2022

Comparison of neural networks and tree-based ensemble methods in detecting correlates of breast cancer survival.

Katam, Ruth Jepchirchir
Strathmore Institute of Mathematical Sciences
Strathmore University

Comparison of Neural Networks and Tree-based Ensemble Methods in Detecting Correlates of Breast Cancer Survival

135214, Katam Ruth Jepchirchir

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Statistical Science of of Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

October, 2022

Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: ..... **Katam Ruth Jepchirchir** .....

Signature: ....

Date: ..... September 8, 2022 .....

Approval

The thesis of Katam Ruth Jepchirchir was reviewed and approved by the following:

Dr Linda Chaba

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu

Dean,

Institute of Mathematical Sciences, Strathmore University.

Dr. Bernard Shibwabo

Director,

Office of Graduate Studies, Strathmore University.

Abstract

Breast cancer is common among women impacting about 2.1 million women each year, and causing a big number of cancer-related deaths. Most times doctors have a struggle in diagnosing the stage to determine accurately and needed medication. Therefore, accurate detection of correlates of breast cancer survival is paramount. This study sought to compare the performance of Neural Networks and Tree-based Ensemble methods to predict breast cancer survival, elucidating on factors causing breast cancer based on clinical data for timely intervention. The accuracy score, recall score, precision score, Area under Receiver-Operating Characteristic Curve, and F1 score were used to evaluate the performance of each model in discerning between breast cancer survivors and non-survivors. XGboost and LSTM exhibited an outstanding performance in the classification of Breast cancer patients. However, XGboost was the most optimal model. The results depicted that age at diagnosis, pam50+ claudin low subtype her2, 3 gene classifier subtype high, profile,radiotherapy,Nottingham prognostic index,type of breast surgery breast conserving, type of breast surgery mastectomy, mutation count, lymph nodes examined positive, tumor stage, tumor size, 3 gene classifier subtype low profile, pre inferred menopausal state and Post inferred menopausal state. among others were the most important correlates of survival from breast cancer.

Keywords

Breast cancer, Breast cancer survival, Ensemble Tree-based, Neural Networks, Machine Learning

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BC: Breast Cancer

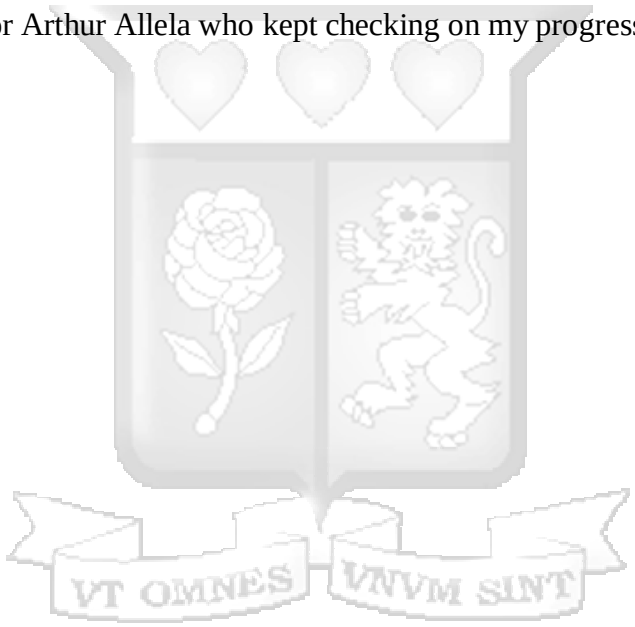
LSTM: Long Short-Term Memory

MLP: Multi-Layer Perceptron

CNN: Convolutional Neural Network

Acknowledgement

Thanking God for this far he has brought me. My experience could not have been completed without the support and guidance of people around me to ensure completion of this study. First and foremost, very special gratitude goes to my supervisor who has been a mentor and has given me guidance all through Dr. Linda Chaba. To my family members; my husband, parents and siblings who never stopped encouraging me to push on and the prayers that made me come this far. I acknowledge the support I have received from the lecturers and my fellow students. My mentor Arthur Allela who kept checking on my progress. God bless you all.



Dedication

To my support system my family, friends, lecturers and fellow students.

Chapter 1

Introduction

1.1 Background to the study

Cancer is one of the killer diseases and it continues to rise globally. Breast cancer is a common cause of cancer-related deaths among women. In Africa, noncommunicable diseases have been sidelined and focus is more on communicable diseases. Limited resources, research, and lack of awareness have affected the rate of cancer diagnosis. According to the World Health Organization, breast cancer is the most common malignancy among women (https://www.who.int/news-room/fact-sheets/detail/breast-cancer). In 2020, 2.3 million women were affected with 685,000 deaths around the globe. World health organization Statistics 2020 showed that 7.8 million women diagnosed with breast cancer in the past 5 years were alive making it the world's most prevalent cancer.

According to Sung et al. (2021), an estimated 19.3 million new cancer cases (18.1 million excluding nonmelanoma skin cancer) and almost 10.0 million cancer deaths and 9.9 million excluding nonmelanoma skin cancer occurred in 2020 worldwide. Female breast cancer has surpassed lung cancer as the most commonly diagnosed cancer, with an estimated 2.3 million new cases. In sub-Saharan Africa cancer is expected to rise to over 85

In Kenya, the number of new breast cancer cases affecting all ages was reported at 6.799 (25.6%) in 2020 and ranked 1 among all cancer cases reported (Sung et al., 2021). Some of the key risk factors of breast cancers are age, gender, affluence, family history, breast conditions, alcohol consumption and obese (Breast Cancer Deadline 2020; Report to the nation-breast cancer).

Breast cancer is caused by uncontrolled growth of breast cells. The body consists of cells that reproduce and die for new ones to be formed. Sometimes the damaged cells or abnormal cells grow and multiply when they should not. This forms lumps of tissues (tumor) that can be cancerous (malignant) or non cancerous (benign). Malignant tumors can spread to other parts of the body and form new tumors, a process called metastasis while benign tumors do not.

Early detection and treatment of breast cancer helps lower mortality rate. Doctors use the Tumor, Node, Metastasis (TNM) method for breast cancer diagnosis and scan to give the stage of the cancer.The cancer stage describes the size of a tumor how far it has spread from where it originated. Prognosis helps check the recurrence of disease, predict survival rate and establish a treatment plan by looking at the likely course and predicting the outcomes of it. The main treatments for breast cancer are surgery, radiotherapy, chemotherapy, hormone therapy and biological therapy.

Diagnosis, prognosis and survival rate of breast cancer is hard to tell and determine. Doctors are expected to read and interpret large amounts of imaging data when mammography is done which decreases accuracy.

The use of Artificial Intelligence and Machine Learning techniques has helped improve detection and prediction of breast cancer and has helped reduce breast cancer mortality. This study will thus compares the performance of Neural Network and tree-based ensemble methods in detecting correlates of breast cancer survival. The study was motivated by Ganjisaffar et al. (2011). He studied Tree-based Ensemble models in predicting breast cancer which depicted better performance in the learning technique based on public data evaluations. Zeng (2017) suggested that using Tree-based models and especially boosting to choose relevant predictors is a viable and competitive approach in predicting an aggregate. It was also found that all the top teams ranked by Yahoo learning challenge all utilized Tree-based Ensemble methods (Ganjisaffar et al., 2011).

Neural Networks are also widely used Machine Learning methods to learn data (feature) representations at multiple levels of abstractions. They are powerful tools utilized for building cancer prediction models from micro-array data (Daoud and Mayo, 2019). The

Neural Networks which have gained traction in cancer prediction are Multi Layer Perceptron (MLP), Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) (Daoud and Mayo, 2019). LSTM has been said to be the most robust Neural Network in predicting common cancers such as breast cancer, liver cancer and lung cancer (Andjelkovic et al., 2022).

1.2 Problem statement

Breast cancer is common among women impacting 2.1 million women each year, and causes a big number of cancer related deaths (Omar et al., 2020). According to Rivera-Franco and Leon-Rodriguez (2018) cancer survival is dependent on affordability of diagnostics that determines the staging. Timely and affordable treatment is essential to ensure survival. Most times doctors have a struggle in diagnosing the stage to determine accurate and needed medication. Therefore, an accurate detection of correlates of breast cancer survival is paramount.

Nevertheless, the accurate and rapid detection of correlates of breast cancer currently needs a great deal of effort and experience with the processing and analysis of clinical data. Machine Learning models have been touted to enrich the insights hitherto foreseen or found with traditional models as Machine Learning and could uncover hidden patterns in data (Gupta et al., 2016). Thus, Machine Learning techniques are the most ideal in detecting the correlates of breast cancer survival particularly Neural Networks and ensemble tree-based models which have depicted outstanding performance in prediction of common cancer diseases (Andjelkovic et al., 2022). Although progress has been made utilizing above Machine Learning techniques, the classification and prediction performance of the existing methods is still far from satisfactory, and therefore a room for further improvement still exists. Moreover, some Neural Networks and Ensemble Tree-based models such as Convolutional Neural Network (CNN), Gradient Boosting and Adaboost have not been well utilized in detecting correlates of breast cancer survival.

1.3 Objectives

General Objective:

The main aim of the study is to compare the performance of Neural Networks and Tree-based Ensemble methods in classification of breast cancer survival.

Specific Objective:

1. To evaluate and compare the performance of three Neural Networks(Multilayer perceptron, Convolutional Neural Networks, Long Short-Term Memory) and five Tree-based Ensemble models(Random forest, Extra tree classifier, Gradient boost, AdaBoost, and XGBoost) in the prediction of breast cancer survival.
2. To determine correlates of breast cancer prognosis and survival.
3. To determine survival status of breast cancer patients.

1.4 Research Questions

1. What is the performance of various tree-based ensemble models in the prediction of breast cancer survival?.
2. What is the most robust model in predicting breast cancer prognosis and survival?
3. What are the factors associated with breast cancer survival?
4. What are the survival rates of breast cancer patients?

1.5 Significance of study

An accurate detection of breast cancer correlates is imperative. This research will help find improved and better ways to detect the correlates (Treatment,Prognosis,demographic factors

etc) of breast cancer. The study will add to the literature that focuses on breast cancer, help doctors get good and accurate detection on breast cancer for better treatment.

Chapter 2

Literature review

Artificial Intelligence and Machine Learning plays an important role in the healthcare field. It has ensured improved health care. Machine learning, deep learning are branches of Artificial intelligence and have been used widely in detection and classification of tumors e.g breast and brain (Fu et al., 2019).

The author Kirubakaran et al. (2017) in their study looked at breast cancer detection and diagnosis. The main objective of their study was to look at correctness in classifying data with respect to effectiveness and efficiency. The algorithms they used included Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) datasets and their focus was on the accuracy, precision, sensitivity and specificity. The result showed that SVM outperformed other models.

Using Machine Learning and data mining (Magboo and Magboo, 2021) did a comparison on popular Machine Learning models (Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and Support Vector Machines) classifying breast cancer recurrences on four different configurations: a) only scaling applied, b) scaling with PCA, c) scaling with PCA and oversampling of minority class and d) using only selected features with high correlation. The results showed that Logistic Regression performed well in the different metrics used which included precision, recall, accuracy, f1 score (weighted), AUROC, AUPROC, and Cohen Kappa statistics. Support Vector Machines was the second best, then by K-Nearest Neighbors. Naive Bayes performed poorly especially in the scaling with PCA configuration. They concluded that the Logistic Regression model serves as a potential model for predicting breast cancer recurrence. This will enable doctors to give proper treatment based on a patient's features that correspond to prognosis.

Lou et al. (2020) looked at the accuracy of Artificial Neural Networks (ANN), KNN, SVM, NB, and Cox Proportional Hazards Regression Models (COX) to predict breast cancer recurrence in a span of 10 years after breast surgery. Artificial Neural Networks had the highest prediction performance. The authors had a set of selected predictors. Demographic characteristics, clinical characteristics, quality of care, and preoperative quality of life were significantly associated with recurrence of breast cancer within 10 years after surgery. Surgeon volume was the best predictor of recurrence within 10 years after surgery, followed by hospital volume and tumor stage. From their conclusion, Machine Learning algorithms improve precision in managing patients after breast cancer surgery and help in understanding of risk factors that lead to recurrence after breast cancer surgery.

Abreu et al. (2016) did a systematic review on prediction of breast cancer recurrence using machine learning. In their study they were focusing on metastasis and recurrence in line to survival. Using Wisconsin Prognostic Breast Cancer Data they used (Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and Support Vector Machines).

In their study Roberto Cesar et al. (2020) looked at several methods to detect breast cancer in post-surgical patients.They used decision trees, Naïve Bayes,SMO Poly-kernel and Support Vector Machines models. Simple K-Means algorithm was then used as a cluster method. It was integrated with the classification methods to ensure best results.The best results was obtained from SMO poly=kernel with simple k-means.From their study and results obtained they concluded that breast cancer in patients that have undergone surgery can be detected with the use of data mining methods.

Fatima et al. (2020) did a comparative study and gave the use of machine learning,data mining and Machine Learning techniques in detecting breast cancer recurrences shading light and providing information. A summary was given on the different deep learning,machine learning,and data mining algorithms for the prediction of breast cancer.

The proposed metric for evaluating the machine learning model on breast cancer survival projection is Area Under the Curve (Kaur et al., 2022). Other metrics that have been proposed by other scholars include; accuracy, sensitivity (recall) and precision (Montazeri et al., 2016).

Chapter 3

Methodology

Areas that will be handled in this section include data description, data preprocessing, machine leaning models, evaluation metrics and data imbalance. The study compares Neural Networks and Tree-based Ensemble methods to arrive at the objectives.

3.1 Data.

Dataset was obtained from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database, a Canada-UK Project which contains targeted sequencing data of 1,980 primary breast cancer samples.Clinical and genomic data was downloaded from cBioPortal https://www.cbioportal.org/.

Name	Description
age_at_diagnosis	Age of the patient at diagnosis time
type_of_breast_surgery	Breast cancer surgery type: 1- **Mastectomy**, which is surgery to remove all breast tissue from a breast as a way to treat or prevent breast cancer. 2- **Breast Conserving,** which is a surgery where only the part of the breast that has cancer is removed
cancer_type_detailed	Detailed Breast cancer types: 1- Breast Invasive Ductal Carcinoma 2- Breast Mixed Ductal and Lobular Carcinoma 3- Breast Invasive Lobular Carcinoma 4- Breast Invasive Mixed Mucinous Carcinoma 5- Metaplastic Breast Cancer
chemotherapy	If or not the patient had chemotherapy as a treatment (yes/no)
pam50_+_claudin-low_subtype	Pam 50: is a tumor profiling test that helps show whether some estrogen receptor-positive (ER-positive), HER2-negative breast cancers are likely to metastasize (when breast cancer spreads to other organs). The claudin-low breast cancer subtype is defined by gene expression characteristics, most prominently: Low expression of cell–cell adhesion genes, high expression of epithelial–mesenchymal transition (EMT) genes, and stem cell-like/less differentiated gene expression patterns
cohort	Cohort is a group of subjects who share a defining characteristic (It takes a value from 1 to 5)
nottingham_prognostic_index	It is used to determine prognosis following surgery for breast cancer. Its value is calculated using three pathological criteria: the size of the tumor; the number of involved lymph nodes; and the grade of the tumor.
inferred_menopausal_state	Whether the patient is post-menopausal or not (post/pre)
lymph_nodes_examined_positive	To take samples of the lymph node during the surgery and see if there were involved by the cancer
overall_survival_months	Duration from the time of the intervention to death
overall_survival	Target variable whether the patient is alive of dead.
radiotherapy	Whether or not the patient had radio as a treatment (yes/no)
3-gene_classifier_subtype	Three Gene classifier subtype It takes a value from 'ER-/HER2-', 'ER+/HER2- High Prolif', nan, 'ER+/HER2- Low Prolif','HER2+'
tumor_size	Tumor size measured by imaging techniques
tumor_stage	Stage of the cancer based on the involvement of surrounding structures, lymph nodes and distant spread

Table 3.1: Data Description

3.2 Statistical Analysis.

The analysis was carried out based on the following steps

1. Exploratory Data Analysis: Numerical variables were summarized using median and interquartile range categorical variables were summarized using frequency and percentages.

2. Variable selection: Variables to be used for further analysis will be selected using an independent T-test for numerical variables, chi-square, and ANOVA for categorical variables.

3. Model evaluation: The machine learning methods for classification were implemented using Python. The libraries used were: NumPy, pandas, and matplotlib among others. Data were split into training and test set percentages of 80% and 20% The class imbalance was taken care of by using upsampling. Methods were compared using evaluation metrics which include Confusion matrix, Accuracy, F1 score, recall score, and precision Score.

4. Determination of breast cancer correlates: Correlates of breast cancer were determined based on the optimal method chosen from step 3.

3.3 Machine Learning Techniques.

3.3.1 Neural Networks

Neural Networks behaves like the human brain and allows computer to recognize patterns. Neural Network consists of three layer the input layer, hidden layer and output layer. Neural Networks will were utilized Multilayer Perceptron, Convolutional Neural Networks and Long Short-Term Memory.

Multilayer Perceptron (MLP)

Is a feed-forward Neural Network. It works by utilizing parallel and advanced structures in the prediction of the target variables by neurons processing through input, hidden and output layers that are linked together through weights (Bazrafshan et al., 2022). Multilayer Perceptron has the characteristic of matrix multiplication, which takes into account every node in the layer. The summation function is used to provide output of the hidden neurons.

$$m_j = \sum_{i=1}^{n_0} w_{ij} x_i + b_j \tag{3.1}$$

$$y_j = f(m_j) = (1 + e^{-z_i})^{-1} \tag{3.2}$$

Where m_j is the input of the j^{th} neuron in the hidden layer, b_j is the bias of the j^{th} neuron in the hidden layer, w_{ij} is the weight value between the i^{th} input neuron and the j^{th} neuron in the hidden layer, y_j is the output of the j^{th} neuron and $f(z_i)$ is the activation function.

Output of the last layer is given as follows:

$$p_k = \sum_{j=i}^{n_1} w_{jk} y_j + b_k \tag{3.3}$$

$p_k$is the output of the neurons in the q^{th} output, w_{jq} is the weight value between the neuron in the j^{th} hidden layer and the neuron in the q^{th} output layer, and $n1$ is the number of neurons in the hidden layers.In as much as MLP mainly utilizes back propagation algorithm to find MLP parameters, it may be limited to local optimums.

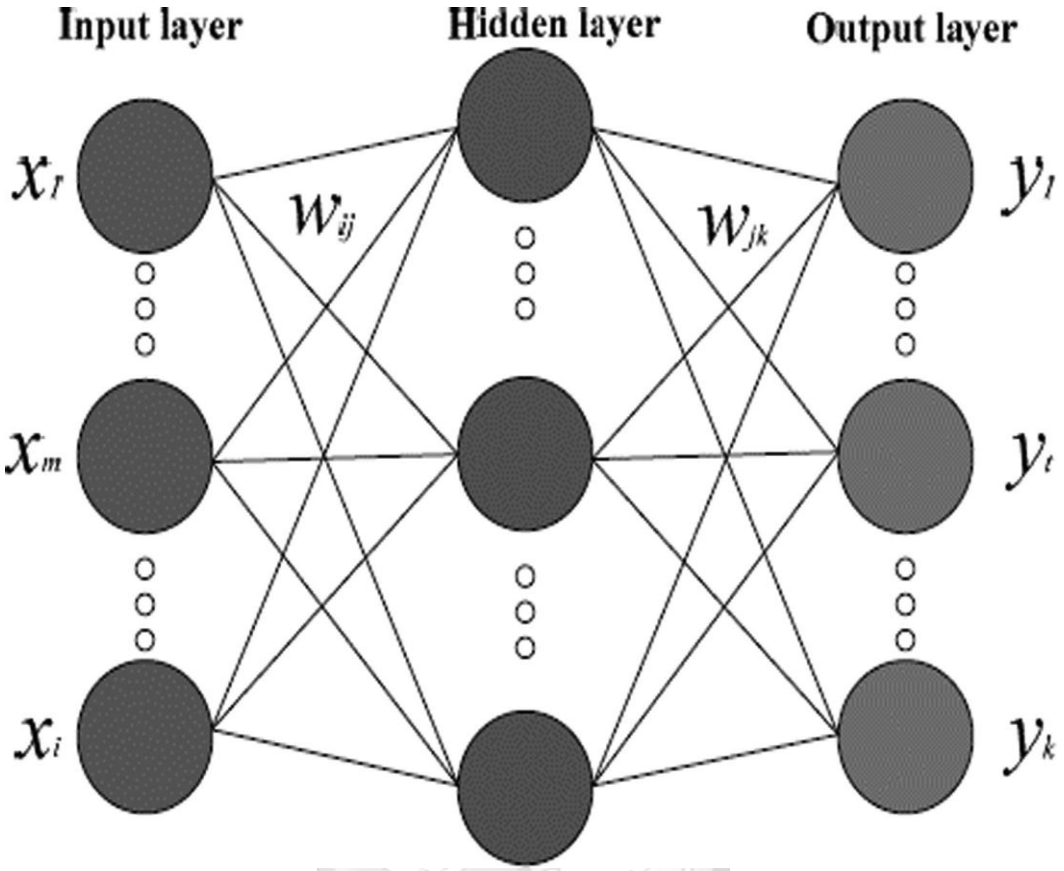


Figure 3.1: Multilayer Perceptron representation with an input layer, a hidden layer and an output layer

Convolutional Neural Networks (CNN)

Is a classification algorithm which takes in an input features, assign important weights and biases to various objects in the variables and differentiate one from the other (Ren et al., 2022).The structure has neurons which increase the low-level features through the learnable weights and deviations.

$$Y_i = \tanh \sum_{n=1}^{p} w_n x_i - n + p + \beta \qquad (3.4)$$

x_i denotes the input features, w_n denotes the weight matrix of the convolution kernel, β indicates the deviation value, p indicates the number of convolution kernels, Y_i shows the output.

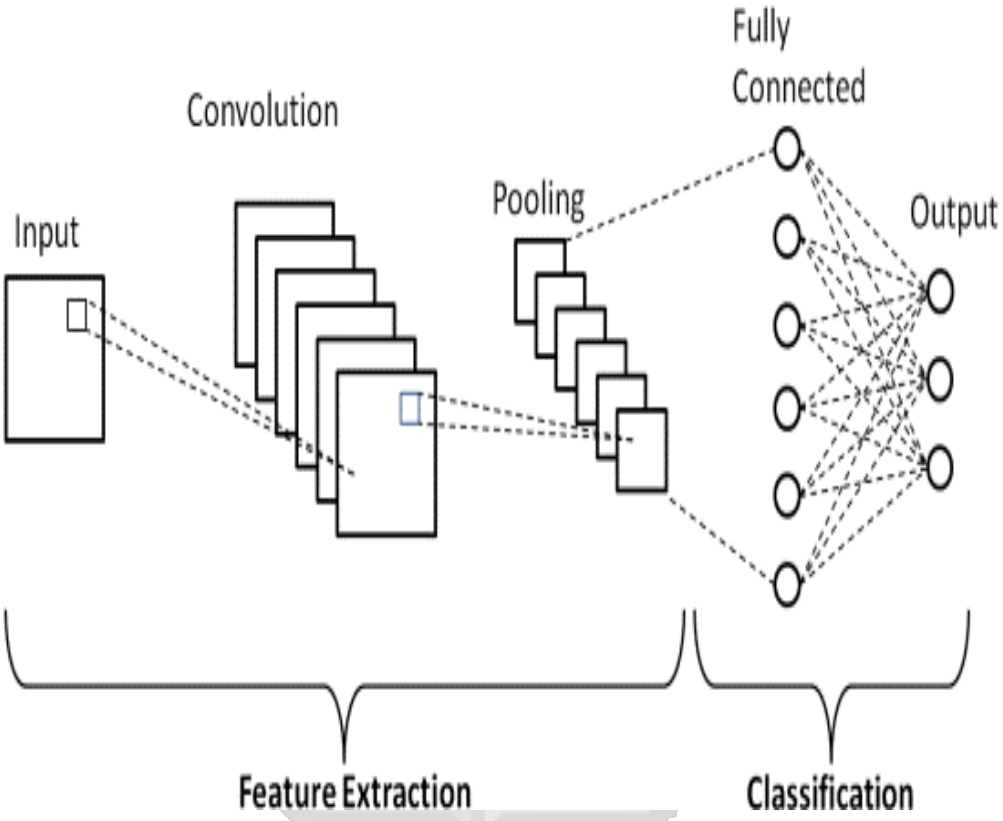


Figure 3.2: Convolutional Neural Networks Structure

Long Short-Term Memory (LSTM)

It is a special kind of recurrent Neural Networks (RNN). RNN has the problem of gradient disappearance and gradient explosion, the LSTM has the ability to solve the problem. LSTM has a strong ability to learn the long-term dependence (Ren et al., 2022). LSTM network behavior is to recall long-term dependencies. It has a cell state which helps transfer the information through the cells with small information changes (Acikmese and Alptekin, 2019).

LSTM structure includes three gates that determine which information will go to the next cell. The gates are forget gate,input gate and output gate. First sigmoid activation function is

the forget gate which selects the information that needs to be forgotten from the previous cell state $(Ct-1)$ and is not needed. The second is the input gate with second sigmoid and tanh activation function which selects values to be updated or saved. The third and last is the output gate with last sigmoid function which determines the information that should be going to the next hidden state and what the output of the cell will be Acikmese and Alptekin (2019). The gates coefficient value ranges from $[0, 1]$.

The structure of LSTM:

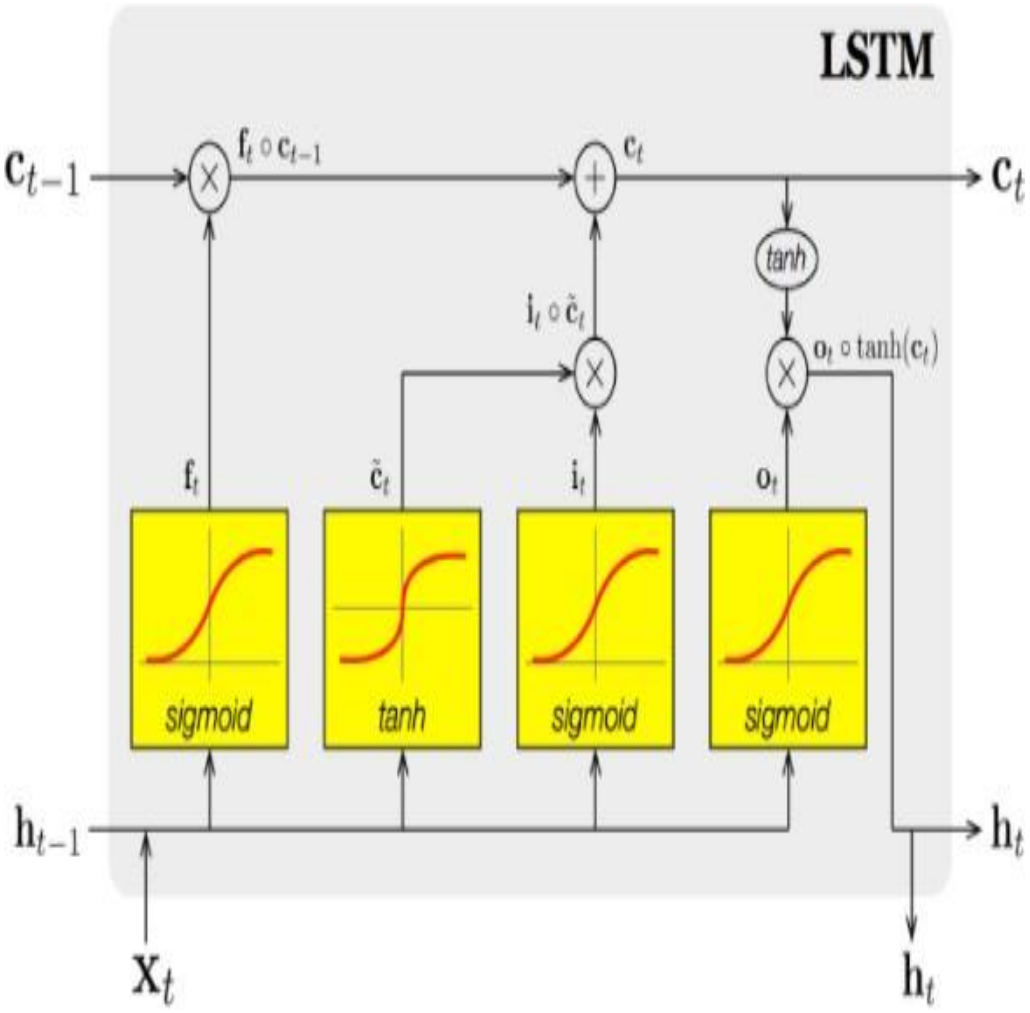


Figure 3.3: LSTM Structure

Cell state equation:

$$c_t = f_t \, o \, c_{t-1} + i_t \, o \, \tilde{c}_t \tag{3.5}$$

Where: c_t is the cell state memory, f_t is the forget gate activation vector, c_{t-1} is the previous state value, i_t is the input gate.

Forget gate equation:

$$f_t = \sigma \; w_f [h_{t-1}, X_t] + b_t \tag{3.6}$$

Where: f_t is the forget gate activation vector, σ is the sigmoid function, w_f weight of respective gate, h_{t-1} output of previous LSTM block, X_t current input, b_t is the bias.

Input gate equation:

$$i_t = \sigma \; w_i [h_{t-1}, X_t] + b_i \tag{3.7}$$

Where: i_t is the input gate, σ is the sigmoid function, w_f weight of respective gate, h_{t-1} output of previous LSTM block, X_t current input, b_i is the bias.

Input modulation gate equation:

$$\tilde{c}_t = \tanh \; w_c [h_{t-1}, X_t] + b_c \tag{3.8}$$

Where: $\tilde{c}_t$ is the current neuron candidate, w_c weight of respective gate, h_{t-1} output of previous LSTM block, X_t current input, b_c is the bias.

Output Gate Equation:

$$o_t = \sigma \; w_o [h_{t-1}, X_t] + b_o \tag{3.9}$$

Where: o_t is the output gate, σ is the sigmoid function, w_o weight of respective gate, h_{t-1} output of previous LSTM block, X_t current input, b_o is the bias.

3.3.2 Tree-based Models

Tree-based ensemble classifiers was considered for the study. Classifiers are used because the response variable was categorical with two classes survival or death. The ensemble method is a Machine Learning technique where learning models combine multiple learners to achieve the best performance. Ensemble methods reduce bias and variance by the use of bagging and boosting techniques. Bagging decreases variance in the data by generating additional data for training from the data set in use using combinations that have repetitions. Boosting on the other hand builds strong predictive models. A base model depends on previous ones. The tree-based ensemble models that will be utilized are random forest, extra tree classifier, gradient boost, AdaBoost, and XGBoost.

Random Forest

Random forest is a supervised Machine Learning algorithm. It contains a large number of individual decision trees trained using bagging method. Each tree has a class of prediction. Random forest will utilize an ensemble of decision trees to predict breast cancer survival. Random forest allows each individual tree to sample with replacement from the data set. Random forest deals with classification problems,by determining individual tree forecasts and taking the response category which occur most frequently in the similar terminal node as the test case being considered (Sage et al., 2020). The response variable is calculated by averaging the prediction of the individual decision trees.

$$Y = \frac{1}{N}\sum_{n=1}^{N} \varphi_n(x_i) \tag{3.10}$$

Y is the output, N represents the number of base learners, x_i input variables, φ_n is the output of the decision tree trained on a bootstrap of the data with a random subsample of the variables (Mittendorf et al., 2022).

Extra Tree classifier

As suggested by Geurts et al. (2006) Extra tree is developed from random forest. Extra tree classifier is an ensemble method which collects de-correlated trees collected in a forest to give classification results. It is similar to Random forest but constructed in a different manner. It splits the given nodes by selecting cut-points at random and then uses the training data and not bootstrap to grow the trees. It ensures no overfitting and gives better accuracy. It decreases bias by use of the original training sample and not bootstrap samples. One major strength is computational efficiency (Ampomah et al., 2020).

$$g(x, \beta_1, \beta_2 \ldots\ldots \beta_k) = \frac{1}{K} \sum_{k=1}^{K} g(x, \beta k) \tag{3.11}$$

Where $g(x, \beta_k)$ indicates the g^{th} predicting tree where β denotes uniform independent distribution vector dispensed before the growth of the tree, x is the input variable and K is the number of variables. All the trees are aggregated (Hammed et al., 2021).

AdaBoost

AdaBoost is an ensemble method also called adaptive boosting used to increase the efficiency of binary classifiers. It uses a repetitive approach to learn from weak classifiers and turn them to strong classifiers. The algorithm repeats by applying the initial base classifier on the training set with a new weight. Finally the classification model obtained is a linear combination of the models obtained from the iterations done (Chengsheng et al., 2017).

$$H_x = sign\left(\sum_{t=1}^{T} \alpha_t . h_t(x)\right) \tag{3.12}$$

Where: α_t denotes trust level, k is the size of training data, t is successive iteration, h_t is the weak classifier selected for the iteration , x is input features while H_x is the output.

Gradient Boost

Gradient boost is a Machine Learning method for regression and classification. Gradient boost assigns weight to the data. Same to Adaboost it combines weak trees and each model tries to predict errors from the previous model. The trees are connected and each strives to reduce the error. They are slow to learn but very accurate.The final model presented is a result of each step and a strong model is realized. Gradient boost uses a loss function to get final output and ensure loss is minimized. The loss function optimization is done using gradient descent. The gradient boosting is taken to calculate the classification value by training a model f via a least-squares regression. Adding an estimator further improves the model in a forward stage-wise strategy:

$$f_t(x) = f_t - 1(x) + Y_t z_t(x) \tag{3.13}$$

where f_t denotes gradient boosting classifier consisting of t decision trees, t represents the total number of decision trees Y_t denotes the learning rate $z_t(x)$ denotes the weak learners. A new decision tree is added to the original gradient boosting model for each boosting iteration (Chen et al., 2020).

Extreme Gradient Boosting(Xgboost) Xgboost is also known as extreme gradient boosting. It improves efficiency and performance. The algorithm minimizes errors of previous models to produce a new model(Chen and Guestrin, 2016). XGBoost is an implementation

of gradient boosted decision trees and used for supervised learning where input features X_i are used to predict response variable Y. The projected score of each tree is summed up to obtain the final score, which is evaluated by using M additive functions to predict the final outcome. This can be indicated as follows:

$$Y = \sum_{c=1}^{M} f_c(X_i),\ f_c \varepsilon\ f_c in \mathsf{R} \tag{3.14}$$

Where M is the number of classification trees, R is the space of classification trees and f_c is a function in the functional space R (Islam et al., 2022).

3.3.3 Evaluation Metrics

Metrics plays an imperative role in optimizing and quantifying Machine Learning model performance. In this study, the evaluation metrics was used to evaluate the performance of the tree based ensemble models. These include: accuracy score, Precision score, area under ROC curve,recall score and F1 Score.

Confusion matrix

Confusion matrix is a contingency table of actual class compared to model predictions.

- True Positive (TP): Is when predicted values are positive and turns out to be true. For instance, the number of cases correctly identified that breast cancer patients survive.
- False positive (FP) is when values predicted as positive and turns out to be false. For instance, the number of cases incorrectly identified that breast cancer patients will survive.
- False Negative (FN) is when values predicted as negative and turns out to be false. This is the number of cases incorrectly identified that breast cancer patients will not survive.

- True Negative (TN): is when values predicted as negative and turns out to be negative, the number of cases correctly identified that breast cancer patients will not survive.

Accuracy Score Accuracy is the ratio of observations which are correctly predicted to the total observations (Yego et al., 2021).

Accuracy=

$$\frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{3.15}$$

Recall Score Recall or sensitivity is the ratio of correctly predicted positive values to the all values in the true class (Yego et al., 2021).

Recall =

$$\frac{TP}{(TP + FN)} \tag{3.16}$$

Precision Score Precision is the ratio of observations which are correctly predicted positive to the total number of observations predicted as positive (Yego et al., 2021).

Precision =

$$\frac{TP}{(TP + FP)} \tag{3.17}$$

F1 Score $F1$ Score is the weighted average of recall and precision score (Yego et al., 2021).

F1 Score=

$$2 * \frac{(Recall * Precision)}{(Recall + Precision)} \tag{3.18}$$

Chapter 4

Results And Discussion

4.1 Results And Discussion

Exploratory Data Analysis was done in order to analyze the dataset using graphs and tables. This was essential to show the relationship between correlates and response variable (breast cancer survival status). After exploring the data, performing data preprocessing and selecting relevant correlates, the dataset was split into training set and test set. For training set 80algorithms. The training set provide a biased sense of model efficacy since actual samples were used to build the model. For the test set 20% of dataset were hold back from training of the model and were used to give an unbiased sagacity of a final model efficacy. The test set was locked away till fine-tuning of the model was complete thereafter an unbiased evaluation of the final hypothesis was obtained (Yego et al., 2021).

4.1.1 Explaratory Data Analysis

Variables	Did Not Survive	Survive
Inferred Menopausal State		
Post	931 (48.9%)	562 (29.5%)
Pre	172 (9.0%)	239 (12.6%)
Type of Breast Surgery		
Breast Conserving	351 (18.6%)	404 (21.5%)
Mastecomy	738 (39.2%)	389 (20.7%)
Pam50 and claudin-low subtype		
Basal	111 (5.8%)	88 (4.6%)
Her2	155 (8.1%)	65 (3.4%)
LumA	364 (19.1%)	315 (16.5%)
LumB	303 (15.9%)	158 (8.3%)
NC	5 (0.3%)	1 (0.1%)
Normal	76 (4%)	64 (3.4%)
claudin-low	89 (4.7%)	110 (5.8%)
3 Gene Classifier Subtype		
ER+/HER2- High Prolif	391 (23%)	212 (12.5%)
ER+/HER2- Low Prolif	317 (18.6%)	302 (17.8%)
ER-/HER2-	146 (8.6%)	144 (8.5%)
HER2+	114 (6.6%)	74 (4.4%)
Tumor stage		
Zero	1 (0.1%)	3 (0.2%)
One	215 (15.3%)	260 (18.5%)
Two	482 (34.4%)	318 (22.7%)
Three	86 (6.1%)	29 (2.1%)
Four	8 (0.6%)	1 (0.1%)
Radio therapy		
Yes	465 (35.5%)	408 (31.2%)
No	277 (21.2%)	159 (12.1%)
Chemotherapy		
Yes	212 (11.1%)	184 (9.7%)
No	891 (46.8%)	617 (32.4%)
Hormone therapy		
Yes	694 (36.4%)	480 (25.2%)
No	409 (21.5%)	321 (16.9%)
Average and median and Interquartile Range for continuous Variables		
Variables	Central Tendencies	Did not Survived
age at diagnosis	Mean	64.4
	Median	66.4
	Interquartile Range	17.9
mutation count	Mean	6
	Median	5
	Interquartile Range	4
lymph nodes examined positive	Mean	2.6
	Median	1
	Interquartile Range	3
tumor size	Mean	28.4
	Median	25
	Interquartile Range	2
nottingham prognostic index	Mean	4.2
	Median	4.1
	Interquartile Range	2

Table 4.1: Frequency distribution of all input features used in the analysis and their relationship with survival status

Table 4.1 above depicts the relationship between survival status and other correlates. A summary is done between the two classes of patients, those who survived and patients who did not survive.

We can see that a number of the patients who are in a premenopausal state had a higher chance of survival than those who were in a postmenopausal state. There is less chance for patients postmenopausal to survive. This implies breast cancer women who had attained menopause had a lower chance of survival.

The number of patients who survive after undergoing breast-conserving surgery was more compared to those who did not survive after undergoing the same surgery. Also, most patients who underwent mastectomy did not survive. This depicts that there was a higher chance

of survival after undergoing breast-conserving surgery and less chance of survival after undergoing mastectomy.

We can also see a higher percentage of patients who did not survive on the variable palm50 were in LumA(Luminal A), lumB(Luminal B), and Basal compared to those who survived in the same group. However, patients in claudin-low had more survivors compared to those who did not survive in the same group.

3 gene classififer subtype shows that most of the patients did not survive in all groups. This shows that breast cancer patients with the three receptor cells had a low chance of survival.

The tumor stage shows that the number of patients decreases as the patients move from one class to another. E.g Patients at stage zero and stage 1 had more survivors. On the other hand stages two, three, and four had a small number of patients who survived as compared to those who did not survive.

Table 4.1 also shows that the number of patients who did not survive after undergoing chemotherapy, hormonal therapy, and radiotherapy was more compared to those who survive after undergoing the same therapy. This depicts that there was a low chance of survival after undergoing chemotherapy, hormonal therapy, and radiotherapy.

4.1.2 Variable selection

(Chi-square)

The first step of selecting correlates that contributes to breast cancer survival was done using the filter method. The variables were filtered to remain with relevant features. The Chi-Square is utilized to determine whether there is some association or relationship between two categorical variables or not. This can be quantified using the p-value approach. When the outcome of the p-value of the chi-square is below 0.05, we conclude that the two values are correlated or rather associated with each other. In a situation where the p-value is greater than 0.05 then the two features that were compared are not correlated or associated. The Chi-square method was used to select important variables since the label of this study was

categorical. This was done by setting a threshold (p-value below 0.05) to eliminate less relevant features. The Chi-square p-value was utilized to determine the extent to which survival status is correlated with other features.

Variables	PValuesOfChi_Square
Inferred menopausal state(Pre)	1.00e-07
Type of breast surgery BREAST CONSERVING	1.90e-06
pam50 + claudin-low subtype Her2	2.13e-05
Type of breast surgery MASTECTOMY	3.57e-05
3-gene classifier subtype ER+/HER2-Low Prolif	7.57e-05
Tumor stage	1.93e-03
Inferred menopausal state(Post)	3.60e-03
3-gene classifier subtype ER+/HER2-High Prolif	1.02e-02
Radiotherapy	4.14e-02

Table 4.2: chis-quare

Table 4.2 above shows the correlation of input features with the response variable (survival status). Inferred menopausal state pre was the most correlated with the target, followed by the type of breast surgery breast-conserving, pam50+ claudin low subtype Her2, type of breast surgery mastectomy, 3 gene classifier subtype ER+/HER2 Low Prolif, Nottingham prognostic index, tumor stage, inferred menopausal state post and so on.

T test

Correlates	DidNotSurvive	Survive	Pvalue
Age at diagnosis	66.4	56.7	0
Mutation count	5.0	5.0	0
Lymph nodes examined positive	1.0	0.0	0
Lumor size	25.0	20.0	0
Mottingham prognostic index	4.1	4.0	0

Table 4.3: T test

The independent t-test was used to determine whether there was a statistically significant difference between means of continuous correlates of those patients who survived and those who did not survive. When the P-value of the T statistics is lower than 0.05 it implies that the variable is a significant predictor of the other variable(the two means are independent).

From our table the P-values of T statistics of the mean for those who survived and those who did not survive for Age at diagnosis, Mutation count, Lymph nodes examined positive, Tumor size and Mottingham prognostic index were 0. Since the P values are less than 0.05, shows that the variables are independent and hence can be used to predict the survival status of the patients.

4.1.3 Machine Learning Model Results

Results on balanced data.

X	Model	Precision.score	Recall.score	F1_score	Accuracy	ROC_AUC
1	Random Forest	0.8727	0.8436	0.8473	0.8523	0. 928
2	Extra Tree Classifier	0.8078	0.8008	0.8025	0.8054	0.926
3	Adaboost	0.7864	0.7811	0.7824	0.7852	0.855
4	Gradient Boosting	0.8336	0.8288	0.8303	0.8322	0.893
5	XGBoost	0.9215	0.9078	0.9112	0.9128	0.939
6	MLP	0. 7444	0. 8375	0. 7882	0. 7584	0.751
7	CNN	0. 8438	0. 8500	0. 8473	0. 8425	0.877
8	LSTM	0. 8764	0.85	0. 8609	0. 9127	0.908

Table 4.4: Balanced Data

Table 4.4 shows the precision, recall, F1 scores, and accuracy for the validation data for the random forest, extra tree classifier, AdaBoost, gradient boost, XGBoost, MLP, CNN, and LSTM on balanced data. From Table2 the model with highest recall score was XGBoost (0.9078), followed by LSTM (0.85), CNN (0.85), random forest (0.8436), MLP (0. 8375), gradient boosting (0.8288), Extra Tree Classifier (0.8008), and model with lowest recall score is AdaBoost (0.7811). The model with highest, F1 score was XGBoost (0.9112), followed by LSTM (0.8609), random forest (0. 8473), CNN (0. 8473), gradient boosting (0.8303), Extra Tree Classifier (0.8025), MLP (0. 7882), and one with lowest F1 score is AdaBoost (0.7824). The accuracy score for each of the models were: XGBoost (0.9128), followed by LSTM (0.9127), random forest (0. 8523), CNN (0. 8425), gradient boosting (0.8322), Extra Tree Classifier (0.8054), AdaBoost (0.7852) , and one with lowest accuracy score is MLP (0.7582).The most robust model on balanced data was XGboost because it had the highest accuracy, precision score, recall score, and F1 score.

Comparison of tree-based ensemble using ROC AUC

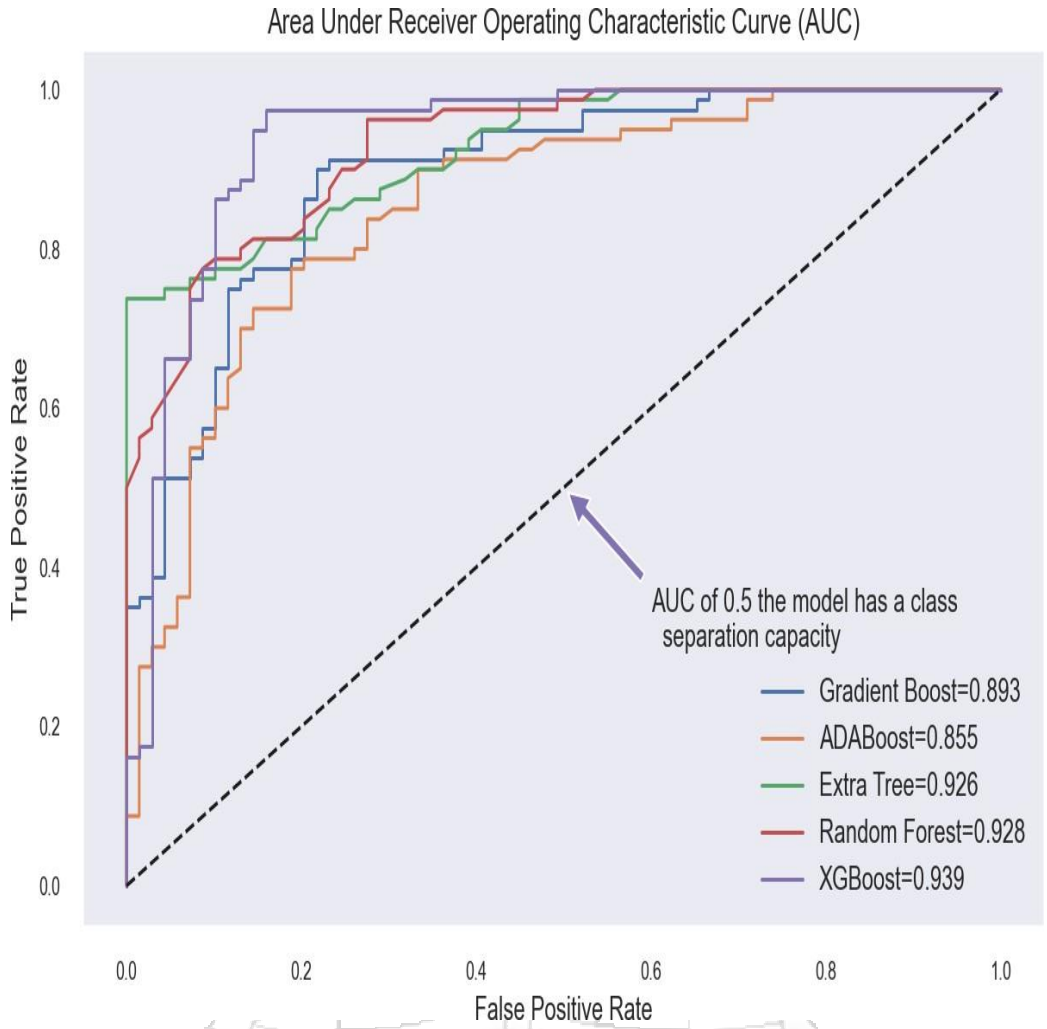


Figure 4.1: Comparing ROC AUC of ML models

Figure 4.1 is ROC AUC which depicts how the Machine Learning models were able to distinguish between the true positives (breast cancer patients that were correctly identified that they will survive) and true negatives (breast cancer patients that were correctly identified that they will not survive) (Kipkogei et al., 2021). The model with the highest ROC AUC was XGBoost (0.939), followed by random forest (0. 928), Extra Tree Classifier (0.926), LSTM (0.908), gradient boosting (0.893), CNN (0. 877), Adaboost (0.855), and one with lowest accuracy score is MLP (0.751). The most robust model with the highest ROC AUC was XGBoost classifier.

ROC AUC results of our XGboost model (0.94) outperform the AUC of the best model (0.87) used by Kaur et al. (2022), the most recent study on breast cancer survival prediction on METABRIC dataset.

Feature Importance

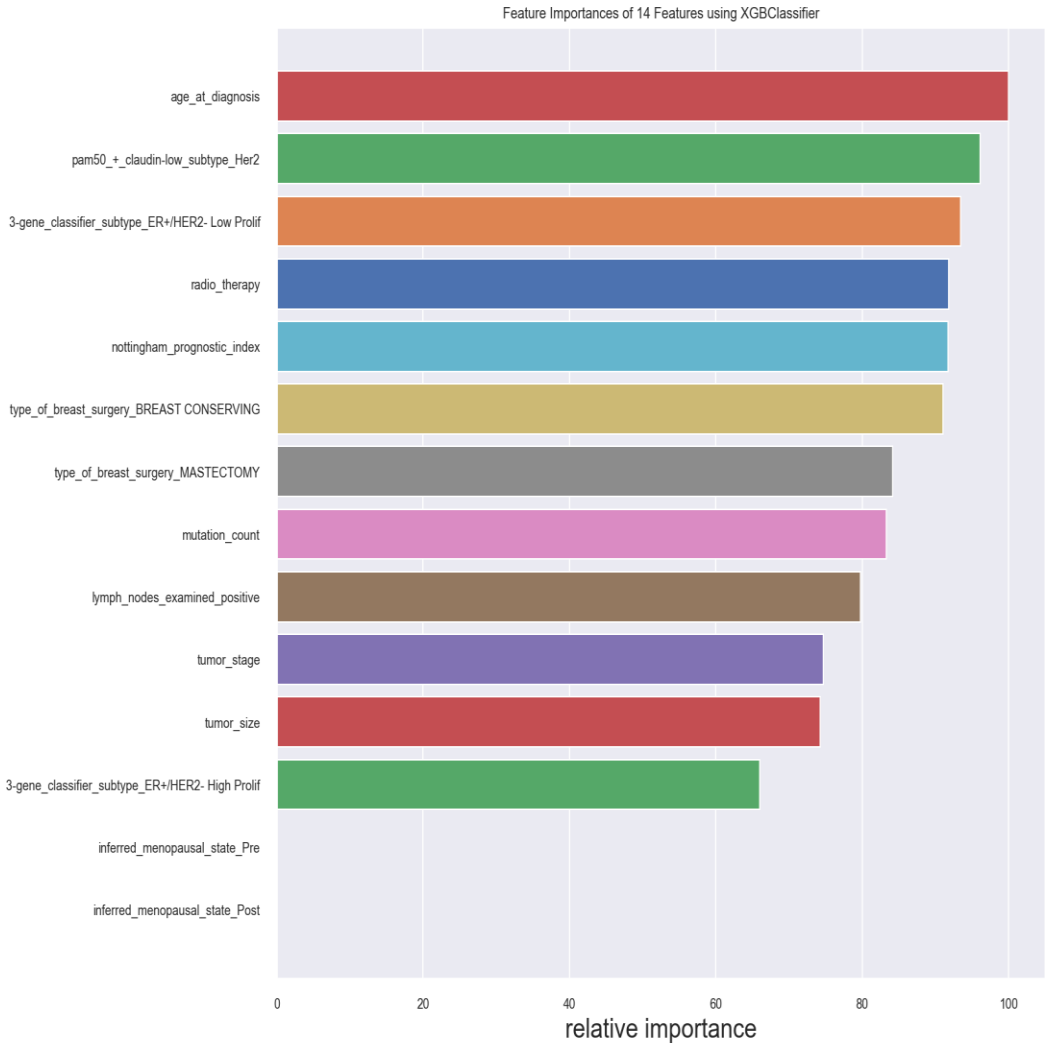


Figure 4.2: Feature Importance

Figure 4.2 depicts the feature importance of 15 variables using XGBoost. After comparing the three Neural Networks and five ensemble tree-based models, the most robust model was used to predict the most important correlates that contribute to breast cancer survival. Since XGBoost had the highest recall score, accuracy score, precision score, ROC AUC, and F1 score, it was, therefore, used to predict important features. The most important features that

contribute to breast cancer survival were; age at diagnosis, pam50+ claudin low subtype her2, 3 gene classifier subtype high, profile,radiotherapy,Nottingham prognostic index,type of breast surgery breast conserving, type of breast surgery mastectomy, mutation count, lymph nodes examined positive, tumor stage, tumor size, 3 gene classifier subtype low profile, pre inferred menopausal state and Post inferred menopausal state

4.1.4 Discusion

In this study, random forest, extra tree classifier, AdaBoost, gradient boosting, XGBoost, MLP, CNN and LSTM were compared as to their respective performance in predicting breast cancer survival. The most optimal classifier was selected as the one with the highest recall score, ROC AUC, F1 score, and accuracy. XGBoost was the most robust classifier based on results from Table 4.4 and Figure 4.1. However, From Table 4,4 results it was evident that performance classifiers improved after balancing the data (upsampling the dataset). XGBoost, LSTM and random forest exhibited outstanding performance on balanced data. Since evaluation metrics of the eight models that were used in this study improved after balancing the data, the results were reported based on the upsampled dataset. Therefore, the most optimal model in Table 4 was used to predict breast cancer survival. On Table 4.4, robust model on balanced data was XGboost because it had the highest accuracy, ROC AUC, precision score, recall score, and F1 score. It was followed by an LSTM, random forest, CNN, gradient boosting, extra tree classifier, AdaBoost, and lastly MLP. Since the study was focusing on predicting breast cancer survival, recall score is a paramount metric since it computes the ratio of breast cancer patients who were correctly identified that will survive to all patients in true class. The ROC AUC is also an essential metric since it depicts how the models were able to distinguish between the true positives (breast cancer patients that were correctly identified that they will survive) and true negatives (breast cancer patients that were correctly identified that they will not survive) (Kipkogei et al., 2021). The model whose ROC AUC was closest to the upper left corner, had highest the recall rate. The model with the highest ROC AUC had the least classification errors. Based on the ROC AUC we concluded that the most optimal model for predicting breast cancer survival is XGBoost.

In Figure 4.2 The most important variable contributing to breast cancer survival was age at diagnosis. From Table 4.3 we could see that the average age of victims who did not survive is 66.4 while those who survived is 56.7. This depicts that aged breast cancer victims had lower chances of survival compared to younger patients.Hence chances of surviving from breast cancer decreases with age. From Table 4.1 we could see that 35% of patients who underwent radiotherapy did not survive while 31% of the patients survived after undergoing radiotherapy. This depicts that there is a slightly low chance of survival for patients who undergo radiotherapy. The third most important variable contributing to breast cancer survival is the mastectomy type of breast surgery. From the table, we could see that most of the patients who underwent mastectomy breast surgery did not survive. On other hand most of the patients who underwent breast-conserving surgery survived. Therefore, mastectomy breast surgery may not be more optimal than breast surgery as compared to breast conservation. Onitilo et al. (2015) found out that breast-conserving surgery was superior to mastectomy.

Chapter 5

Conclusion and Recommendations

5.1 Conclusion and Recommendations

In this study, it was found that the most important correlate contributing to breast cancer survival was age at diagnostic. There is a need for future research to predict the survival time of breast cancer patients based on correlates found in this study such as tumor stage, tumor size time of intervention, and types of treatment and surgeries. Radiotherapy was found as a less optimal treatment for some breast cancer patients and more effective for other breast cancer patients who survived. Thus, we recommend future research to find the correlates which could have influenced the survival status of breast patients who underwent radiotherapy. This includes age, tumor stage, and tumor size among others at the time patient was undergoing radiotherapy. The study also found that XGBoost and LSTM depicted an outstanding performance in breast cancer survival prediction. It was found that balancing data was the most appropriate approach to reduce misclassification and improve model performance. Nevertheless, XGBoost was most optimal than others including LSTM, random forest, CNN, gradient boosting, extra tree classifier, AdaBoost, and lastly MLP models in predicting breast cancer survival. Thus, the successful application of XGBoost in predicting breast cancer survival could be an antecedent for tackling a broad class of pattern detection of correlates that contributes to breast cancer survival.

The methods employed in this research serve as an instance that could be automated and applied to other healthcare predictions. The developed XGBoost model can achieve timely, accurate, and reliable prediction and therefore could be utilized to detect correlates of breast cancer survival (such as age at diagnosis, pam50+ claudin low subtype, 3 gene classifier subtype low profile, Nottingham prognostic index, type of breast surgery mastectomy, tumor

size, tumor stage, radiotherapy, among others.), helping medical industry and governments, to make informed decisions on breast cancer prognosis. Future researchers may evaluate the capability of such techniques to identify correlates of breast cancer survival missed by ensemble tree-based and Neural Networks models in this study to translate them into improved breast cancer survival predictions. We also recommend future scholars harness and compare the performance, KNearst Neighbors (KNN), Naïve Bayes, and Gated Recurrent Unit (GRU) models in predicting breast cancer survival. Future research may employ simulation studies to check characteristics of the data and try other survival analysis techniques in detecting correlates of breast cancer.

References

Abreu, P. H., Santos, M. S., Abreu, M. H., Andrade, B., and Silva, D. C. (2016). Predicting breast cancer recurrence using machine learning techniques: a systematic review. *ACM Computing Surveys (CSUR)*, 49(3):1–40.

Acikmese, Y. and Alptekin, S. E. (2019). Prediction of stress levels with lstm and passive mobile sensors. *Procedia Computer Science*, 159:658–667.

Ampomah, E. K., Qin, Z., and Nyame, G. (2020). Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. *Information*, 11(6):332.

Andjelkovic, J., Ljubic, B., Hai, A. A., Stanojevic, M., Pavlovski, M., Diaz, W., and Obradovic, Z. (2022). Sequential machine learning in prediction of common cancers. *Informatics in Medicine Unlocked*, page 100928.

Bazrafshan, O., Ehteram, M., Latif, S. D., Huang, Y. F., Teo, F. Y., Ahmed, A. N., and El-Shafie, A. (2022). Predicting crop yields using a new robust bayesian averaging model based on multiple hybrid anfis and mlp models. *Ain Shams Engineering Journal*, 13(5):101724.

Chen, R.-C., Caraka, R. E., Arnita, N. E. G., Pomalingo, S., Rachman, A., Toharudin, T., Tai, S.-K., and Pardamean, B. (2020). An end to end of scalable tree boosting system. *Sylwan*, 165(1):1–11.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chengsheng, T., Huacheng, L., and Bing, X. (2017). Adaboost typical algorithm and its application research. In *MATEC Web of Conferences*, volume 139, page 00222. EDP Sciences.

Daoud, M. and Mayo, M. (2019). A survey of neural network-based cancer prediction models from microarray data. *Artificial intelligence in medicine*, 97:204–214.

Fatima, N., Liu, L., Hong, S., and Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, 8:150360–150376.

Fu, X., Pereira, R., De Angelis, C., Veeraraghavan, J., Nanda, S., Qin, L., Cataldo, M. L., Sethunath, V., Mehravaran, S., Gutierrez, C., et al. (2019). Foxa1 upregulation promotes enhancer and transcriptional reprogramming in endocrine-resistant breast cancer. *Proceedings of the National Academy of Sciences*, 116(52):26823–26834.

Ganjisaffar, Y., Caruana, R., and Lopes, C. V. (2011). Bagging gradient-boosted trees for high precision, low variance ranking models. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 85–94.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Gupta, P., Sharma, A., and Jindal, R. (2016). Scalable machine-learning algorithms for big data analytics: a comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(6):194–214.

Hammed, M. M., AlOmar, M. K., Khaleel, F., and Al-Ansari, N. (2021). An extra tree regression model for discharge coefficient prediction: Novel, practical applications in the hydraulic sector and future research directions. *Mathematical Problems in Engineering*, 2021.

Islam, M. D., Li, B., Islam, K. S., Ahasan, R., Mia, M. R., and Haque, M. E. (2022). Airbnb rental price modeling based on latent dirichlet allocation and mesf-xgboost composite model. *Machine Learning with Applications*, 7:100208.

Kaur, P., Singh, A., and Chana, I. (2022). Bsense: A parallel bayesian hyperparameter optimized stacked ensemble model for breast cancer survival prediction. *Journal of Computational Science*, page 101570.

Kipkogei, F., Kabano, I. H., Murorunkwere, B. F., and Joseph, N. (2021). Business success prediction in rwanda: a comparison of tree-based models and logistic regression classifiers. *SN Business & Economics*, 1(8):1–19.

Kirubakaran, R., Jia, T. C., and Aris, N. M. (2017). Awareness of breast cancer among surgical patients in a tertiary hospital in malaysia. *Asian Pacific journal of cancer prevention: APJCP*, 18(1):115.

Lou, S.-J., Hou, M.-F., Chang, H.-T., Chiu, C.-C., Lee, H.-H., Yeh, S.-C. J., and Shi, H.-Y. (2020). Machine learning algorithms to predict recurrence within 10 years after breast cancer surgery: A prospective cohort study. *Cancers*, 12(12):3817.

Magboo, V. P. C. and Magboo, M. S. A. (2021). Machine learning classifiers on breast cancer recurrences. *Procedia Computer Science*, 192:2742–2752.

Mittendorf, M., Nielsen, U. D., and Bingham, H. B. (2022). Data-driven prediction of added-wave resistance on ships in oblique waves—a comparison between tree-based ensemble methods and artificial neural networks. *Applied Ocean Research*, 118:102964.

Montazeri, M., Montazeri, M., Montazeri, M., and Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. *Technology and Health Care*, 24(1):31–42.

Omar, A., Bakr, A., and Ibrahim, N. (2020). Female medical students' awareness, attitudes, and knowledge about early detection of breast cancer in syrian private university, syria. *Heliyon*, 6(4):e03819.

Onitilo, A. A., Engel, J. M., Stankowski, R. V., and Doi, S. A. (2015). Survival comparisons for breast conserving surgery and mastectomy revisited: community experience and the role of radiation therapy. *Clinical medicine & research*, 13(2):65–73.

Ren, J., Yu, Z., Gao, G., Yu, G., and Yu, J. (2022). A cnn-lstm-lightgbm based short-term wind power prediction method based on attention mechanism. *Energy Reports*, 8:437–443.

Rivera-Franco, M. M. and Leon-Rodriguez, E. (2018). Delays in breast cancer detection and treatment in developing countries. *Breast cancer: basic and clinical research*, 12:1178223417752677.

Roberto Cesar, M.-O., German, L.-B., Paola Patricia, A.-C., Eugenia, A.-R., Elisa Clementina, O.-M., Jose, C.-O., Marlon Alberto, P.-M., Fabio Enrique, M.-P., and Margarita, R.-V. (2020). Method based on data mining techniques for breast cancer recurrence analysis. In *International Conference on Swarm Intelligence*, pages 584–596. Springer.

Sage, A. J., Genschel, U., and Nettleton, D. (2020). Tree aggregation for random forest class probability estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(2):134–150.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249.

Yego, N. K., Kasozi, J., and Nkurunziza, J. (2021). A comparative analysis of machine learning models for the prediction of insurance uptake in kenya. *Data*, 6(11):116.

Zeng, J. (2017). Forecasting aggregates with disaggregate variables: does boosting help to select the most relevant predictors? *Journal of Forecasting*, 36(1):74–90.

Appendices

Appendix A: Python Codes

```
#Basic libraries
import numpy as np
import pandas as pd
from scipy import stats
# Visualization libraries
import matplotlib.pyplot as plt
import matplotlib
import seaborn as sns
import yellowbrick as yb
from matplotlib.colors import ListedColormap
from yellowbrick.classifier import ROCAUC
from matplotlib_venn import venn3
import matplotlib.patches as mpatches
# Statistics, EDA, metrics libraries
from scipy.stats import normaltest, skew
from sklearn.preprocessing import LabelEncoder, StandardScaler

#Import Data
df = pd.read_csv('METABRIC_RNA_Mutation.csv', delimiter=',')
# create a new dataframe for clinical attributes only
clinical_features_to_drop = df.columns[31:] # non clinical attributes
clinical_df = df.drop(clinical_features_to_drop, axis=1)
clinical_df.head()
# a function that takes a dataframe and transforms it into a standard f
orm after dropping nun_numirical columns
def to_standard (df):

    num_df = df[df.select_dtypes(include = np.number).columns.tolist()]

    ss = StandardScaler()
    std = ss.fit_transform(num_df)

    std_df = pd.DataFrame(std, index = num_df.index, columns = num_df.c
olumns)
    return std_df
```

Python Codes

```
# Visualization

fig, ax = plt.subplots(ncols=3, figsize=(15,3))
fig.suptitle('The Distribution of Continuous Clinical Attributes', font
size = 18)

ax[0].hist(survived['age_at_diagnosis'], alpha=0.9, color=sns.color_pal
ette(color)[5], label='Survived')
ax[0].hist(died['age_at_diagnosis'], alpha=0.9, color=sns.color_palette
(color)[0], label='Died')
#ax[0].legend()

ax[1].hist(survived['tumor_size'], alpha=0.9, color=sns.color_palette(c
olor)[5], label='Survived')
ax[1].hist(died['tumor_size'], alpha=0.9, color=sns.color_palette(color
)[0], label='Did not survive')
#ax[1].legend()

ax[2].hist(survived['lymph_nodes_examined_positive'], alpha=0.9, color=
sns.color_palette(color)[5], label='Survived')
ax[2].hist(died['lymph_nodes_examined_positive'], alpha=0.9, color=sns.
color_palette(color)[0], label='Died')
ax[2].legend()

ax[0].set_xlabel('Age in Year')
ax[0].set_ylabel('Number of patients')
ax[1].set_xlabel('Tumour diameter in (mm)')
ax[1].set_ylabel('')
ax[2].set_xlabel('Number of positive nodes')
ax[2].set_ylabel('')

plt.show()
```

Python Codes

```
# data splitting

X = dff.drop(['overall_survival'], axis=1)

y = dff['overall_survival']

# using stratify for y because we need the distribution of the two clas
ses to be equal in train and test sets.

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3
3, random_state=42, stratify = y)
```

```
# ML Models

clfs = {

'RandomForest': RandomForestClassifier(random_state=2),

 'Extra Tree':ExtraTreesClassifier(random_state=5),

  'AdaBoost':AdaBoostClassifier(),

'Gradient Boosting': GradientBoostingClassifier(),

'XGBoost': XGBClassifier()

}
```

Python Codes

```
 #code block to test all models in clfs and generate a report
models_report = pd.DataFrame(columns = ['Model', 'Precision_score', 'Re
call_score','F1_score','Accuracy'])


for clf, clf_name in zip(clfs.values(), clfs.keys()):
    clf.fit(x_train,y_train)
    y_pred = clf.predict(x_test)
    y_pred1 = clf.predict_proba(x_test)[:, 1]
    y_score = clf.score(x_test,y_test)


    #print('Calculating {}'.format(clf_name))
    t = pd.Series({
                   'Model': clf_name,
                   'Precision_score': metrics.precision_score(y_test,
 y_pred,average='macro'),
                   'Recall_score': metrics.recall_score(y_test, y_pre
d,average='macro'),
                   'F1_score': metrics.f1_score(y_test, y_pred,averag
e='macro'),
#'log_loss':metrics.log_loss(y_test, y_pred1),
#'Brier': metrics.brier_score_loss(y_test, y_pred1),
#'ROC AUC':metrics.roc_auc_score(y_test, y_pred),


'Accuracy': metrics.accuracy_score(y_test, y_pred)})
    models_report = models_report.append(t, ignore_index = True)
models_report



#Balancing Using Smote
from imblearn.over_sampling import SMOTE
import imblearn
sm=SMOTE()
Xs, Ys=sm.fit_resample(X,y)
```

Python Codes

```
x_train,x_test,y_train,y_test=train_test_split(Xs, Ys,test_size=0.15, r
andom_state=300)

 #code block to test all models in clfs and generate a report

models_report = pd.DataFrame(columns = ['Model', 'Precision_score', 'Re
call_score','F1_score','Accuracy'])


for clf, clf_name in zip(clfs.values(), clfs.keys()):

    clf.fit(x_train,y_train)

    y_pred = clf.predict(x_test)

    y_pred1 = clf.predict_proba(x_test)[:, 1]

    y_score = clf.score(x_test,y_test)


    #print('Calculating {}'.format(clf_name))

    t = pd.Series({

                    'Model': clf_name,

                    'Precision_score': metrics.precision_score(y_test,
 y_pred,average='macro'),

                    'Recall_score': metrics.recall_score(y_test, y_pre
d,average='macro'),

                    'F1_score': metrics.f1_score(y_test, y_pred,averag
e='macro'),

#'log_loss':metrics.log_loss(y_test, y_pred1),

#'Brier': metrics.brier_score_loss(y_test, y_pred1),

#'ROC AUC':metrics.roc_auc_score(y_test, y_pred),

                    'Accuracy': metrics.accuracy_score(y_test, y_pred)
} )

 models_report = models_report.append(t, ignore_index = True)

models_report


#Using Neural Networks

from sklearn.neural_network import MLPClassifier

mlp = MLPClassifier(random_state=1, max_iter=300).fit(x_train, y_train)

mpred=mlp.predict(x_test)
```

Python Codes

```
# accuracy: (tp + tn) / (p + n)
accuracy = metrics.accuracy_score(y_test, mpred)
print('Accuracy: %f' % accuracy)
# precision tp / (tp + fp)
precision = metrics.precision_score(y_test, mpred)
print('Precision: %f' % precision)
# recall: tp / (tp + fn)
recall = metrics.recall_score(y_test, mpred)
print('Recall: %f' % recall)
# f1: 2 tp / (2 tp + fp + fn)
f1 = metrics.f1_score(y_test, mpred)
print('F1 score: %f' % f1)
auc = metrics.roc_auc_score(y_test, mpred)
print('ROC AUC: %f' % auc)
from imblearn.over_sampling import SMOTE
import imblearn
sm=SMOTE()
Xs, Ys=sm.fit_resample(X,y)
x_train, x_test, y_train, y_test = train_test_split(x_up,y_up, test_siz
e=0.1, random_state=42)


# import model
#using CNN
model=Sequential()
# layers
model.add(Conv1D(filters=32, kernel_size=2, activation='relu', input_sh
ape=X_train[0].shape))
model.add(BatchNormalization())
model.add(Dropout(0.2))
model.add(Conv1D(filters=64, kernel_size=2, activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.2))
```

Python Codes

```
# Adding neural networks

model.add(Flatten())

model.add(Dense(64, activation='relu'))

model.add(Dropout(0.5))

model.add(Dense(1, activation='sigmoid'))


# compile model

model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['a
ccuracy'])

history=model.fit(X_train,y_train,epochs=20,validation_data=(X_test,y_t
est))

pred=model.evaluate(X_test,y_test, verbose=5)

pred=model.evaluate(X_test,y_test, verbose=5)

# predict probabilities for test set

yhat_probs = model.predict(X_test, verbose=0)

# predict crisp classes for test set

yhat_classes = (model.predict(X_test) > 0.5)*1

# accuracy: (tp + tn) / (p + n)

accuracy = accuracy_score(y_test, yhat_classes)

print('Accuracy: %f' % accuracy)

# precision tp / (tp + fp)

precision = precision_score(y_test, yhat_classes)

print('Precision: %f' % precision)

# recall: tp / (tp + fn)

recall = recall_score(y_test, yhat_classes)

print('Recall: %f' % recall)

# f1: 2 tp / (2 tp + fp + fn)

f1 = f1_score(y_test, yhat_classes)

print('F1 score: %f' % f1)

auc = roc_auc_score(y_test, yhat_probs)

print('ROC AUC: %f' % auc)
```

Python Codes

```
#Using LSTM

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import Dense,Flatten,Conv1D,BatchNormaliza
tion,Dropout

from keras.models import Sequential

from keras.layers import Dense

from keras.layers import Dropout

from keras.layers import LSTM


# summary

model.summary()

model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['a
ccuracy'])

history=model.fit(X_train,y_train,epochs=20,batch_size=300, validation_
data=(X_test,y_test))

pred=model.evaluate(X_test,y_test, verbose=5)

# predict probabilities for test set

yhat_probs = model.predict(X_test, verbose=0)

# predict crisp classes for test set

yhat_classes = (model.predict(X_test) > 0.5)*1
```

Appendix B: Additional Figures

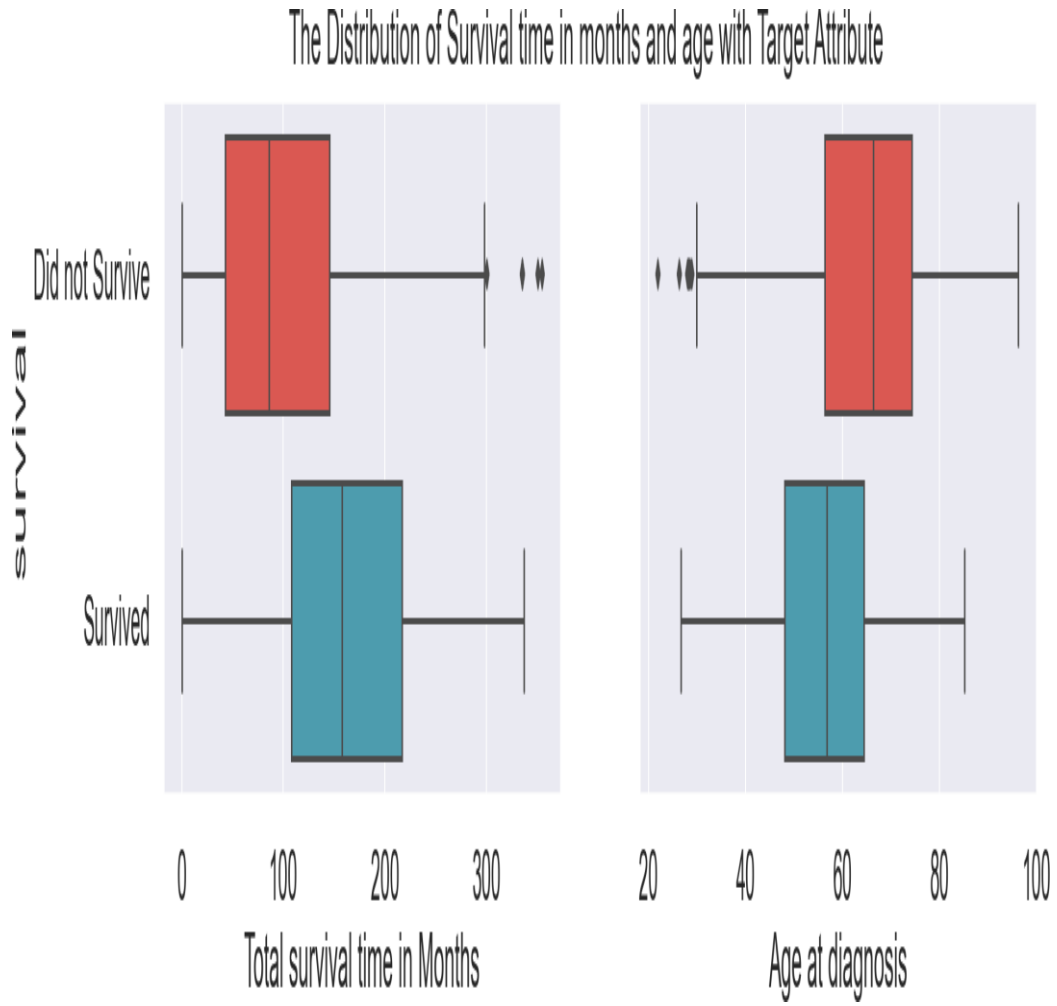


Figure B.1: Total Survival Time and Age at Diagnosis

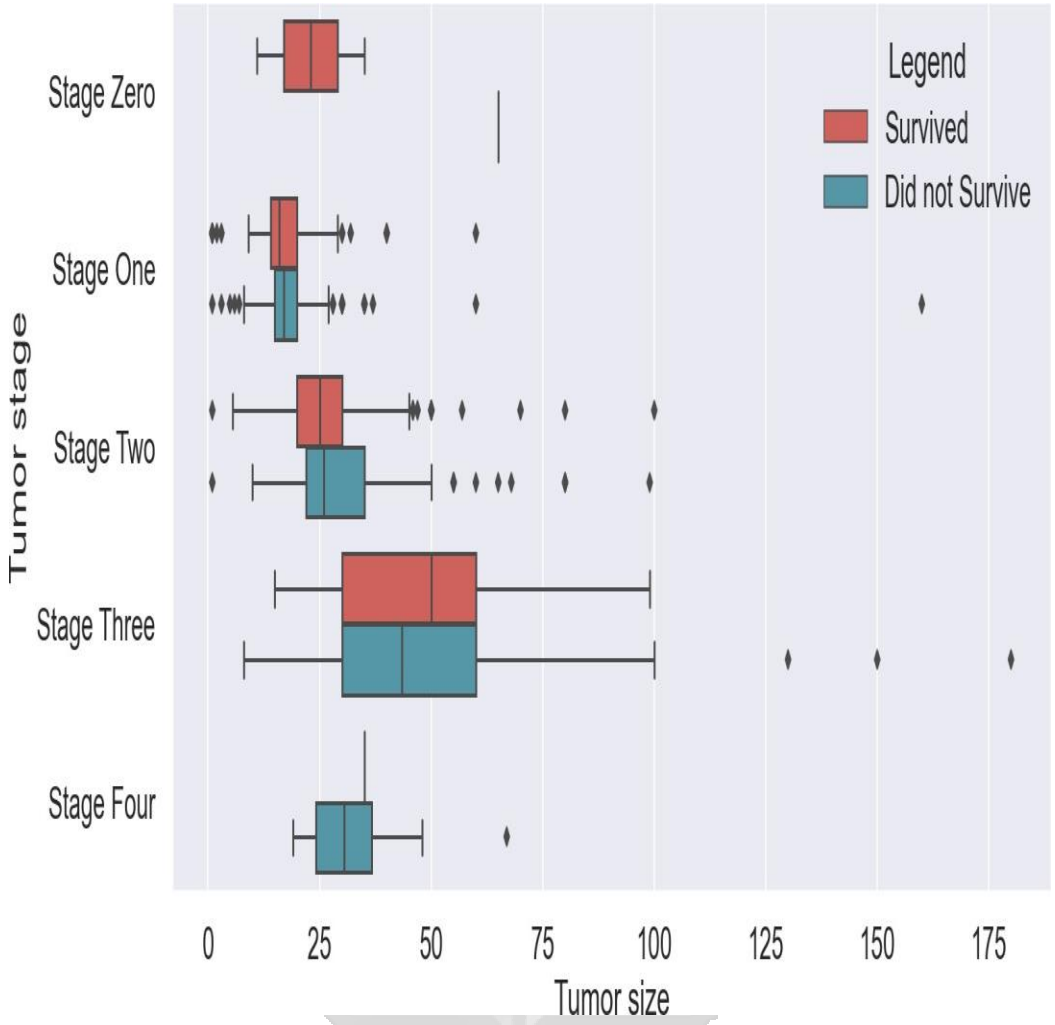


Figure B.2: Breast cancer Stage and Tumor Size by survival status

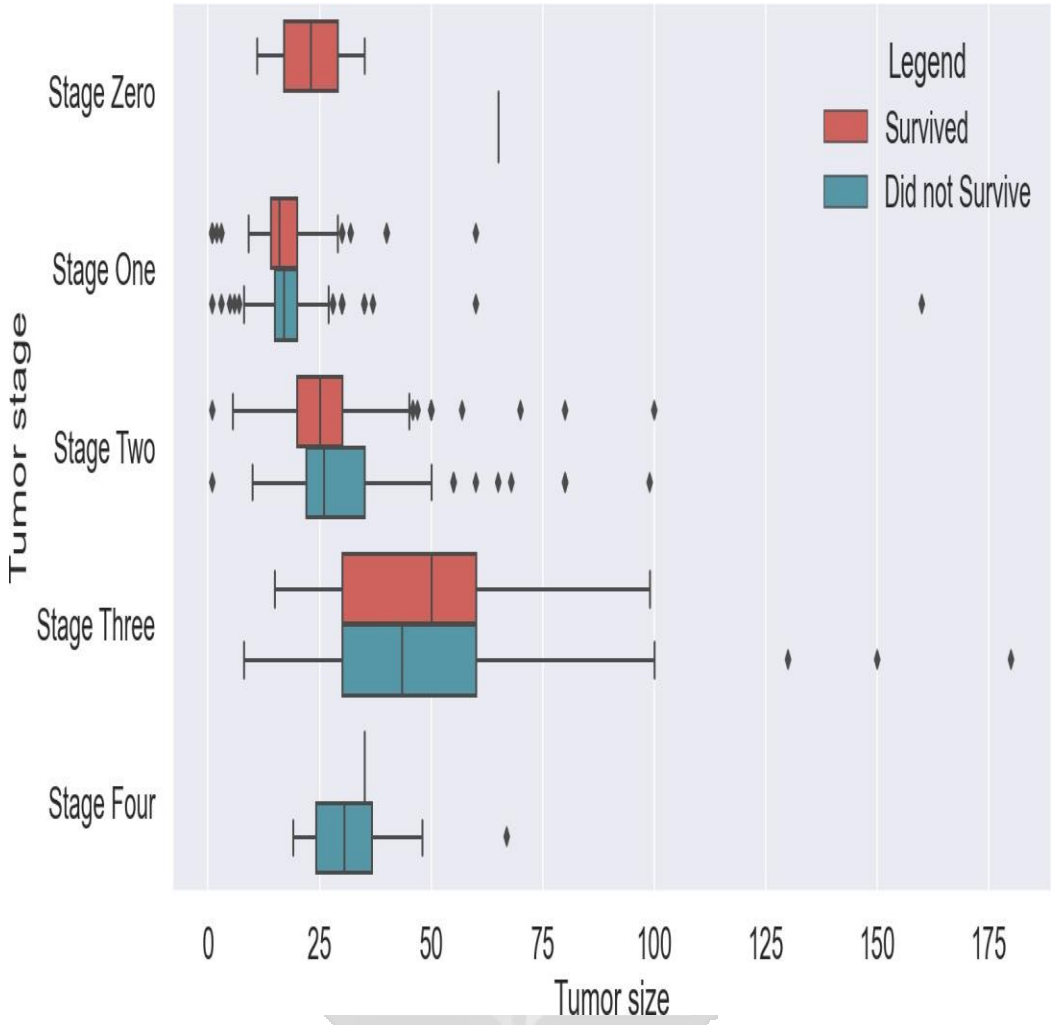


Figure B.3: Age, diameter of tumor and number of positive lymph nodes

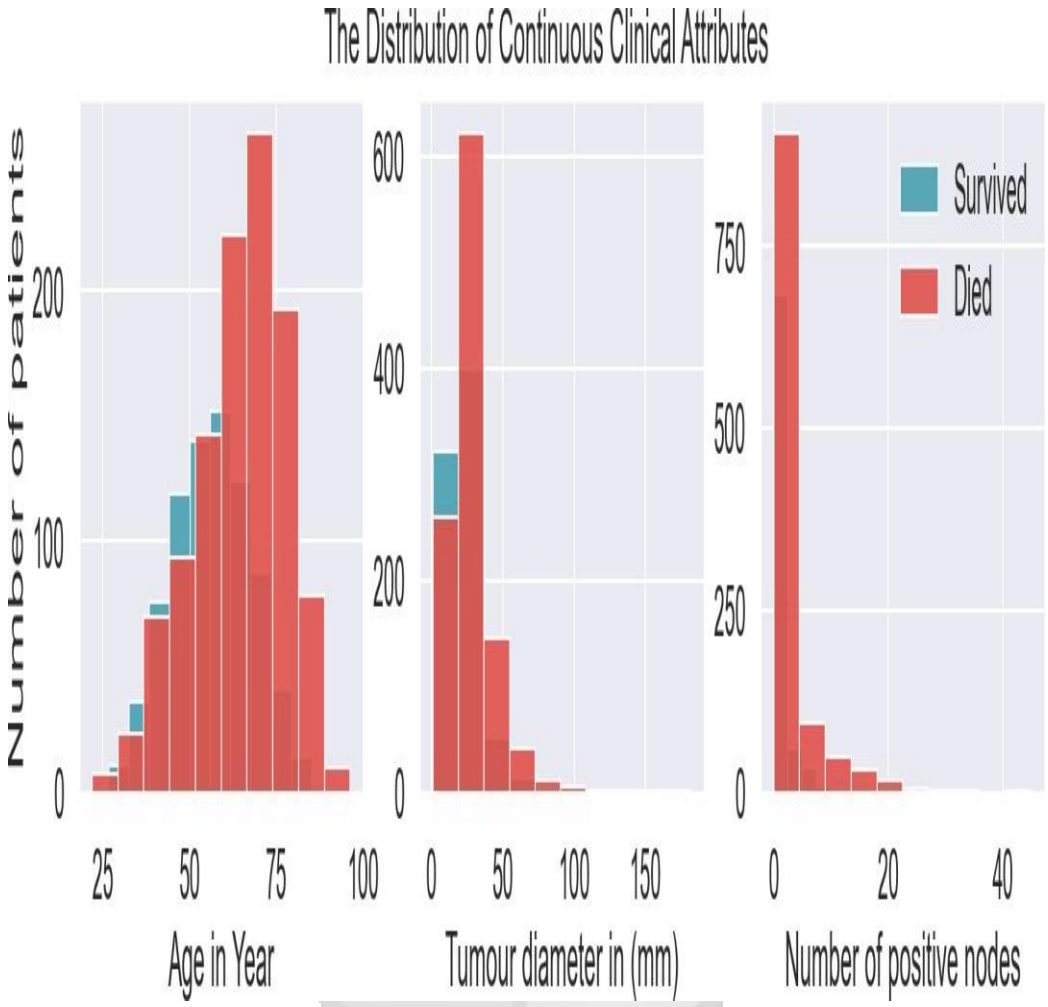


Figure B.4: Age, diameter of tumor and number of positive lymph nodes

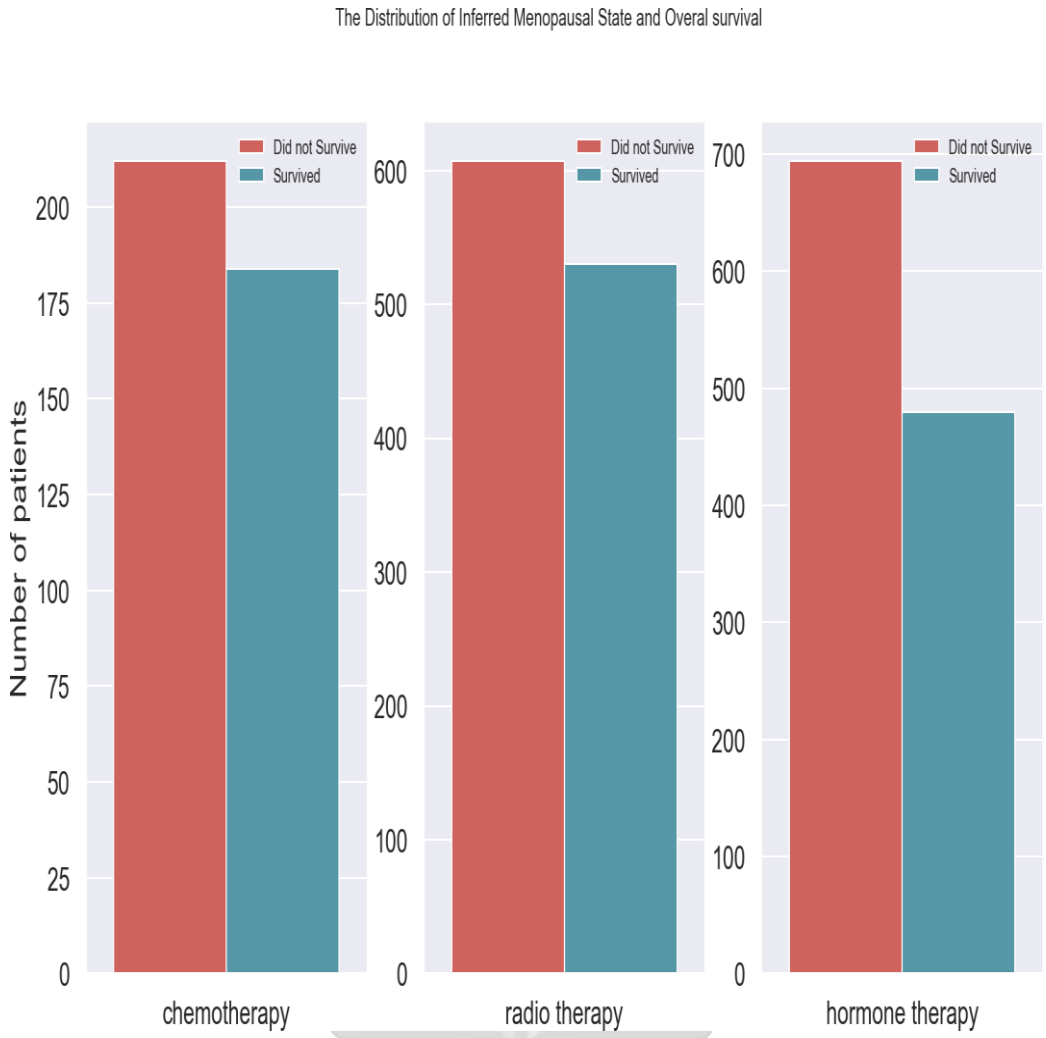


Figure B.5: Chemotherapy,Hormonal therapy, and Radio therapy

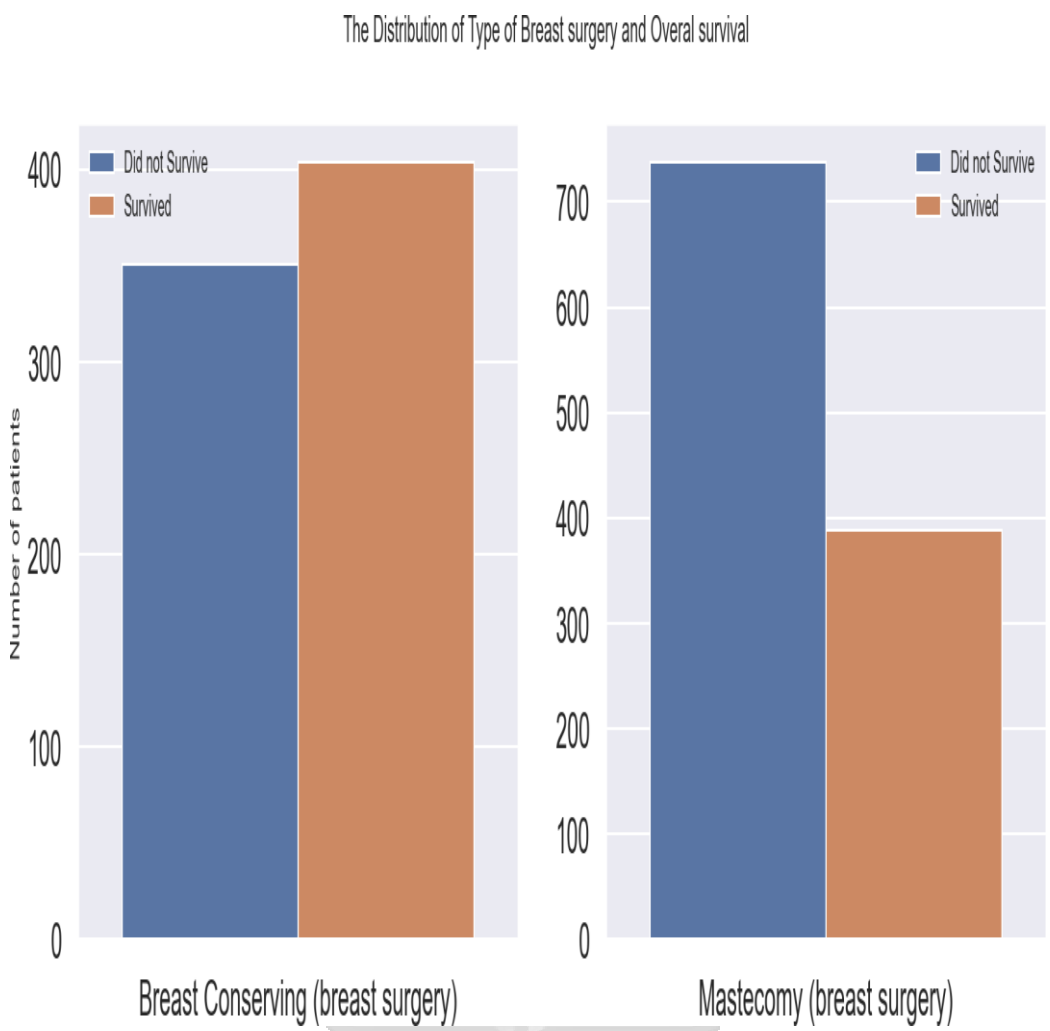


Figure B.6: BC Surgery

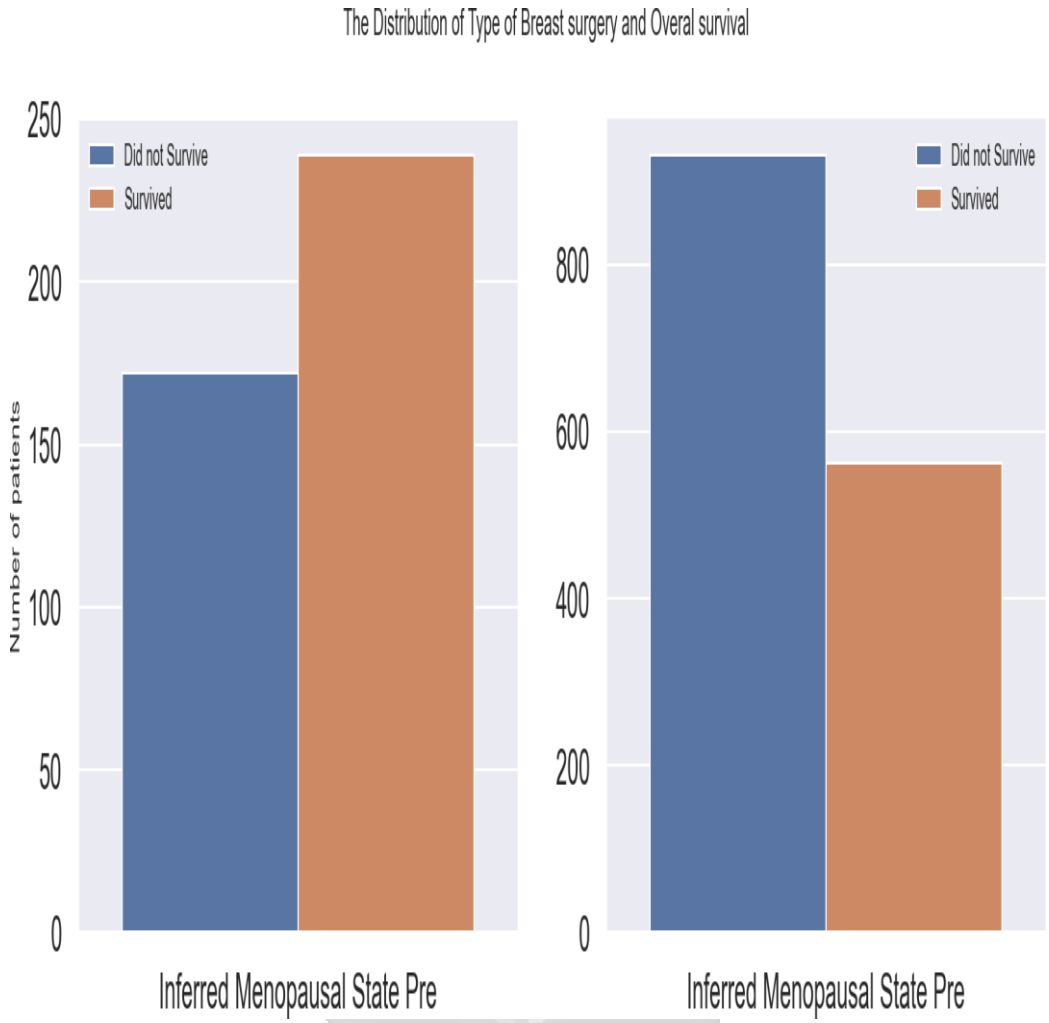


Figure B.7: BC Surgery

Appendix C: Ethical Review Report

6th June 2022

Mrs Jepchirchir, Ruth
ruth.katam@strathmore.edu

Dear Mrs Jepchirchir,

RE: Comparison of Neural Networks and Tree-based Ensemble methods in detecting correlates of breast cancer survival.

This is to inform you that SU-IERC has reviewed and **approved** your above **SU Masters'** research proposal. Your application reference number is **SU-IERC1367/22.** The approval period is **6th June 2022 to 5th June 2023.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

for: **Dr Ben Ngoye,**
Secretary; SU-IERC

Cc: Prof Fred Were,
Chairperson; SU-IERC

Appendix D: Similarity Report

Ouriginal by Turnitin

Document Information

Analyzed document	135214_RuthJepchirchir Research.pdf (D141481336)
Submitted	6/30/2022 1:22:00 AM
Submitted by	
Submitter email	Ruth.Katam@strathmore.edu
Similarity	10%
Analysis address	library.strath@analysis.urkund.com

Sources included in the report

SA	**14354530.pdf** Document 14354530.pdf (D54758629)	1
W	URL: https://chowdera.com/2021/04/20210409080850910k.html Fetched: 6/24/2022 5:01:15 PM	1
W	URL: https://stevenfelix.github.io/kaggle-churn.html Fetched: 6/30/2022 1:23:08 AM	4
W	URL: https://jejjohnson.github.io/rbig_eo/notebooks/spatial_temporal/3.1_exp2_results_prelim_models/ Fetched: 1/5/2022 1:53:14 PM	1
SA	**lab+TM4+week+43.pdf** Document lab+TM4+week+43.pdf (D57685509)	9
SA	**14354206.pdf** Document 14354206.pdf (D54758572)	8
SA	**B9IS107_Data Analytics and Visualisation_10574812.pdf** Document B9IS107_Data Analytics and Visualisation_10574812.pdf (D135763124)	15
SA	**CCMVI2085U_Exam.pdf** Document CCMVI2085U_Exam.pdf (D110846828)	2
SA	**Exam_in_Machine_Learning_for_Predictive_Analytics_in_Business__S119074_.pdf** Document Exam_in_Machine_Learning_for_Predictive_Analytics_in_Business__S119074_.pdf (D119852730)	3
SA	**lab_3.pdf** Document lab_3.pdf (D114450670)	4
W	URL: https://datascience.stackexchange.com/questions/46885/how-to-reshape-xtrain-array-and-what-about-input-shape Fetched: 4/11/2020 1:45:28 PM	3
SA	**14304655.pdf** Document 14304655.pdf (D54491330)	1