Breast Cancer Analysis and Prognosis Using Machine Learning

MSc Research Project
Data Analytics

Simitha Sitaram Shetty
Student ID: X18134980

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Simitha Shetty
Student ID:	X18134980
Programme:	Data Analytics
Year:	2019-2020
Module:	MSc Research Project
Supervisor:	Dr. Muhammad Iqbal
Submission Due Date:	12/12/2019
Project Title:	Breast Cancer Analysis and Prognosis Using Machine Learning
Word Count:	6522
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	25th January 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Breast Cancer Analysis and Prognosis Using Machine Learning

Simitha Shetty
x18134980

Abstract

With increasing number of cases and deaths every year, breast cancer is one of the most common health problem in today's society and prime cause of death in females. This research focuses on predicting breast cancer severity i.e. "benign" or "malignant" at an earlier phase using Machine Learning algorithms so that appropriate treatment can be provided to reduce number of fatalities in women. This study proposes classification models that will help to identify the severity of disease with the use of two datasets containing breast mass and breast tissue samples. The research aims at balancing the dataset first by applying Synthetic Minority Over-Samppling Technique and then compares seven models, Support Vector Machine, Naive Bayes, Logistic Regression, K-Nearest Neighbout, Classification and Regression Tree, Artificial Neural Network and Extreme Gradient Boosting. As the study focuses on the sensitivity of each model i.e. the True Positive Rate, the output shows that in the first dataset, K-Nearest Neighbour performed best with sensitivity of 97.10% whereas in the second dataset, Extreme Gradient Boosting performed better with 97.81% sensitivity. However on analyzing Figure 10, it can be seen that Artificial Neural Network and K-Nearest Neighbour performed good on both the datasets and these models can be used in predicting the breast cancer severity.

1 Introduction

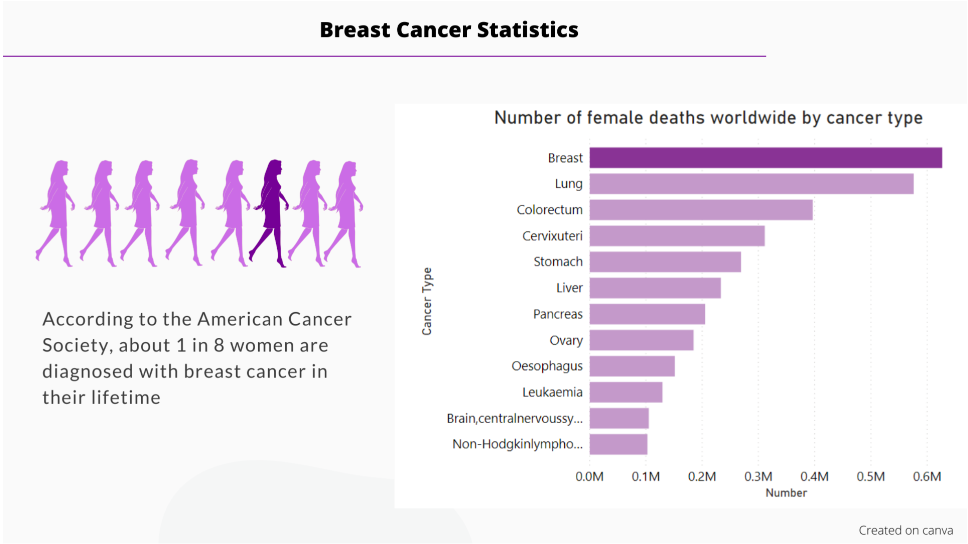
Breast Cancer is one of the most leading type of cancer in females all across the world. Both men and female have the chances of developing the disease, but it is more common in women. Breast cancer develops when the breast tissue spreads in an unusual way making the cells to multiply faster than normal. There are two types of tumors that develop in breast cancer i.e. "Benign" and "Malignant". Benign tumors are the non-cancerous tumors as they do not spread to other body parts whereas Malignant tumors are the cancerous ones Chaurasia et al. (2018). With proper medication benign tumors can be stopped from turning into cancerous tumours at later stages, while malignant tumors just multiply and spread to other parts of body. So, it is important to detect the type of tumor at an early phase and prevent it to turn into cancerous tumors with precise treatment.

There are various factors that can lead to developing breast cancer, some of which are Gender, Heredity, Age i.e. higher the age, more is the chance of developing the disease, High consumption of alcohol and Exposure to harmful radiation. It is not necessary that all these factors definitely lead to cancer, but it is important to know the factors in order to reduce the risk. Breast Cancer cannot be avoided, but with proper medication the chances of curing it can be increased. According to Keles (2019), early detection and treatment might help in effectively treating the disease and reducing the number of deaths in females as it makes various treatment options available to treat cancer. On the other end, early detection also requires proper and accurate results which helps the physicians to segregate benign tumours from the malignant tumours and treat the disease efficiently and improve the survival rate. Nowadays, various machine learning techniques are adapted in the medical domain as they help in finding the hidden patterns in the data. Machine Learning also seems to be appropriate in the medical data due to the complexity in the attribute information available in this field Abed (2018).

Over the time, there are various data that are made available for public use by several breast cancer foundations or cancer hospitals. With the availability of data, the need to analyse them and train the model so as to diagnose the disease to treat it and bring down the death rate has increased substantially. The study carried out in this research aims at identifying a model that will help the practitioners in the medical field in precisely predicting breast cancer at an early stage by classifying the tumor type.

1.1 Motivation

There are different types of cancer and all these types lead to death if not identified at a proper time. As there are awareness for regular check-ups, there needs to be a robust model in place to detect the tumour accurately. If the model fails to detect the tumor, this will too lead to death as the patient will not be aware of it until it is very late.



According to the American Cancer Society, about 1 in 8 women are diagnosed with breast cancer in their lifetime

Created on canva

Figure 1: Cancer Fatalities by Type

Figure 1 shows the number of deaths caused in females in the year 2018 due to cancer by its type. According to the graph, breast cancer is the most dominant disease causing highest number of female deaths [1]. Also, as per the American Cancer Society[2], about every 1 in 8 women has the risk of developing breast cancer in their lifetime. If the cancer

[1] https://www.statista.com/statistics/1031250/number-of-cancer-deaths-females-worldwide-by-type/

[2] https://www.breastcancer.org/symptoms/understand$_bc/statistics$

and its severity is diagnosed at an earlier stage, the number of deaths could be reduced by giving treatments and medication to the patients. This makes it very important to identify the model that accurately predicts the disease so that with appropriate treatment, the risk of fatalities in women can be reduced.

1.2 Research Question

”To what extent can Machine Learning techniques help in accurately diagnosing breast cancer at an early stage?”

1.3 Research Objectives

1. Application of Feature Selection algorithm to select important features.
2. Understanding the nature of the data if the data is balanced and if not, balancing the data to train the model precisely.
3. Implementation of Machine Learning algorithms on balanced datasets and comparison of model performance on these datasets.
4. To train the model to accurately classify the type of tumor as “Benign” or “Malignant”.

2 Related Work

This section discusses various work carried out in the breast cancer field with application of several models and their performance and justifies the research performed by the authors with evidence. The section is further categorized into different objectives that has been tried to achieve by various authors.

2.1 Prediction of Breast Cancer :

Khourdifi and Bahaj (2018) stated that early prediction and detection of breast cancer can help in reducing the chances of death. They used machine learning algorithms like Random Forest, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbour (KNN) in order to predict the disease on an early basis. The objective of the study was to select the best technique to predict the disease. The dataset used for the study is the publicly available data from the University of Wisconsin Hospitals Madison Breast Cancer Database. Later, 10 fold cross validation test was applied to evaluated the classifiers. The output showed that SVM gave the highest accuracy of 97.9%. Graja et al. (2018) used the similar dataset and applied Data Mining Logistic Regression technique with 10-fold cross validation and achieved an accuracy of 99.48% on 70-30 train-test split. Mushtaq et al. (2019) also worked on the Wisconsin dataset which was taken from the UCI Machine Learning Repository. They not only applied supervised classification techniques such as Decision Tree (DT), K-Nearest Neighbour (KNN), Logistic Regression (LR), Naïve Bayes (NB) and Support Vector Machine (SVM) but also combined it with popular PCA dimension reduction techniques such as Linear, Sigmoid, Cosine, Poly and Radial basis function. As per the results, Sigmoid based NB performs best with the highest accuracy of 99.20%. KNN also showed good performance with all PCA techniques. They further suggest using appropriate feature selection methods to improve the performance of the models. Dai et al. (2018) diagnosed breast cancer using the Random Forest algorithm. In this study, ensemble technique was carried where multiple weak classifiers were combined to improve the accuracy of the model and the Wisconsin Breast Cancer (Diagnostic) Dataset from the University of California, Irvine had been used for the same. The results showed good and improved model performance with the accuracy of 99.3% whereas Jivani and Shah (2013) compared three Data Mining Classification techniques such as Decision Tree, Bayesian Network and K-Nearest Neighbour (KNN) on the same dataset where Naïve Bayes outperformed the rest two with 95.99% accuracy and 0.02 seconds of execution time as compared to Decision Tree and KNN.

M and S (2017) applied both the Data Mining algorithms i.e. Naïve Bayes and Decision Tree on the Wisconsin Breast Cancer Dataset and compared the performance of both these techniques where Decision Tree gave not that good but better accuracy than Naïve Bayes i.e. about 76%. Basha and Iyenger (2018) worked on analysing and predicting breast cancer using three different datasets from the UCI Machine Learning Repository. The Machine Learning algorithms applied to do same were Decision Tree, Random Forest (RF) and Support Vector Machine (SVM). The result showed that Random Forest gave best performance with the accuracy of about 98% whereas Mudgil et al. (2019) compared six classification algorithms like Naïve Bayes, K-Nearest Neighbour (KNN), Decision Tree, Logistic Regression, Random Forest, Support Vector Machine (SVM) on Wisconsin Breast Cancer dataset to predict breast cancer. The result proved that KNN performed best with the accuracy of 98.03%. Padmapriya and Velmurugan (2014) worked on analysing breast cancer and discusses the use of the Data Mining Classification algorithms ID3 and C4.5 in the disease. This study mainly focussed on analysing various works and the technique that was useful in predicting breast cancer. As per the study, various algorithms were compared and C4.5 proved to give better and high accuracy and precision rate as compared to other algorithms.

Wang and Yoon (2015) used Dimensionality Reduction and Data Mining techniques to predict Breast Cancer. Principal Component Analysis (PCA) to reduce dimensions and four Data Mining techniques such as Support vector Machine (SVM), Artificial Neural Network (ANN), Naïve Bayes and AdaBoost tree were applied on two breast cancer datasets i.e. Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995). Later, 10-Fold Cross Validation method was implemented to calculate the performance like test error of each of the model. Comparison between the performance of the models with and without PCA were analysed out of which PCA outperforms which means the model gives better accuracy when used along with PCA. Williams et al. (2015) conducted a study to predict breast cancer on Nigerian patients by using Data Mining Classification techniques like Naïve Bayes and J48 Decision Tree and the most effective model was evaluated. The work was carried out on the LASUTH breast cancer dataset which was taken from the Cancer Registry of Lasuth, Nigeria. The results showed that J48 performed better with high accuracy of 94.2% and less error rate as compared to Naïve Bayes which is 82.6% of accuracy.

Rajesh and Anand (2012) diagnosed breast cancer using Data Mining Classification technique on the SEER breast cancer dataset. Data Mining technique C4.5 was used to classify the data into groups “Carcinoma in situ” or “Malignant potential” which gave an accuracy of about 93%. C4.5 was chosen based on the analysis of other models like C-RT, CS-MC4, C4.5, ID3, KNN, LBA, Naïve Bayes, PLS-LDA, RND-TREE and SVM, out of which C4.5 gave less error rate and performed better. Singhal and Pareek (2018) in their study made use of Machine Learning algorithm such as Artificial Neural Network (ANN) using Wisconsin Breast Cancer (Diagnostic) Dataset provided by University of Wisconsin Hospitals. For testing the performance of the model, Feed-Forward and Back-Propagation algorithms were considered and hence gave better performance and proved to be successful in predicting the disease with good accuracy.

Mekha and Teeyasuksaet (2019) compared Deep Learning techniques with activation functions like Tanh, Rectifier, Maxout and Exprectifier with Machine Learning algorithms such as Naïve Bayes (NB), Decision Tree(DT), Support Vector Machine (SVM), Vote (DT+NB+SVM), Random Forest (RF) and Adaboost. The Wisconsin dataset considered for the study was downloaded from UCI Machine Learning Repository. Results showed that Deep Learning with exprectifier function gave high accuracy of 96.99%. The study also stated that Deep Learning helps in automatically selecting features that are important for classification and only selects those features that will give better performance. Delen et al. (2005) worked on SEER Cancer Incidence Public-Use Database for the years 1973-2000. The datafiles to use had been request from the SEER website and Data Mining methods like Artificial Neural Networks (ANN), Decision Trees (DT) and Logistic Regression (LR) along with a 10-Fold Cross Validation were applied to predict the survivability of Breast Cancer patients. Results showed that DT performed best in predicting the survivability with the accuracy of 93.6% followed by ANN and lastly LR with accuracy 91.2% and 89.2% respectively.

2.2 Prediction of Breast Cancer Recurrence :

As per the studies done by Mulatu and Gangarde (2017) on different Data Mining techniques that can effectively help in predicting breast cancer recurrence, Data Mining algorithms like Bayes Net, Support Vector Machine (SVM) and Decision Tree (J48) were considered to be the best in prediction as they gave high performance and good results.

Ojha and Goel (2017) aimed at predicting breast cancer recurrence by using Data Mining techniques on the Wisconsin Prognostic Breast Cancer Dataset from the UCI Machine Learning Repository. The study uses clustering and classification algorithms to compare the performance of these models. Clustering algorithms like K means, Expectation Maximization (EM), Partitioning around Medoids (PAM), Fuzzy c-means and classification algorithms like K Nearest Neighbour (KNN), Support Vector Machine (SVM), Naïve Bayes (NB) and C5.0 were applied. The performance measure used i.e. accuracy, sensitivity and specificity to evaluate the output of the model showed that C5.0 and SVM proved to perform better with the accuracy of around 81% which is more than other models used. Ahmad LG and AR (2013) used the Iranian Centre for Breast Cancer (ICBC) program dataset where the patients were registered from 1997-2008 and predicted breast cancer recurrence by using three Machine Learning techniques i.e. Decision Tree (C4.5), Support Vector Machine (SVM) and Artificial Neural Network (ANN) along with 10-Fold Cross Validation technique to predict unbiased accuracy of every model. All the models were implemented using the WEKA tool and the performance was evaluated based on accuracy, sensitivity and specificity metrics. The results showed that SVM outperformed C4.5 and ANN with the accuracy of about 96%.

2.3 Prediction of Breast Cancer Survivability :

Alhaj and Maghari (2017) predicted breast cancer survivability using two classification techniques I.e. Rule Induction and Random Forest on the Gaza Strip 2011 cancer patient's dataset. RapidMiner tool was used to carry out the implementation process. The results showed that Random Forest performed better with less execution time and better accuracy of 74.6% while Bellaachia and Guven (2006) used SEER Public-Use Dataset to predict breast cancer survivability using Naïve Bayes, back-propagation Neural Network and C4.5 where C4.5 performs better with 87% accuracy.

M.Lundin et al. (1999) in their work predicted the survival chances of breast cancer patients by considering 951 breast cancer patients data including 5-, 10- and 15- cancer specific survival and dividing the dataset into train and test data i.e. 651 and 300 respectively. Artificial Neural Network (ANN) and Logistic Regression (LR) models were applied of which ANN proved to be more accurate in predicting the survival with the accuracy of about 91%. Sadoughi et al. (2015) also predicted breast cancer survival by implementing the Knowledge Discovery in Databases (KDD) methodology using the SEER dataset. Prediction models like Support Vector Machine (SVM), Bayes Net and Chi-Squared Automatic Interaction Detection (CHAID) were used and these models were run in IBM SPSS Modeler 14.2. The results showed that SVM outperformed all other models in breast cancer survival prediction with the accuracy of 96.7%. Zand (2015) also used SEER public-use dataset and aimed at predicting and diagnosing survivability rate of breast cancer patients using Data Mining Techniques. The Data Mining Techniques like Naïve Bayes, Artificial Neural Network (ANN) and C4.5 Decision Tree were implemented and the WEKA tool was used to implement these models. C4.5 gave better accuracy as compared to Naïve Bayes and ANN which is 86.7%.

2.4 Prediction of Breast Cancer Severity :

Vivek Kumar and Abhishek Verma (2019) applied Data Mining techniques to predict Benign and Malignant breast cancer. The Wisconsin dataset used to experiment the study was taken from the UCI Machine Learning Repository and twelve Data Mining techniques like AdaBoost, Decision Table, J-Rip, J48, Lazy IBK, Lazy K-star, Logistic Regression, Multiclass Classifier, Multilayer-Perceptron , Naïve Bayes, Random Forest and Random Tree were implemented. The results showed that Random Tree, Random Forest, Lazy K-star and Lazy IBK performed the best with the accuracy of about 99%. Chaurasia et al. (2018) conducted a study to predict Benign and Malignant breast cancer by using Data Mining algorithms. Three popular Data Mining algorithms like Naïve Bayes, RBF Network and J48 Decision Tree on the Wisconsin Breast Cancer dataset provided by UCI Machine Learning Repository were used to perform the study. 10-Fold Cross Validation technique to measure unbiased accuracy was also applied. The output proved that Naïve Bayes gave best accuracy of 97.36% followed by RBF Network and J48 with the accuracy of 96.77% and 93.41% and the attribute “class” indicated to be the most important predicting factor.

NEMISSI et al. (2018) carried analysis on the Wisconsin Breast Cancer Dataset to diagnose the disease at an early age. They proposed Neural Network system with single hidden layer using several activation functions and trained the model using Extreme Learning Machine (EML) Algorithm. Later, 10-Fold Cross Validation was applied so that the performance can be generalised where 9 of 10 randomly segregated dataset were used for training and remaining as test data. When compared with Conventional Extreme

Learning Network, the proposed method gave better results with less hidden neurons. Hence, this proved to overcome one of the main problem of EML as it has highest hidden neurons. Keles (2019) suggests early prediction of breast cancer can help in treating the disease more effectively. The study aims at using Data Mining Classification techniques on an antenna dataset considering skin layer, fat layer and fibro-glandular layer to predict and diagnose the disease as early as possible. 10-Fold Cross Validation was applied to achieve accurate results. The performance of the algorithms were analysed in the WEKA tool and the classification algorithms like Bagging algorithm, IBk, Random Committee, Random Forest and Simple Classification and Regression Tree (SimpleCART) algorithms gave better performance with the accuracy higher than 90% and Random Forest with highest accuracy of 92.2%.

3 Methodology

When working on any research project, it is very necessary to have a defined methodology to obtain good output. There are several methods that can be followed for a research such as KDD, CRISP-DM, SEMA, etc. For this research, the methodology considered is the Knowledge Discovery in Databases (KDD) methodology. KDD helps in deriving useful knowledge from the data. KDD can be useful in medical study as it helps to seek hidden information and patterns in a dataset with many features Sadoughi et al. (2015).

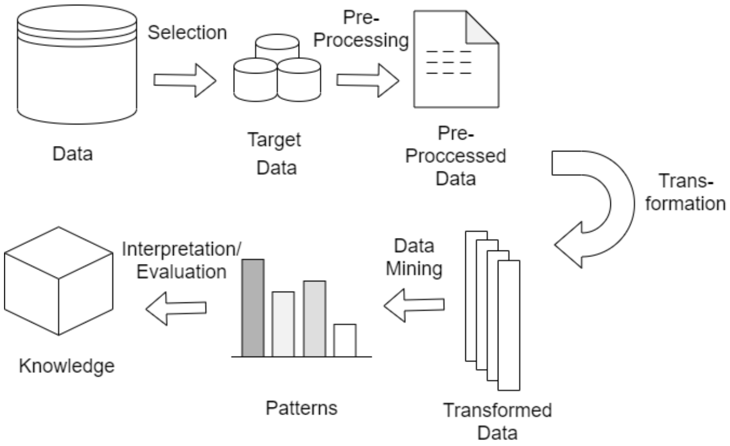


Figure 2: KDD Process

Figure 2 shows the KDD process flow. Following are the steps that the KDD methodology follows :

3.1 Data Selection

There is large amount of data available, of which some subset of data is selected known as the target data. Appropriate data selection is important as it must satisfy the aim of the research. For this research, two datasets i.e.

- Dataset 1 : The Wisconsin Breast Cancer dataset from data.world [3]. The dataset has 570 observations with 32 attributes.It contains information about the breast mass.

[3] https://data.world/health/breast-cancer-wisconsin

Table 1: Attributes of Dataset 1

Code(Attributes)	**Description**	**Domain(Values)**
diagnosis	tumor type	B-Benign or M-Malignant
radius	mean of distances from center	decimal
texture	standard deviation of gray-scale values	decimal
smoothness	local variation in radius lengths	decimal
concavity	severity of concave portions of the contour	decimal
concave points	number of concave portions of the contour	decimal
fractal dimension	design of object and relation between them	decimal

Table 1 describes some of the attributes of the first dataset. Ten Features were computed for each cell nucleus of breast mass from a digitized image by Wisconsin hospitals. Three measure results i.e. mean, standard error and worst for all the features were also calculated which resulted into 30 features altogether in the dataset.

- Dataset 2 : The Breast Cancer dataset from the Hospitals of Wisconsin taken from Kaggle [4]. It consists of 699 rows with 11 attributes. Table 2 shows some of the features from the second dataset. All these attributes gives information about the breast cell tissue.

Table 2: Attributes of Dataset 2

Code(Attributes)	**Description**	**Domain(Values)**
class	tumor type	2-Benign or 4-Malignant
Clump Thickness	checks if cells are single or multi-layer	1-10
Uniformity of Cell Size	Density in cell size	1-10
Uniformity of Cell Shape	similarity of cell shape	1-10
Mitosis	level of cell reproductivity	1-10
Single Epithelial Cell Size	size of epithelial, if they are big	1-10
Normal Nucleoli	if nucleoli is small, large or dense	1-10

Both these data were originally derived from the real-time scenarios from the University of Wisconsin Hospitals.

3.2 Data Pre-processing

Some data might contain noise and this can affect the performance of the models. This step focuses on removing such noise from the dataset by performing some cleaning by

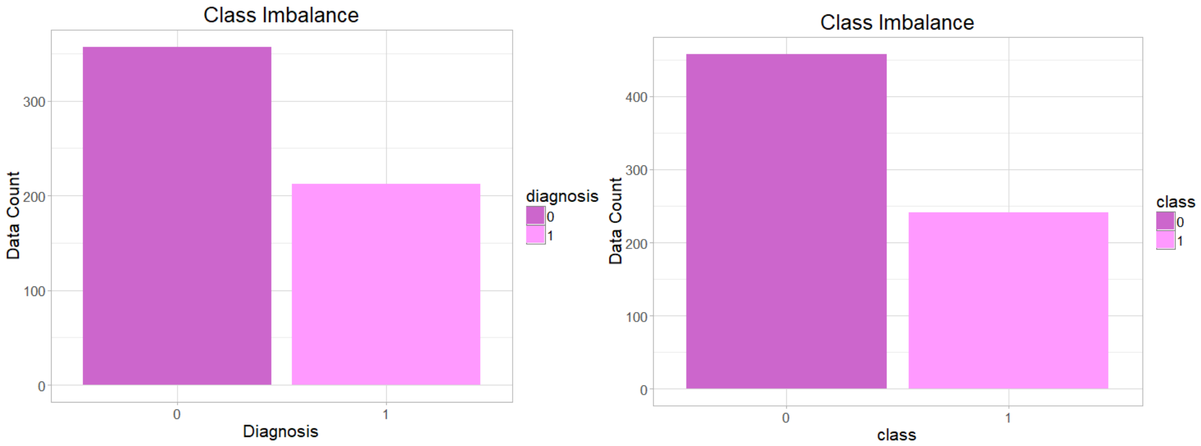
[4] https://www.kaggle.com/roustekbio/breast-cancer-csv

eliminating null rows or special characters to make the data more accurate for better results. While the Wisconsin dataset did not contain any missing values, the second dataset i.e. the Breast Cancer dataset taken from Kaggle had special characters which were replaced by the mean of that column. Removing those rows with special values might result in loss of some useful information, instead an alternate option of replacing it with mean was chosen.

3.3 Data Transformation

Data Transformation is the task of changing the data from one form to other. In this step, the data is transformed as per the requirement of the model. In our case, the target variable considered in the Wisconsin dataset i.e. “diagnosis” has categorical values of B for Benign tumor and M for Malignant tumor which was transformed to 0 and 1 respectively. Similarly, the second dataset had the target variable “class” as 2 and 4 where 2 is for Benign tumor and 4 is for Malignant tumor which was transformed to 0 and 1 respectively for simplicity.

3.4 Exploratory Data Analysis



(a) Imbalance in Dataset 1

(b) Imbalance in Dataset 2

Figure 3: Imbalance in the Datasets (0 : Benign Tumor, 1 : Malignant Tumor)

Figure 3a and Figure 3b shows that there is imbalance in both the datasets i.e. the first dataset has 357 of ”benign” cases and 212 of ”malignant” cases and the second dataset has 458 and 241 of ”benign” and ”malignant” cases. The class with fewer number of samples are known as the minority class whereas the class with comparably larger samples are known as the majority class. Having an imbalanced dataset causes the machine learning models to be more biased towards the majority class.

3.5 Feature Selection

Sometimes the data consists of many features, some of them might not be useful in achieving the results. Selecting the important features that contributes the most to the target variable is important to obtain the desired output as having unnecessary attributes might reduce the performance of the models by training it on the basis of irrelevant attributes. The Wisconsin Breast Cancer dataset consists of total 32 attributes which is quite high and might contain unnecessary features. But the main question is how do we know which feature to select. There are many feature selection algorithms available but the one used in this study is the Recursive Feature Elimination (RFE) feature selection technique as it is best suited for smaller datasets.

3.6 Model Application

This step of methodology is a very crucial part as selecting the algorithms as per the dataset and the aim of the study is the most important part. Various classification algorithms such as Support Vector Machine, Naïve Bayes, K-Nearest Neighbour, Logistic Regression, Classification and Regression Tree, Artificial Neural Network and Extreme Gradient Boosting are implemented in this research.

1. Support Vector Machine (SVM) : SVM is a supervised learning classification algorithm analyses the data used for classification or regression purposes. It is considered to be one of the capable machine learning algorithm in classifying problems Ahmad LG and AR (2013). SVM can classify breast cancer more precisely and can be used in breast cancer prediction problems Kumar (2013). It is one of the most widely used classification model in breast cancer problems Wang and Yoon (2015). SVM works well for smaller datasets since it won't take much time to process[5].

2. Naïve Bayes (NB) : NB is a machine learning algorithm which is applied with a strong naïve presumptions within the attributes. NB is considered to be one of the simplest Bayesian model. NB computes the probability of target variable by checking the times it has occurred in the training set which is known as "prior probability" Kumar (2013). One of the advantage of using NB is that it does not require large amount of data to calculate the mean and differences of variables which is important for classification M and S (2017).

3. K-Nearest Neighbour (KNN) : KNN model is used for classification and regression problems. KNN considers the similar instances and classifies them. Classification is done by considering majority of neighbouring instances Kumar (2013). KNN algorithms help to easily interpret the data.

4. Logistic Regression (LR) : Logistic Regression can be used when the target variable is categorical or binary in nature. LR can only be used if the target variable has not more than 2 classes. As the aim of this study is to predict the severity of breast cancer i.e. if the tumor is Benign or Malignant, LR is used as the data is categorical.

5. Classification and Regression Tree (CART) : CART is a decision tree algorithm in machine learning which helps in the prediction of target variable considering other variables. Decision Tree develops classification or regression models in a tree format. Decision tree are easy to understand as the problem is solved in a tree representation.

6. Artificial Neural Network (ANN) : ANN are the brain-inspired machine learning model which contains the input and the output layers. It also consists of one to

[5] https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589

many hidden layers which modifies the input layer into something that the output can use.

7. Extreme Gradient Boosting (Xgboost) : Xgboost is a machine learning algorithm used for classification and regression. Using Xgboost in medical domain can be beneficial to patients as well as physicians as it will help patients by diagnosing cancer early whereas physicians will be able to provide fine medical care to their patients Abed (2018).

4 Implementation

Figure 4 describes the framework on step-by-step processes that will be carried out. All these processes will be further explained in depth.

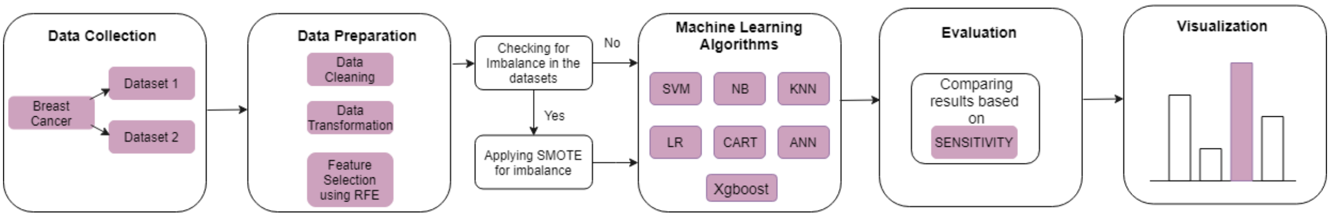


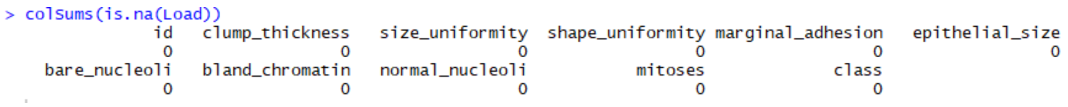
Figure 4: Framework

4.1 Data Preparation

1. Data Pre-processing :
   Since the Wisconsin dataset did not contain any missing values, the main focus was on the second Breast Cancer dataset as a column "bare-nucleoli" contained 16 cells with special character, denoted as "?". Following are the steps followed for cleaning the data :

   - Removing special character : The gsub() function is used to remove the special value. The advantage of using gsub() is it removes all the occurrences at a single time. Once the code is run, all the cells with "?" was replaced by a null value.
   - Calculating the mean of the column : As, deleting the row with null value might result in important data loss, the mean for the column "bare-nucleoli" was computed using mean() function.
   - Replacing the null value with the mean : After getting the mean value for the column, all the null values are replaced with the mean value.

   As shown in Figure 5, the dataset is now ready to use for further processes as it does not have any special or null value in it.



```
> colSums(is.na(Load))
             id  clump_thickness  size_uniformity shape_uniformity marginal_adhesion   epithelial_size
              0                0                0                0                 0                 0
  bare_nucleoli  bland_chromatin  normal_nucleoli          mitoses             class
              0                0                0                0                 0
```

Figure 5: Cleaned Data

2. Data Transformation :
   After the cleaning, transformation of the data as per the necessity is done. Transformation is an important step as the data need to be transformed as per the model requirement for it to function properly. As mentioned earlier, both the dataset are transformed.

   - In the Wisconsin data, the target variable “diagnosis” was transformed from B and M to a factor of binary variables 0 and 1, 0 for benign tumor and 1 for malignant tumor. As the ANN model works on numeric data the factor was then converted to numeric data-type.
   - The second dataset had the target variable “class” as 2 and 4 i.e. 2 for benign and 4 for malignant which was transformed to 0 and 1 respectively for better understanding. It was later converted to numeric data-type for the ANN model.

3. Feature Seletion :
   As mentioned above, RFE feature selection is applied on the Wisconsin dataset in order to select important features. The rfe() contains a number of pre-defined functions like linear regression (lmFuncs), naive bayes (nbFuncs), randon forest (rfFuncs) and the one used is the ”rfFuncs” [6]. After applying RFE, 10 features are selected for further classification models, i.e. symmetry-mean, symmetry-se, compactness-worst, concave.points-se, smoothness-worst, smoothness-se, concavity-worst, texture-se, concavity-se, concave.points-worst. However, feature selection was not performed on the second dataset as it consisted of 10 attributes and all of these were selected for model implementation.

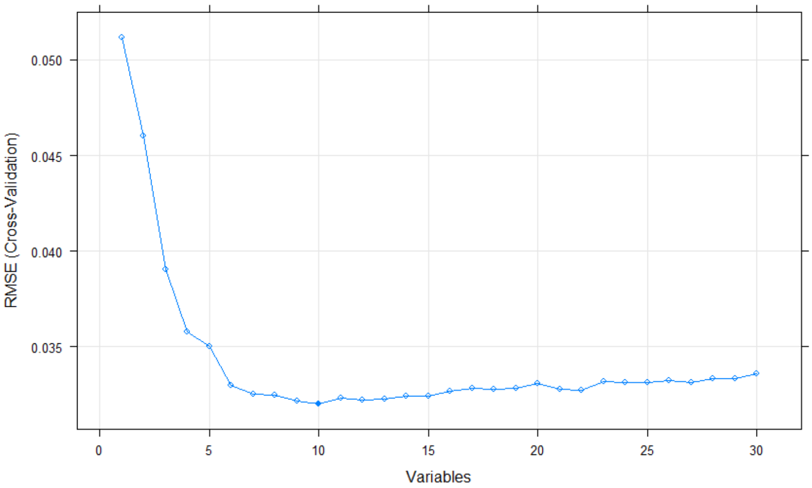
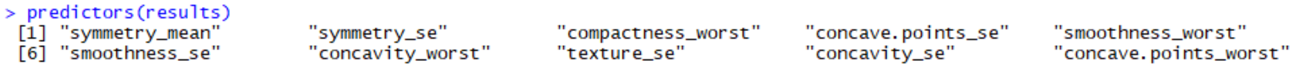


Figure 6: Number of Features Selected - Dataset 1

[6] https://topepo.github.io/caret/recursive-feature-elimination.html



```
> predictors(results)
 [1] "symmetry_mean"        "symmetry_se"          "compactness_worst"    "concave.points_se"    "smoothness_worst"
 [6] "smoothness_se"        "concavity_worst"      "texture_se"           "concavity_se"         "concave.points_worst"
```

Figure 7: Features Selected - Dataset 1

Figure 6 shows that out of 30 features 10 important features were selected on application of RFE algorithm and Figure 7 represent the names of features selected that are chosen in the first dataset. These features are used further in training the algorithms.

4.2 Classification Techniques

Applying and Implementing algorithms is one of the most necessary part of machine learning study. Various classification models have been implemented in this research that include Support Vector Machines, Naïve Bayes, K-Nearest Neighbour, Logistic Regression, Classification and Regression Tree, Artificial Neural Networks and Extreme Gradient Boosting. R software is used to carry out the analysis. Before applying any models, the imbalance in the dataset needs to be solved. There are various methods to deal with this issue such as over-sampling, under-sampling, of which the technique chosen is Synthetic Minority Over-Sampling Technique.

4.2.1 Synthetic Minority Over-Sampling Technique (SMOTE):

To solve the problem of class imbalance, Chawla et al. (2002) introduced the technique of Synthetic Minority Over-Sampling Technique (SMOTE). SMOTE is a process of over-sampling of the minority class and under-sampling the majority class. The percentage of Over-Sampling denoted by "perc.over" tells the number of synthetic samples that needs to be generated and is at all times a multiple of 100. If "perc.over" is set to 100, then the number of minority samples i.e. :perc.under" is doubled. If the "perc.over" is set to 200, then "perc.under" is tripled [7]. There are two methods to apply SMOTE :

1. First approach is to split the dataset into training and test set and then applying SMOTE on just the training set.

2. Second approach is to apply SMOTE on the whole dataset and then divide it into train and test split.

In this research, the first approach of applying SMOTE to just train data is used as it is considered to be the right way to oversample and helps to produce good results whereas the second approach i.e. upsampling the whole dataset and then splitting it into train and test sets might create same observations in both the sets and may not give accurate output [8].

[7] https://medium.com/towards-artificial-intelligence/application-of-synthetic-minority-over-sampling-technique-smote-for-imbalanced-data-sets-509ab55cfdaf

[8] https://beckernick.github.io/oversampling-modeling/

4.2.2 Support Vector Machine (SVM):

SVM algorithm is used to classify breast cancer tumor. The caret package in R helps to implement the SVM model and is used for the same. Caret gives train() method which helps to train the data on several techniques. Before this step, trainControl() method is passed with 2 parameters i.e. method as "cv" and number as 10 iterations as 10-fold Cross Validation is used to check the performance. The trainControl() is then passed as an argument in the train() method along with method as svm, train data, preProcess and tuneLength. The preProcess is set to center and scale as it will help in centering and scaling the data. The tuneLength contains an numeric value and is used for tuning the algorithm.

4.2.3 Naive Bayes (NB):

The e1071 package is used to implement NB. It provides the function "naiveBayes()" which is useful in carrying out Bayes classification. Once the model is trained with the naiveBayes() function it is then tested by using predict() by passing the trained model along with the test data as an argument[9]. 10-fold Cross Validation is performed to evaluate the performance of the algorithm.

4.2.4 K-Nearest Neighbour (KNN)

The KNN classification technique is implemented using the kNN() function[10] and the features selected from the feature selection algorithm. Once the model is trained on the training set using 10-fold Cross Validation, it is then tested on the test set to check how well the model has classified the target variable.

4.2.5 Logistic Regression (LR):

The glm() is used to train LR model which stands for generalized linear models. The glm() is passed with formula, data which is training set and family as "binomial(link="logit")" for logistic regression. The predict() is then used to test the implemented model[11].

4.2.6 Classification and Regression Tree (CART):

CART is a machine learning algorithm which is implemented using the train() method. The method parameter in train() is given the value as "rpart" as the implementation of CART in R is known as RPART. The other parameters are the predictor variables and the prediction variables selected from feature selection.

4.2.7 Artificial Neural Network (ANN):

Before applying ANN, the target variable is converted into numeric datatype as neural network works on numeric data. The "neuralnet" package is used to train the ANN model [12]. The target variable along with the prediction variables are passed as an argument in

[9] https://www.rdocumentation.org/packages/e1071/versions/1.7-3/topics/naiveBayes

[10] https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/kNN

[11] https://www.r-bloggers.com/predicting-creditability-using-logistic-regression-in-r-cross-validating-the-classifier-part-2-2/

[12] https://www.rdocumentation.org/packages/neuralnet/versions/1.44.2/topics/neuralnet

the neuralnet() method with linear.output set to false. The model is trained with less number of hidden layers as the model performs better with lower hidden layers NEMISSI et al. (2018).

4.2.8 Extreme Gradient Boosting (Xgboost):

The Xgboost algorithm[13] is applied using the xgboost package and 10-fold cross validation technique. The train() method is passed with arguments like xgbTree, trainControl with 10-fold cross validation and tuneGrid. The tuneGrid is used to tune the algorithm with various parameters, some of which are "eta" which is the learning rate, max-depth which is the depth of the tree, number of iterations as nrounds.

5 Evaluation

After the implementation of various techniques, the performance of the model is measured using several criteria.The aim of evaluation is to provide understanding about the patterns that is achieved by interpreting the results. Following are the metrics that are used to evaluate the classification techniques :

- Accuracy : Accuracy is the number of correct predictions that the model made to the total number of predictions. The formula to calculate accuracy is given in Equation 1 :

$$Accuracy = \frac{TruePositive + TrueNegative}{Total} \tag{1}$$

- Sensitivity : Sensitivity also known as the True Positive Rate is the ratio of positive classes that are correctly classified as positive with respect to all the positive classes in the data. Sensitivity is calculated with the formula given in Equation 2 :

$$Sensitivity = \frac{TruePositive}{FalseNegative + TruePositive} \tag{2}$$

- Specificity : Specificity also known as the False Positive Rate is the ratio of the negative classes that are wrongly classified as positive with respect to all the negative classes in the data. Specificity is calculated with the formula given in Equation 3 :

$$Specificity = \frac{FalsePositive}{FalsePositive + TrueNegative} \tag{3}$$

Although all these metrics are considered to evaluate the performance of the model, however the best model will be judged based on the sensitivity because in medical data it is important to identify the True Positives i.e. the number of correctly identified patients with cancer out of all the patients.

[13] https://analyticsdataexploration.com/xgboost-model-tuning-in-crossvalidation-using-caret-in-r/

Table 3: Performance - Dataset 1

(Models)	**Accuracy %**	**Sensitivity %**	**Specificity %**
SVM	90.27	90.14	90.48
NB	88.50	87.32	90.48
KNN	94.69	**97.10**	90.09
LR	91.15	91.54	90.47
CART	88.50	87.32	90.48
ANN	89.47	96.34	71.87
Xgboost	92.03	92.95	90.47

Table 3 shows the performance of all the models for Dataset 1 in terms of accuracy, sensitivity and specificity. As the model is judged based on the Sensitivity i.e. the True Positive Rate, the model which has correctly classified maximum number of true positives is KNN with 97.10% followed by ANN with 96.34% of sensitivity.

Table 4: Performance - Dataset 2

(Models)	**Accuracy %**	**Sensitivity %**	**Specificity %**
SVM	94.25	97.08	88.88
NB	95.69	95.62	95.83
KNN	96.17	96.40	95.71
LR	93.30	97.08	86.11
CART	82.78	85.40	77.77
ANN	94.25	96.99	89.47
Xgboost	94.74	**97.81**	88.89

Table 4 shows the model performance for Dataset 2. In Dataset 2, Xgboost performed very well with the sensitivity of 97.81% followed by 97.08% achieved by both SVM and LR algorithms.

5.1 Experiment 1 : Choosing the K-value for KNN technique

For the KNN model, choosing the right k-value is the most important task as the model performance depends on the k-value that is selected. For the same, performance of the KNN algorithm at different k-values is analyzed for both Dataset 1 and 2 in order to select the best value for the KNN implementation.
Figure 8 shows the accuracy and semsitivity with respect to the K-values 10,15,20,25 and 30 for the KNN technique. It can be seen that the k-value 25 and 30 gives high and similar sensitivity, so the k-value considered in the 1st dataset is 25.

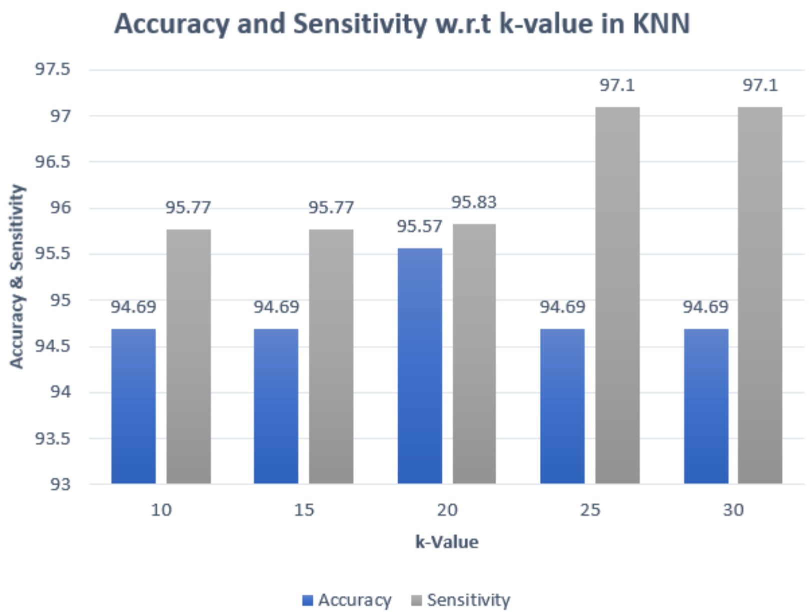


Figure 8: Selecting the k-value for KNN : Dataset 1

In Figure 9, the k-value 20 and 25 both produce similar and highest sensitivity among others i.e. 96.40%. Hence, the value chosen for KNN implementation in 2nd dataset is 25 as the same has been chosen for dataset 1 as well.

As the k-value 25 works good for both the dataset and gives highest rate of true positives, this can help the physicians in the medical domain to effectively identify the tumor type.

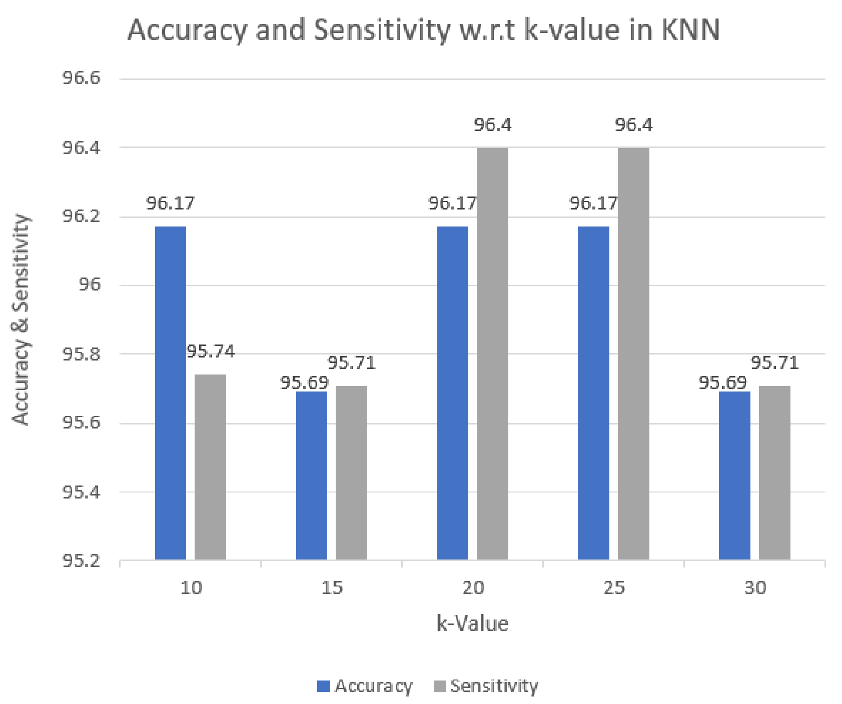


Figure 9: Selecting the k-value for KNN : Dataset 2

5.2 Experiment 2 : Choosing the number of hidden layers for ANN

The ANN model consists of a input layer, a hidden layer and output layer. The number of hidden layers to use needs to be specified else the model takes 1 as default layer. The hidden layer also affects the performance of the model. Hence, several trials considering hidden layer as 2,3,4,5 and 6 are performed to select the best hidden layer for the algorithm.

Table 5: Hidden Layer - Dataset 1

Hidden Layer	**Sensitivity %**
2	**96.34**
3	96.10
4	95.40
5	95.40
6	94.11

According to Table 5, hidden layer 2 gives high sensitivity of 96.34% as compared to others. Hence, the value selected for the 1st dataset is 2. SImilarly, on the 2nd dataset, the same experiment is performed with the different hidden layer i.e. 2,3,4,5 and 6 to choose the best among them.

Table 6: Hidden Layer - Dataset 2

Hidden Layer	**Sensitivity %**
2	95.62
3	96.29
4	93.61
5	**96.99**
6	95.58

As seen in Table 6, hidden layer 5 has achieved maximum number of correct predictions i.e. 96.99 % than hidden layer 2 which is 95.62%. Hence, the value used for hidden layer in 2nd dataset is 5 as it gives the highest sensitivity. This experiment helped in identifying the hidden layer which gives better results and can be useful to the medical practitioners in recognizing the tumor type.

5.3 Discussion

In this research, the severity of the breast cancer is identified using seven machine learning algorithms i.e. SVM, NB, KNN, LR, CART, ANN and Xgboost using 2 breast cancer datasets. Some important features are selected using the RFE feature selection algorithm.

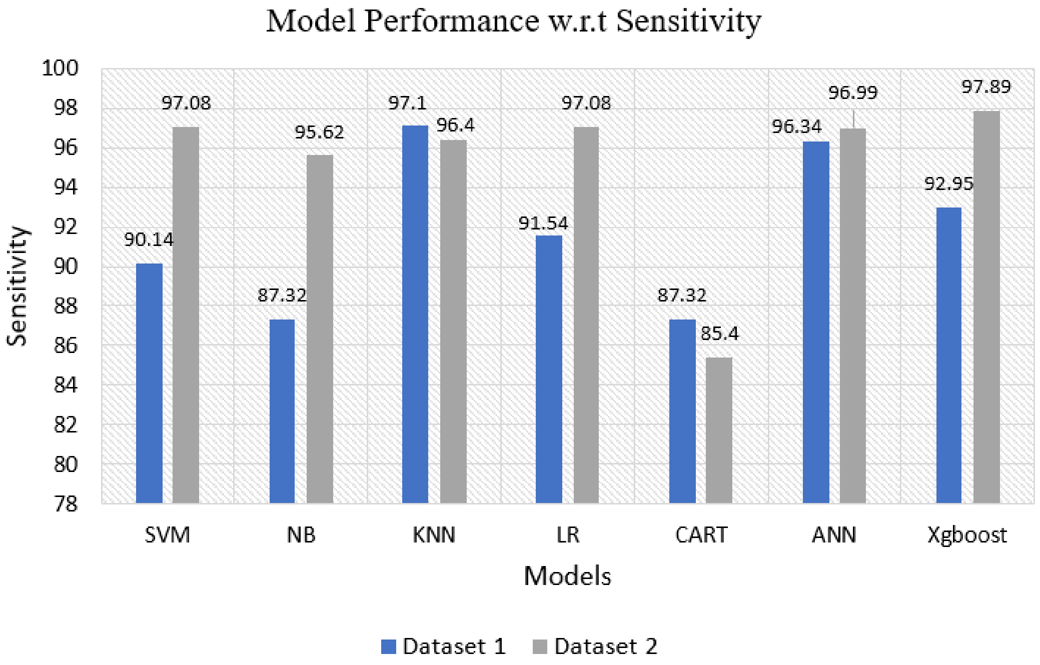


Figure 10: Result w.r.t sensitivity for Dataset 1 and 2

As shown in Figure 10, in the 1st dataset KNN performed best with the sensitivity of 97.10% followed by ANN with very less difference i.e. 96.34% sesitivity, whereas Xgboost outperformed in the second dataset with 97.81% sensitivity. Most of the models performed quite well in the 2nd dataset including ANN, SVM, KNN and LR with the sensitivity all above 96%. This shows that ANN and KNN techniques worked well on both the datasets whereas Xgboost, SVM and LR worked good on the 2nd dataset but not that good on the 1st dataset.

To check the performance of the model, 10-fold Cross Validation is used for these models and the metrics are noted. The reason for using the k-value as 10 is because it was widely used in most of the papers reviewed Chaurasia et al. (2018),Graja et al. (2018). But, to see how the model works with respect to the k-value, the k is changed to 20 and the difference in the results are examined. However, it was seen that there was no difference in the working of the models and generated similar results for both 10-fold and 20-fold cross validation.

Although the evaluation metrics accuracy, sensitivity and specificity is calculated, the main metric on the basis of which the model is evaluated is sensitivity. The reason behind this being a medical data, the need of knowing which model successfully classifies the tumor type correctly becomes very necessary for treating the disease . And as sensitivity is the calculation of True Positives that the model has correctly identified out of all the True Positives, this metric is extremely important.

6 Conclusion and Future Work

The main aim of this study is to recognize a model that will correctly identify the cancer severity in patients. Two dataset are used to conduct the study. As the datasets are not balanced, SMOTE technique is applied to solve the imbalance. Later, seven machine learning algorithms have been implemented and the performance of these models are analyzed with the sensitivity score.

The result shows that while KNN and Xgboost gave best sensitivity i.e. 97.10% and 97.81% on the 1st and 2nd dataset respectively, KNN and ANN worked better on the both the datasets with senitivity above 96% on both the data and hence, can be useful to medical practitioners to detect benign and malignant tumors successfully at an early stage. As there is a great need for diagnosing breast cancer as early as possible as the death rates are increasing, these models can help in achieving this aim efficiently.

In this study, the ANN model is tested with different hidden layers, however in the future more number of layers can be added and the patterns of performance of the model can be analyzed. Also, different feature selection methods can be used to examine any improvement in the model performance.

Acknowledgement

I would like to thank god for being able to complete this project. I am greatly thankful to Dr. Muhammad Iqbal for constantly guiding me and giving me feedback during this 13 week project phase which has served a major contributor towards completion of this project. I would further express my gratefulness to my family for supporting me throughout my academic career. Lastly, my friends for inspiring me in every way possible.

References

Abed, S. I. (2018). Predicting breast cancer using gradient boosting machine, *International Journal of Science and Research (IJSR)* **8**(6): 885–891.

Ahmad LG, Eshlaghy AT, P. A. E. M. and AR, R. (2013). Using three machine learning techniques for predicting breast cancer recurrence, *Journal of Health Medical Informatics* **4**(2): 1–3.

Basha, M. and Iyenger, S. N. (2018). A novel approach to perform analysis and prediction on breast cancer dataset using r, *International Journal of Grid and Distributed Computing* **11**(2): 41–54.

Bellaachia, A. and Guven, E. (2006). Predicting breast cancer survivability using data mining techniques, **58**.

Chaurasia, V., Pal, S. and Tiwari, B. (2018). Prediction of benign and malignant breast cancer using data mining techniques, *Journal of Algorithms & Computational Technology* **12**(2): 119–126.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research* pp. 321–357.

Dai, B., Chen, R., Zhu, S. and Zhang, W. (2018). Using random forest algorithm for breast cancer diagnosis, *2018 International Symposium on Computer, Consumer and Control (IS3C)*, pp. 449–452.

Delen, D., Walker, G. and Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods, *Artificial Intelligence in Medical* **34**(2): 113–127.

Graja, O., Azam, M. and Bouguila, N. (2018). Breast cancer diagnosis using quality control charts and logistic regression, *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*, pp. 215–220.

Jivani, A. and Shah, C. (2013). Comparison of data mining classification algorithms for breast cancer prediction, *International Conference on Computing, Communications and Networking Technologies (ICCCNT)* .

Keles, M. K. (2019). Breast cancer prediction and detection using data mining classification algorithms: A comparative study, *Department of Computer Engineering, Adana Science and Technology University* **26**(1): 149–155.

Khourdifi, Y. and Bahaj, M. (2018). Applying best machine learning algorithms for breast cancer prediction and classification, *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, pp. 1–5.

Kumar, G. R. (2013). An efficient prediction of breast cancer data using data mining techniques, Vol. 2, pp. 139–144.

M, S. A. and S, S. J. (2017). Prediction of breast cancer, *International Journal of Research in Engineering, IT and Social Sciences* **7**(3): 21–28.

Mekha, P. and Teeyasuksaet, N. (2019). Deep learning algorithms for predicting breast cancer based on tumor cells, *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*, pp. 343–346.

M.Lundin, J.Lundin, H.B.Burke, S.Toikkanen and L.Pylkkänen (1999). Artificial neural networks appliedtosurvival predictioninbreastcancer, *Department of Oncology, University of Helsinki, Department of Pathology and Oncology* **57**: 281–286.

Mudgil, P., Garg, M., Chhabra, V., Sehgal, P., Jyoti and Journal, I. (2019). Breast cancer prediction algorithms analysis, *International Journal of Advance Research, Ideas and Innovations in Technology* **5**(1): 424–427.

Mulatu, D. and Gangarde, R. R. (2017). Survey of data mining techniques for prediction of breast cancer recurrence, *) International Journal of Computer Science and Information Technologies(IJCSIT)* **8**(6): 599–601.

Mushtaq, Z., Yaqub, A., Hassan, A. and Su, S. F. (2019). Performance analysis of supervised classifiers using pca based techniques on breast cancer, *2019 International Conference on Engineering and Emerging Technologies (ICEET)*, pp. 1–6.

NEMISSI, M., SALAH, H. and SERIDI, H. (2018). Breast cancer diagnosis using an enhanced extreme learning machine based-neural network, *2018 International Conference on Signal, Image, Vision and their Applications (SIVA)*, pp. 1–4.

Ojha, U. and Goel, S. (2017). A study on prediction of breast cancer recurrence using data mining techniques, *2017 7th International Conference on Cloud Computing, Data Science Engineering - Confluence*, pp. 527–530.

Padmapriya, B. and Velmurugan, T. (2014). A survey on breast cancer analysis using datamining techniques, *EEE International Conference on Computational Intelligence and Computing Research* pp. 1234–1237.

Rajesh, K. and Anand, D. S. (2012). Analysis of seer dataset for breast cancer diagnosis using c4.5 classification algorithm, *International Journal of Advanced Research in Computer and Communication Engineering* **1**(2): 72–77.

Sadoughi, F., Ahmadi, M. and Lotfnezhad, H. (2015). Prediction of breast cancer survival through knowledge discovery in databases, *Global Journal of Health Science* **7**(4): 392–397.

Singhal, P. and Pareek, S. (2018). Artificial neural network for prediction of breast cancer, *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on*, pp. 464–468.

Vivek Kumar, Brojo Kishore Mishra, M. D. and Abhishek Verma (2019). Prediction of malignant benign breast cancer: A data mining approach in healthcare applications, pp. 1–8.

Wang, H. and Yoon, S. W. (2015). Breast cancer prediction using data mining method.

Williams, K., Idowu, P., Balogun, J. and Oluwaranti, A. (2015). Breast cancer risk prediction using data mining classification techniques, *Transactions on Networks and Communications* **3**(1).

Zand, H. K. K. (2015). A comparitive survey on data mining techniques for breast cancer diagnosis and prediction, *Indian Journal of Fundamental and Applied Life Sciences* **5**(3): 4330–4339.