diagnosis dataset*

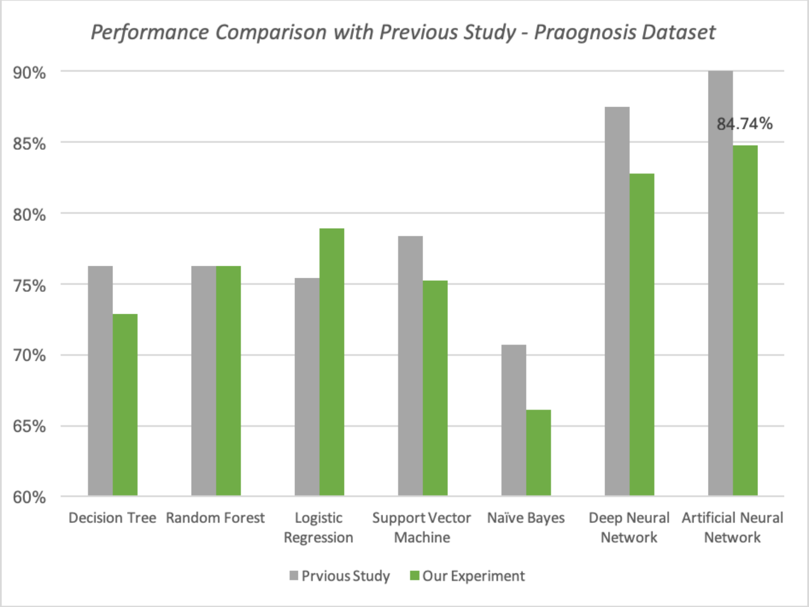


*Figure 23 : Peer model performance comparison - Prognosis dataset*

Most models on diagnosis dataset are improved compared to the previous study, especially SVM model performs the best, achieving an accuracy of 98.83%. However, only one model (Logistic Regression) beat the previous model as the graphs above illustrate. Since not every literature gave the details of their proposed models, the model was built with the default hyper-parameters setting for the comparative analysis.

- Streamlined Model Performance Comparison - Diagnosis Dataset

As the label class of diagnosis is unbalanced and the size of the data is small, we applied upsampling techniques to counter these issues. Looking at the results of the streamlined model trained on diagnosis dataset (Table 9), each model has an accuracy range of 95% to 97% without upsampling being applied to the training dataset. However, after applying the upsampling technique, we can see that each model's accuracy has improved to an average of 98.24%. We can conclude that the upsampling technique has a positive effect on the model's performance and a slight improvement of 2% to 3% on the accuracy. On the other hand, the different models used in each combination has a positive improvement on the model overall performance. From the table, we can see that by replacing a Naïve Bayes classifier with a Decision Tree classifier improved the accuracy from 97.07% to 97.66%.

The model has an accuracy of 98.8% which is slightly higher than the rest of the model and it has a precision score of 1.0 which indicates that It also has a recall of 0.968 and f1-score of 0.983. On the other hand, the confusion matrix shows that the model predicted the cancer to be benign a total of 110 times and malignant 61 times. Whereas the actual results show that there are 108 cases to be benign and 63 cases to be malignant. It also predicted 2 cases to be benign, but it turns out to be malignant. However, it correctly predicted all malignant cases. More details can be found from Figure 18 and Table 10 in result part above.

- Streamlined Model Performance Comparison - Prognosis Dataset

Similarly, for the prognosis dataset the class label is highly unbalanced as well. Thus, in addition to upsampling, we have also included a balanced class weight parameter to improve model performance on prognosis classification. For the streamlined model that was trained on the prognosis dataset (Figure 17), we can observe that each model that does not have upsampling technique applied has an accuracy range of 76% to 83%. Similar to the diagnosis dataset, when upsampling is applied, the accuracy of each model has improved to a range of 76% to 88%. An additional parameter was added for the prognosis dataset models, which was to set the class weight to be balanced. We can see by applying this parameter a further improvement of 2% to 6% of accuracy across all the models.

The best streamlined model for prognosis dataset has an accuracy of 89.8% which is significantly higher compared to the rest of the streamlined model trained on prognosis dataset. It also has a precision of 0.75, a recall of 0.60 and f1-score of 0.66. Looking at the confusion matrix, we can observe that the model predicted 51 times that a case is non-recurrent and predicted 8 times that a case is recurrent. Compared to the actual result, there are 49 non-recurrent cases and 10 recurrent cases. The model incorrectly predicted 4 cases to be non-recurrent and 2 cases to be recurrent. More details can be found from Figure 19 and Table 11 above.

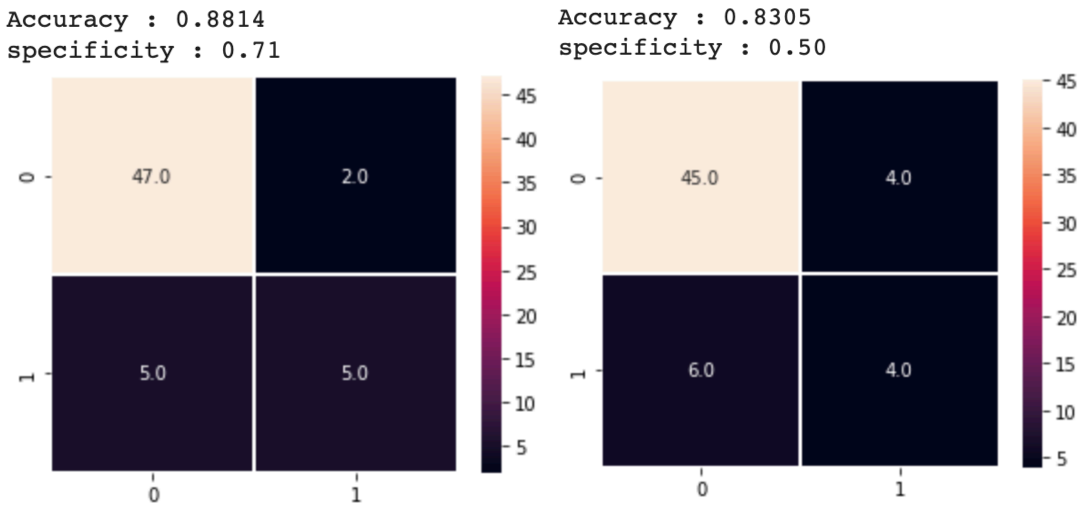
- Further Model Performance Comparison

For the model that was trained on the diagnosis dataset, it showed an increase in performance when upsampling was applied. Table 12 shows that the accuracy of ANN + SVM model improved from 94.73% to 96.49%. However, when upsampling was applied on the model trained on the prognosis dataset, its accuracy dropped from 86.44% to 76.21%, as shown in Table 12. Overall, our streamlined model outperforms over ANN + SVM model on both diagnosis and prognosis dataset.

- Hyper-parameter Comparison

1. K-Fold Cross Validation

When dividing the data into a k number of sections/folds in order to ensures that each fold is used as a testing set using K-Fold cross validation, Choosing an appropriate K must be carefully done to avoid misrepresentation and high variance/ bias. K is usually chosen as 5 or 10, that are demonstrated empirically to yield error rate estimates that suffer neither from high bias nor from high variance. During training our streamlined model, K=5 and 10 have been experimented while keeping every parameters same. The confusion matrixes below prove that the model (SVM + LR + RF + DT trained on prognosis dataset) outperforms when K is 5, rather than 10.



```
Accuracy : 0.8814
specificity : 0.71
```

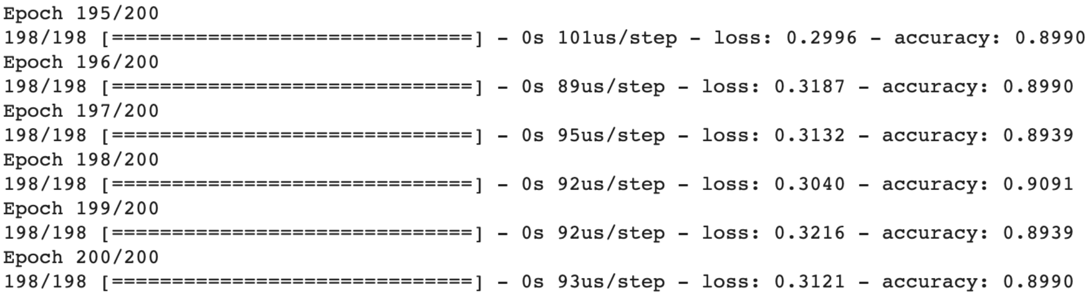
```
Accuracy : 0.8305
specificity : 0.50
```

*k=5* *k=10*

*Figure 24, 25 : Model performance with different K (5, 10)*

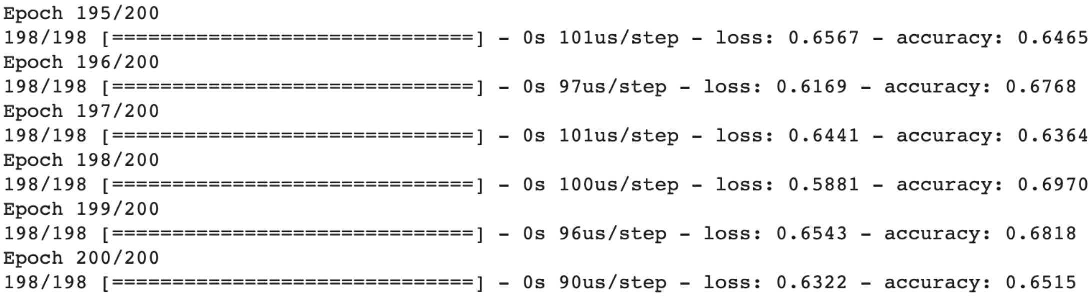
2. Dropout Rate

The dropout hyper-parameter indicates the probability of training a node in each layer, where 1 means no dropout, and 0 means no outputs from the layer at all. In the last stage of the streamlined model (the result of SVM + LR + RF + DT trained on prognosis dataset feeds Artificial Neural Network), different dropout rate of 0.1 and 0.8 have been experimented while keeping every parameters same.



```
Epoch 195/200
198/198 [==============================] - 0s 101us/step - loss: 0.2996 - accuracy: 0.8990
Epoch 196/200
198/198 [==============================] - 0s 89us/step - loss: 0.3187 - accuracy: 0.8990
Epoch 197/200
198/198 [==============================] - 0s 95us/step - loss: 0.3132 - accuracy: 0.8939
Epoch 198/200
198/198 [==============================] - 0s 92us/step - loss: 0.3040 - accuracy: 0.9091
Epoch 199/200
198/198 [==============================] - 0s 92us/step - loss: 0.3216 - accuracy: 0.8939
Epoch 200/200
198/198 [==============================] - 0s 93us/step - loss: 0.3121 - accuracy: 0.8990
```

*Figure 26 : The loss and training accuracy from last six epoch (Dropout rate = 0.1)*



```
Epoch 195/200
198/198 [==============================] - 0s 101us/step - loss: 0.6567 - accuracy: 0.6465
Epoch 196/200
198/198 [==============================] - 0s 97us/step - loss: 0.6169 - accuracy: 0.6768
Epoch 197/200
198/198 [==============================] - 0s 101us/step - loss: 0.6441 - accuracy: 0.6364
Epoch 198/200
198/198 [==============================] - 0s 100us/step - loss: 0.5881 - accuracy: 0.6970
Epoch 199/200
198/198 [==============================] - 0s 96us/step - loss: 0.6543 - accuracy: 0.6818
Epoch 200/200
198/198 [==============================] - 0s 90us/step - loss: 0.6322 - accuracy: 0.6515
```

*Figure 27 : The loss and training accuracy from last six epoch (Dropout rate = 0.8)*

Both loss and training accuracy seem better where the model with dropout rate = 0.1 in comparison with dropout rate = 0.8. Hence, dropout rate of 0.1 was chosen.

- Model Implementation Discussion

Our proposed streamlined model is inspired by neural networks' structures. On the previous stage of streamlined models, each classic machine learning model makes their own classification outcome independently. Then, the result of each of predicted class is fed into Artificial Neural Network as an input. As results from more than on models (basic machine learning algorithms) are combined, the final predictions become more powerful and accurate. Hence, our final streamlined model could be more robust classification model with much lower variance and improved generalisability compared to the predictions from each individual model[28]. At the same time, upsampling and cost-sensitive extensions could significantly improve model performance on both diagnosis and prognosis dataset.

- Project Aims Discussion

Based on the development of data science in the medical industry, this study gives combination models for breast cancer diagnosis and prognosis, which are in the scope of binary classification problems. Through combining several machine learning models, the streamlined models result in higher accuracy for both Breast Cancer Wisconsin diagnosis and prognosis dataset. This positive improvement is also expected to be able to help decreasing uncertainty of breast cancer diagnosis prediction. In addition, our prediction model associates to design more appropriate treatments for patients who are already suffered from breast cancer and elevate the survival rate accordingly.

# 9. LIMITATIONS AND FUTURE WORKS

## 9.1. Limitations

- Lack of manpower due to COVID-19 travel restriction

Due to the travel restriction, one of our team members suspended studying for this semester, and the number of team members decreased to 4 from 5. We supposed the lack of manpower might lead to work overloaded and difficult to follow the project plan.

- Online meeting

Every on-site meeting turned to on-line during this uncertain time. Since the group members and the supervisor & client are never able to meet each other, extra time was required to organise weekly meetings and to discuss even. Moreover, Zoom - a platform we relied on for meetings - used to be lagging and unstable sometimes.

- Hardware limitation

During the Grid Search in order to find the optimal hyper-parameters for Artificial Neural Network, we faced extremely long computation time due to hardware limitation.

- Comparative analysis of the previous study

Some of the papers that have been reviewed through this project do not give readers sufficient information (e.g. hyper-parameters) about their models. This does not allow us to do comparative analysis in an exactly same environment.

- Biased dataset with small sample size

Breast Cancer Wisconsin-Prognosis data is such small and imbalanced data - consists of 151 non-recur outcomes and 47 recur outcome, and after splitting the data into train and test set, test set has only 10 or 11 recur class out of 59 of the total outcome. Since the majority of machine learning algorithms that are designed for prediction assume data has an equal proportion of samples for each class, the results form the models may perform poor accuracy. In other words, the machine learning algorithms are much more likely to classify new observations to the majority class, since the models are supposed to minimise errors and probability of instances belonging to the majority class is significantly high in the imbalanced data set. To tackle this issue, the cost-sensitive extension on several machine learning algorithm and up-sampling for training dataset are implemented as fully described in '4.3.2 Imbalanced data' part earlier. Whereas, we still encounter the small sample size of test data and have difficulty in limited improvement in model performance, especially for Breast Cancer Wisconsin-Prognosis data.

- Overfitting

Due to small sample size of dataset and the complexity machine learning algorithms, we were confronted with the over-fitted model. To ensure the model is neither under-fitted nor over-fitted, some techniques such as K-Fold cross validation and parameter (e.g. batch size or epoch in neural network) were performed that may help model generalisation.

- Unreproducible result

Although Artificial Neural Network(ANN) is trained on the same data, it sometimes produces different results because of its randomness. Even if we remove the randomness from the training and test dataset when splitting the data, we occasionally get an inconsistent outcome with every execution. This is results from randomness in initialisation(e.g. biases and weights), layer, regularisation(e.g. dropout), and optimisation and it may be possible to get the same results by some solutions that fix the problem of randomness. However, GPU server, third-party library, and sophisticated model may cause the unreproducible results.

## 9.2. Future Works

- Hardware improvement

Using the higher performance of hardware to run the optimisation algorithm and fine-tune on neural networks is expected to improve the model.

- Contact the authors of the previous study

As mentioned above, we were not able to gain a full description of how machine learning models form the previous study have been trained (e.g. hyper-parameters, optimisation function, etc.). We might need to contact authors to further discussion of more details about the models and even the techniques they applied and conduct more elaborate or reliable comparative analysis. Moreover, we expect to come up with some ideas of a novel prediction model that is possibly designed further.

- Missing value handling

The data that have missing values in Breast Cancer Wisconsin-Prognosis data have been simply removed in this project. However, other missing value handling skills could be performed, and this might produce a different result. For instance, missing values can be dealt with by simple mean/median/mode imputation or by prediction algorithms for imputation(e.g. Linear Regression and K-Nearest Neighbours).

- Different sampling technique

For this project, K-Fold Cross Validation, that divides the data into a k number of sections/folds and ensures that each fold is used as a testing set, is implemented. Likewise, Stratified Cross-Validation could be possibly compared in the future study. Stratified Cross-Validation splits the data into folds, ensuring that each fold has the same proportion of the class outcome, so this might be preferred over K-Fold Cross Validation when the data shows imbalance distribution.

- More sophisticated models/combination

We intend to design and implement an elaborate and advanced predictive model using other machine learning algorithms. More details will be explained as follows.

[Bayesian Networks]

Bayesian Networks are found to be a technique that is especially suited for capturing and reasoning with uncertainty[29]. They have been applied in biomedicine and health-care for a while now and have become increasingly prevalent for handling the vague diagnoses of disease. As we all know, predicting the future is always uncertain, hence prognosis is more predominant than the diagnosis. However, prognostic Bayesian networks still have a clear general temporal structure(Lucas et al., 2004).

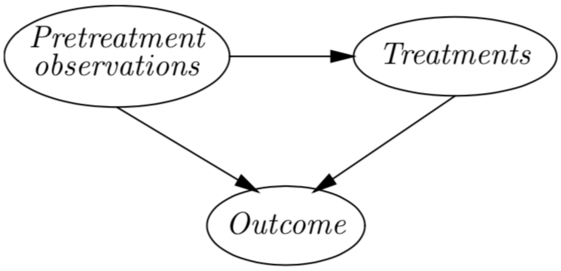


*Figure 28 : General structure of a prognostic Bayesian network(Lucas et al., 2004)*

- Applying different medical dataset

As we successfully carried out designing the predictive model, we further suggest applying different medical dataset of breast cancer that has similar attributes or also add supplementary data that has a demographic feature in order to develop the model adopting more information.

- Different performance measurement

In this study, the prediction model is evaluated based on its accuracy. Nevertheless, classification accuracy is subject to be not enough to measure the performance. For instance, in this project, False-Negative value is probably worse than False Positive value and needs to be minimised, because if a patient is diagnosed as healthy when a patient is actually ill, this is more serious than classifying a healthy patient unhealthy (malignant or recur).

| | | *Predicted/Classified* | |
|---|---|---|---|
| | | Non-recur | Recur |
| *Actual* | Non-recur | 80 | 0 |
| | Recur | 15 | 5 |

*Figure 29 : Confusion matrix example 1*

| | | *Predicted/Classified* | |
|---|---|---|---|
| | | Non-recur | Recur |
| *Actual* | Non-recur | 60 | 20 |
| | Recur | 5 | 15 |

*Figure 30 : Confusion matrix example 2*

Looking at the confusion matrix examples above, the accuracy of the models are 85% and 75%, respectively. If the performance is evaluated by the accuracy only, the best model must be the first model, while it classifies 15 patents of 20 patients who have cancer recurred to non-recur class and let them go home missing an opportunity to get essential treatment. We discover that the first model has higher accuracy than the second model, but its performance measurement is less suitable for medical data. Hence, we further suggest using different performance measurement, such as recall score. Recall calculates the number of True Positives divided by the number of True-Positives and the number of False-Negatives, and it is also named the True Positive Rate or Sensitivity.

# References

[1] Institue, N. C. (2014). *Breast Cancer*. Retrieved from https://web.archive.org/web/20140625232947/http://www.cancer.gov/cancertopics/types/breast

[2] Council, A. C. (n.d.). *Breast Cancer*. Retrieved from https://www.cancer.org.au/about-cancer/types-of-cancer/breast-cancer/

[3] Hiba Asri; Hajar Mousannif; Hassan Al Moatassimec; ThomasNoeld. (2016). *Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis*. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050916302575

[4] Khatib, O. (2006). Guidelines for the early detection and screening of breast cancer. WHO.

[5] Latest global cancer data:Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. Who.int. (2018). Retrieved 24 March 2020, from https://www.who.int/cancer/PRGlobocanFinal.pdf.

[6] *Diet, nutrition, physical activity and breast cancer*. Wcrf.org. (2018). Retrieved 23 March 2020, from https://www.wcrf.org/dietandcancer.

[7] Milosevic, M., Jankovic, D., Milenkovic, A., & Stojanov, D. (2018). Early diagnosis and detection of breast cancer. *Technology And Health Care*, *26*(4), 729-759. https://doi.org/10.3233/thc-181277

[8] *Diet, nutrition, physical activity and breast cancer*. Wcrf.org. (2018). Retrieved 23 March 2020, from https://www.wcrf.org/dietandcancer.

[9] Balentine, J. *Breast Cancer: Signs, Symptoms, Causes, Treatment, Stages & Survival Rates*. RxList. Retrieved 25 March 2020, from https://www.rxlist.com/breast_cancer_facts_stages/article.htm#breast_cancer_facts.

[10] Goel, V. (2018). *Building a Simple Machine Learning Model on Breast Cancer Data*. Medium. Retrieved 23 March 2020, from https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3.

[11] Kharya, S. (2012). Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease. *International Journal Of Computer Science, Engineering And Information Technology*, *2*(2), 55-66. https://doi.org/10.5121/ijcseit.2012.2206

[12] AlamKhan, R., Ahmad, N., & Minallah, N. (2013). Classification and Regression Analysis of the Prognostic Breast Cancer using Generation Optimizing Algorithms. *International Journal Of Computer Applications*, *68*(25), 42-47. https://doi.org/10.5120/11754-7423

[13] Nisbet, R., Miner, G., Yale, K., Elder, J., & Peterson, A. (2009). *Handbook of statistical analysis and data mining applications* (p. Chapter 13 - Model Evaluation and Enhancement).

[14] Agarap, A. (2018). On breast cancer detection. Proceedings Of The 2Nd International Conference On Machine Learning And Soft Computing - ICMLSC '18. https://doi.org/10.1145/3184066.3184080

[15] *1.4. Support Vector Machines — scikit-learn 0.22.2 documentation*. Scikit-learn.org. Retrieved 26 March 2020, from https://scikit-learn.org/stable/modules/svm.html.

[16] Entezari, R. (2018). Breast Cancer Diagnosis via Classification Algorithms. Retrieved 25 May 2020

[17] Maglogiannis, I., Zafiropoulos, E., & Anagnostopoulos, I. (2007). An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. *Applied Intelligence*, *30*(1), 24-36. https://doi.org/10.1007/s10489-007-0073-z

[18] Wikipedia contributors. (2020, January 14). Linear regression. In Wikipedia, The Free Encyclopedia. Retrieved March 25, 2020, from https://en.wikipedia.org/w/index.php?title=Linear_regression&oldid=935782381

[19] Khairunnahar, L., Hasib, M., Rezanur, R., Islam, M., & Hosain, M. (2019). Classification of malignant and benign tissue with logistic regression. *Informatics In Medicine Unlocked*, *16*, 100189. https://doi.org/10.1016/j.imu.2019.100189

[20] Sun, D., Wang, M., & Li, A. (2019). A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data. *IEEE/ACM Transactions On Computational Biology And Bioinformatics*, *16*(3), 841-850. https://doi.org/10.1109/tcbb.2018.2806438

[21] Park, Y., & Lek, S. (2016). Artificial Neural Networks. *Developments In Environmental Modelling*, 123-140. https://doi.org/10.1016/b978-0-444-63623-2.00007-4

[22] Yana, M., & Stoyan, S. (2019). Classifying breast cancer data from the Wisconsin database using Artificial Neural Network. Retrieved 24 May 2020

[23] Zhang, Y. (2019). *Deep Learning in Winonsin Breast Cancer Diagnosis*. Medium. Retrieved 24 March 2020, from https://towardsdatascience.com/deep-learning-in-winonsin-breast-cancer-diagnosis-6bab13838abd.

[24] Ferguson, K., & Ferguson, K. (2020). *Why It's Important to Standardize Your Data - Atlan | Humans of Data*. Atlan | Humans of Data. Retrieved 1 May 2020, from https://humansofdata.atlan.com/2018/12/data-standardization/.

[25] *Violin plot*. En.wikipedia.org. (2020). Retrieved 1 May 2020, from https://en.wikipedia.org/wiki/Violin_plot.

[26] Hintze, J., & Nelson, R. (1998). Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, *52*(2), 181. https://doi.org/10.2307/2685478

[27] *Heat map*. En.wikipedia.org. (2020). Retrieved 1 May 2020, from https://en.wikipedia.org/wiki/Heat_map#History.

[28] Koidan, K. (2019). *7 Effective Ways to Deal With a Small Dataset | Hacker Noon*. Hackernoon.com. Retrieved 11 June 2020, from https://hackernoon.com/7-effective-ways-to-deal-with-a-small-dataset-2gyl407s.

[29] Lucas, P., van der Gaag, L., & Abu-Hanna, A. (2004). Bayesian networks in biomedicine and health-care. *Artificial Intelligence In Medicine*, *30*(3), 201-214. https://doi.org/10.1016/j.artmed.2003.11.001