National College of Ireland

A Comparative study of Breast Cancer Diagnosis and Classification Using Neural Networks and Machine learning models

MSC Research Project
Data Analytics

SAKSHI DUBEY
Student ID: x19201290

School Of Computing
National College of Ireland

Supervisor: Dr. Bharathi Chakravarthi

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Sakshi Dubey

Student ID: X19201290

Programme: Data Analytics **Year:** 2021-2022

Module: MS Research Project

Supervisor: Dr. Bharathi Chakravarthi

Submission Due Date: ……

Project Title: A comparative study of breast cancer diagnosis and classification using neural networks and machine learning models.

Word Count: 9418 **Page Count** 22 pages

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: 29/01/2022

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):	

A Comparative study of Breast Cancer Diagnosis and Classification Using Neural Networks and Machine learning models

Sakshi Dubey
x19201290

1. Abstract

Over the time period, the rise in incidence and mortality rates due to deaths caused from breast cancer is evident that it's turning a terminal disease globally. An early detection of these developments in the female body can reduce the risk of cancer spreading throughout the body and expensive treatment costs. Early diagnosis and modern cancer treatment is essential for better understanding of development of cancerous tissues in the female breast and most importantly prevent deaths from cancer. Hence, it is essential to develop a robust system for early-stage breast cancer diagnosis that can assist the medical professionals to classify cancer tissues in mammograms and thus save lives of patients. In this research, we are implementing nine machine learning algorithms and one artificial neural network model on "Wisconsin Breast cancer" dataset to determine the presence of cancer and classify the type of abnormality as benign and malignant. The primary objective of our research is implementing machine learning and deep learning algorithms on trained data and detect and classify the severity of cancer through effective exploratory data analysis, feature selection and performance metrics of classification models. We intend to provide a robust and reliable approach for early-stage diagnosis of breast cancer using machine learning and deep learning methodologies so there can be enough space for treatment plans and higher chances of survival in patients.
Keywords: Breast cancer, machine learning, deep learning, malignant, benign

2 Introduction

1.1 Project Background and Motivation

1.1.1.1 Brief overview of breast cancer

Over the time period, breast cancer has briefly affected many lives being one of the most common causes of death among women and slowly and gradually equating these statistics with lung cancer. We can identify breast cancer as a non-skin cancer that is diagnosed quite frequently in women considering the risk factors as lifestyle changes, aging, gene mutations and most often a family history of breast cancer. We can identify early signs of breast cancer as it begins with development of a denser breast tissue and slowly grows to a wider area in

one of the breasts. The development of this takes place from the inner lining milk ducts in the breast or basically the lobules provide the ducts with milk. Breast cancer can be best explained as a mutation in Deoxyribonucleic acid and ribonucleic acid that transforms normal cells in the body into cancer cells. In many proposed studies, the common factors giving rise to these mutations include exposure to nuclear radiation and electromagnetic radiation viruses, harmful bacteria, fungi, parasites, consumption of contaminated food and water, cell injuries and lastly aging or modification of cell molecules. Breast cancer is further classified as malignant and benign depending on the fatality caused by the cancer tissues. Despite most tumors occur due to non-cancerous development in breasts but it becomes a serious issue when a benign cancer tissue transforms into malignant cancer. Hence it becomes important to identify the initial stage when the malignant cells have not approached the breasts. An early detection of these developments in the female body can reduce the risk of cancer spreading throughout the body and expensive treatment costs. Early diagnosis and modern cancer treatment is essential for better understanding of development of cancerous tissues in the female breast and most importantly prevent deaths from cancer. If not detected early, there are higher chances of patient going through complex invasive treatments and with very less probability of survival post diagnosis.

The rise in incidence and mortality rates due to deaths caused from breast cancer is evident that it's turning a terminal disease globally. Recent statistics from globocan survey in 2018 breast cancer as one of every four cancers diagnosed in women and stands second in the world causing maximum cancer deaths. Also, the mortality rate was noted as 6.8 per one lakh women and similarly age subjected breast cancer incidence was noted as 23.7 per one lakh cases worldwide. With increase in modern science developments and their contribution in breast cancer research, we still have long way to go towards developing robust treatment plans for combating breast cancer.

1.1.1.2 Traditional global breast cancer screening methods

In this section, we are discussing the traditional global breast cancer screening methods adopted across developed and developing countries. In most cases, nations primarily adopt breast cancer screening mechanisms depending on the availability of resources with them. The clinical breast cancer examination and breast self-examination methodologies are

adopted by developing countries for screening of breast cancer. Similarly, the procedure of mammography is prevalent in developed countries. In most cases, obesity, family history, bad lifestyle habits, bad nutrition and lack of exercise are highlighted as red flags for breast cancer. With update in treatment plans in healthcare industry and presence of enough data to conduct research, it becomes the prime responsibility of researchers to discover cure for cancer.

1.1.1.3 Traditional global breast cancer detection methods

In this section, we are discussing the traditional global breast cancer detection methods adopted across developed and developing countries. The most effectively used detection test is the triple assessment test often referred as gold standard for breast cancer detection. It comprises of three tests that includes mammography/ ultrasonography that is radiological imaging and pathology that includes core needle biopsy. If any of these tests positive then the patient is diagnosed with malignancy while if three of them test negative then its a benign breast cancer condition. The traditional methodology of breast cancer screening often follows a linear path in understanding the risk factors using regression. Machine learning proves to be efficient in breast cancer identification and diagnosis since it doesn't assume linearity in the procedure.

1.1.2 Implementation of machine learning in predicting breast cancer.

1.1.2.1 Significance of ML based breast cancer prediction system for screening and detection

Data mining and machine learning methodologies prove to be efficient in understanding and interpreting data. The patient data obtained from any of the screening or detection tests can be utilized to develop machine learning models that can further assist in predicting breast cancer. The data can be anything, rights from results obtained from mammography or ultrasonography and even the core needle biopsy tests. Sometimes getting access to patient data can be an expensive affair, in this case using data from a patient's anthropometric blood tests proves to be a better approach. Once a machine learning model performs effectively with patient's blood data then it can be further used to develop an AI tool that if approved by the food and drug administration can be to identify symptoms of breast cancer in patients by clinicians. In this procedure, the experts can obtain the output by inputting patient's blood result data in the tool that can identify the stage or abnormality of cancer. This prototype of breast cancer detection by machine learning models proves to have more potential if compared to the traditional screening and detection tests for breast cancer.

1.1.2.2 Motivation

Over the time period, the ratio of deaths in women from any form of cancer has increased. It is essential to diagnose cancer at an early stage to identify the symptoms and derive efficient treatment plans for the patient. This involves collective work by researchers, medical professionals and communities to actively participate in developing awareness, effective treatment plans conducting significant research to help prevent the fatality caused by cancer. Figure 1 shows fatalities cause by different types of cancer among women in the year 2018 and as we can see that breast cancer proves to be dominant causing maximum deaths worldwide. As a data science aspirant working on breast cancer research, it is essential to develop a robust model that helps in breast cancer classification and detecting the tumour accurately. If the models fail in detection of malignancy, then the patient may die without

being aware of the symptoms. If we are successful in determining cancer at early stage then it can result in saving many lives and giving way for effective treatment plans. Hence it becomes important to develop models that are capable of rapidly predicting these symptoms so that we can construct treatment plans for patients and thus reduce the risk of fatalities.

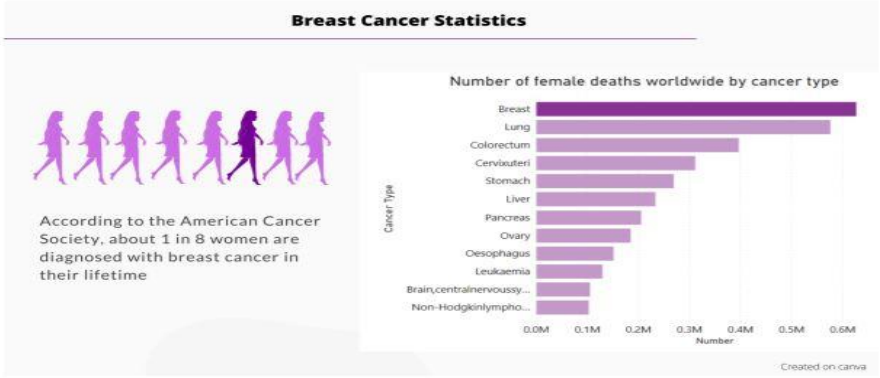

Figure 1: Cancer Fatalities by Type

The motivation for this study is to enable medical professionals to adopt machine learning and deep learning methodologies in order to increase/ make the rate of breast cancer detection and classification more rapid and reliable. This study extensively involves literature review of twenty-one research papers that cover the work done by data science scholars in the field of breast cancer detection and classification adopting machine learning and deep learning methodologies.

1.1.2.3 Project Requirement Specification

1.1.3 Research Question

RQ: "To what extent can machine learning and neural network models assist in accurate and early diagnosis of breast cancer?"

1.1.4 Research objectives

1] Application of feature selection algorithms to select significant features contributing in breast cancer classification.

2]Understanding the nature of data through effective exploratory data analysis and visualizations and balancing the data precisely to train the model.

3]Implementation of machine learning and deep learning models on trained data and evaluating performance metrics of these models on parameters of accuracy, sensitivity, specificity and f1 score.

4]Training the models accurately for classifying the tumour as "Benign" or "Malignant"

2 Related Work

The healthcare industry has been one of the most promising fields of research for data science scholars due to the presence of significant data and suitable data types. If we take into

account the patient data in hospitals that contains continuous records of symptoms, category of diseases, death history and also patients admitted in numerical values. Health industry can be defined as a system flexible for research and studies and one which welcomes and implements improvement methodologies for welfare of community. We have a detailed literature survey highlighting amazing research work, studies, surveys and review conducted by data science scholars and researchers that includes implementation of machine learning and deep learning methodologies towards breast cancer classification and diagnosis.

1]In this study, the author has focused on presenting an overview of machine learning and deep learning methodologies and their applications in breast cancer classification and diagnosis. The author has used multi modalities imaging to classify the breast cancer data and further determine the presence of cancerous and non-cancerous masses in breast. The author has summarized a brief overview of research databases that influenced the study and has explained the variation in results obtained from machine learning and deep learning models in detection and classification of cancer cells.2] In this study, the author has followed a different approach for breast cancer diagnosis by incorporating a differential evolution algorithm of radial based function kernel extreme learning machines. The network structure of RBF-KELM model contains two significant parameters 'C' and 'sigma' which determines the efficiency of the model. The author has investigated the performance of differential evolution algorithms on two datasets and was successful in fetching significant insights in breast cancer classification.3] In this study, the author has focused on developing machine learning models and studying their performance in detection of breast cancer. The subject of author's examination was diagnosis cancer tissues and further classifying them into malignant and benign category that can help in determining the possibility and life span of patient. The author kept their primary focus on understanding features contributing in diagnosis and also evaluating performance of models based on their accuracy, affectability, explicitness and exactness to determine results. 4] In this study, the author has focused on segmenting breast cancer based on its malignancy from an ultrasound image data by implementing backpropagation neural network models. The author has conducted brief research on patient data that comprises 184 images in total and was successful in obtaining high rates of precision in characterizing patient data into cancerous and non-cancerous with the help of deep learning models.5] On similar lines, this study includes implementation of artificial neural network models for classification and diagnosis of breast cancer data. The author has classified tumor cells based on their degree of malignancy by defining backpropagation and radial basis neural network algorithms.On further analysis of patient data, researchers were successful in conducting automatic classification of cancer cells that lead to breast cancer in patients by comparing performance of Radial based and backpropagation neural network models. The accuracies of both classification models were noted as 50% and 70%.6]In this study, the author has used multimodal ultrasound images for determination of molecular subtypes of breast cancer that facilitates good treatment plans and also improves patient prognosis. The author has implemented this task of determining molecular subtypes by using three multimodal, dual modal and monomodal convolutional neural network models and further examining their performance based on accuracies. On further implementation, it was observed that multimodal CNN model outperformed the other two models in classifying five specific molecular subtypes leading to breast cancer and also triple negatives from non-triple negative breast cancer data. Thus, a multi modal Convolutional neural network model proves to be successful in breast cancer prediction and classification.7] As we go further, the primary focus of this study is conducting image analysis of clinical patient data and predicting breast cancer by implementing multi-input classification model. In many studies, a multi-input classification model is proved to be

highly beneficial than convolutional neural network models. At first, the researchers determined patterns in thermal images and further moved to clinical data with an aim to improve the model performance. The best performing model on the mastology research dataset was a multi input convolutional model with an accuracy of 97% and specificity as 100%.8] The researchers in this study followed an interesting approach of deep learning methodologies for segmentation of breast lesions found in mammographic images into malignant and non-malignant categories. The first approach includes utilizing images patches from region of interest and second approach includes taking whole image data into considerationThe author was successful in determining significant features that assisted in classification of breast lesions from mammograms and all the models demonstrated best results with 99% accuracy, sensitivity and specificity.9)]The author has significantly focused improvement of current breast cancer detection methodologies by implementing a combined convolutional neural network that includes graph convolutional neural network and convolutional neural network. Two specific improvement techniques were used that included dropout and batch normalization. An 8-layer CNN model was developed that included rank based stochastic pooling, improvement techniques and another two-layer graph convolutional neural network model for classification of malignant tissues from mammograms.10] This study focuses on developing a convolutional neural network model for breast cancer classification from different categories of images that includes thermal, mammograms and ultrasound images. The author has focused on building a five layered convolutional layer that includes a fully connected neural network layer for feature extraction and parameters affecting the classification. The convolutional neural network models demonstrated an accuracy of 96% and specifically the data augmentation techniques were successful in reducing errors and provides good generalization in the model. 11]In this study, the researchers aim at helping medical professionals by developing convolutional neural network models for breast cancer classification and diagnosis. The performance of 8 level neural network models were examined on medical images obtained from patient data to determine the presence of cancer cells and their degree of malignancy. On further analysis, the CNN models were able to successfully distinguish the patient data into malignant, benign or none and assisting the healthcare professionals for segregating breast cancer treatment plans for patients. All the deep learning models demonstrated best performance with respect to accuracy, sensitivity and specificity.12] This study primarily focuses on developing a neural network framework for breast cancer classification by utilizing the concept of transfer learning on two publicly available datasets that include 7000 microscopic breast images. The author has developed a “MultiNet” framework in order to enhance the efficiency of process of breast cancer classification. Three pretrained neural network models were used to conduct feature extraction on the microscopy image dataset and further were fed into the network layer to develop a neural network model. The classification models performed best with “MultiNet” framework yielding an accuracy of 98% with respect to accuracy, sensitivity and specitivity.13] This study involves a survey primarily focused on examining and understanding the functionality of computer assisted detection systems and Convolutional neural network models in breast cancer classification. The author also examined quantitative results obtained from the models and concluded that the approach of convolutional neural networks proves to be more reliable than traditional computer detection systems for cancer diagnosis.14] This study primarily focuses on application of machine learning algorithms in classification of cancer tissues present in breast mass and further investigating the performance of models on the image data based on feature extraction, accuracy, sensitivity and specificity.15] This study focuses on a similar objective of developing algorithms based on biosensors and machine learning to identify cancer tissues in microscopic images. The author has implemented different machine learning algorithms that includes naïve bayes,

support vector machine, k nearest neighbors and random forests classifier on variety of datasets that includes fine needle aspirations of breast mass, infrared images, mammography, microscopic images, thermal and ultrasound images. The researchers further evaluated the performance of machine learning models on the basis of detection accuracy, response time and feature extraction. The study also included literature work highlighting applications of machine learning and biosensors algorithms in breast cancer research and how effectively they have improved the process.16]This study gives a detailed overview of different stages of breast cancer classification using machine learning models. The researchers have implemented the models and further segmented the process into three stages namely that includes: data preprocessing, feature extraction, feature scaling, model building and classification. They have further investigated the impact of these algorithms in classification process of mammograms. In the end, the study concludes the process of evaluating performance of models in detection and classification of cancer cells on the parameters of sensitivity, specificity and exactness.17] This study primarily focuses on evaluating the performance of different classification models on a standard dataset containing data of breast cancer patients. It includes different classification models such as decision tree, support vector machine, logistic regression, naïve bayes and k nearest neighbors. We know that application of different classifiers yields different results as they vary in accuracy, precision and exactness. After applying different classification models on the dataset, the author examined a frequent change in accuracies and also different patterns in classification results. The research work includes comparative analysis of breast cancer classification results from different classification models. 18]In this study, the researcher has conducting and interesting hypothesis test of checking whether reduction in noise from research data leads to higher accuracies in classification models. The author has further compared several classification methodologies on breast cancer dataset in order to briefly understand the pattern of malignant and non-malignant segmentation. On implementation, the author was able to gain significant insights that includes higher efficiency and accuracy observed in classification models post removal of noise from the data which explains the role of dimensionality reduction in reducing noise before implementation of ML algorithms.19]This study focuses on implementing neural network classification models in order to classify the type of breast cancer into malignant or benign. The author primarily focuses of different breast cancer types that includes mucinous carcinoma, tubular adenoma, ductal carcinoma, lobular carcinoma, phyllodes tumor, adenosis and papillary carcinoma. On further implementation, the author was able to gain significant insights that included successful performance of classification models on the image dataset and a 97% accuracy malignant and nonmalignant classification.20]This study focuses on using a deep learning methodology for breast cancer classification and conduct a preliminary research based on the findings. The author has implemented several hybrid convolutional neural network algorithms on a standard breast cancer dataset and “mias” mammography image dataset. On further implementation, the author was able to gain significant insights that included successful performance of classification models with an accuracy of 97% in malignant and nonmalignant classification. The results clearly explained that a hybrid neural network model can also yield accurate prediction of cancer cells and also obtain their classification based on degree of malignancy

3 Research Methodology

While working on a research project, it becomes essential to follow a systematic methodology that can assist in fetching significant results. There are different methodologies available to conduct analysis, classification and research such as CRISP-DM, SEMA, KDD

etc. The research methodology followed in this project is the Knowledge discovery in databases methodology (KDD).

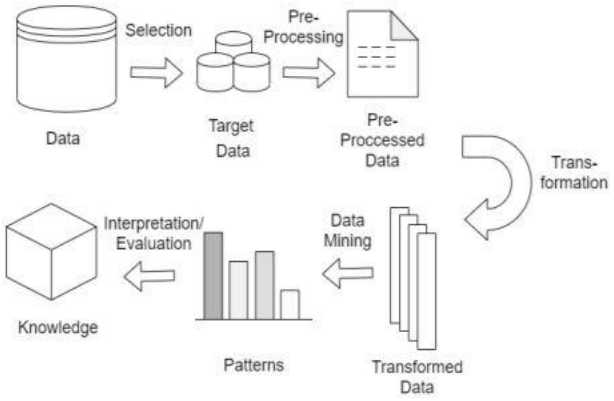

Figure 2: KDD Methodology

We can define KDD as an iterative methodology that involves data selection, pre-processing and transformation of data belonging to databases and also different other sources such images, data warehouses, text etc. It primarily focuses on applying data mining algorithms in order to fetch significant, valid and useful insights from the data. The KDD methodology is highly useful in breast cancer research as it helps in determining and discovering significant patterns and features from the data.

3.1 Data source information

In this study, we have used a public data source from Kaggle that contains information of breast cancer data of University of Wisconsin hospital from Dr William h. Walberg. This data included information of patients of Wisconsin hospital containing cancerous and non-cancerous breast mass. We have used this data in our classification, visualizations and several deep learning and machine learning models that included random forest classifier, k nearest neighbors, decision tree classifier and support vector machine.

Attributes	Description	Domain (values)
Diagnosis	tumour type	B -Benign or M - Malignant
radius	mean of distances from centre	decimal
texture	standard deviation of grey scale values	decimal
smoothness	local variations 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
perimeter	sum of length of all sides	decimal
area	region enclosed within the perimeter	decimal
compactness	(perimeter^2/area - 1.0)	decimal
symmetry	balanced proportions between two masses	decimal

Table 1: Attributes of dataset 1

Table 1 shows all the ten attributes in dataset 1 that were calculated by Wisconsin hospitals for cell nucleus of breast mass for each cell from a digitized image. All the features in the dataset are calculated from a digitized image of a fine needle aspirate of a breast mass. All these features define the attributes of the cell nuclei present in the image. It consists of 699 rows and 11 attributes in total. The class distribution of malignant and benign breast mass includes 357 benign and 212 malignant. Three measure results were calculated that includes mean, standard error and worst /largest of these features for each image that resulted in total 30 features.

3.2 Data pre-processing (includes data cleaning and removing outliers)

Data preprocessing can be defined as the process of transformation of raw data into clean and insightful format. In an ideal situation, we deal with databases present with raw data entries with missing values, noise, incomplete and incorrect data entries which makes it unfit for research or analysis process and also affects the performance of applied models. In this scenario, we need to implement a process to resolve these issues and make data ready for research and analysis. Data preprocessing plays a significant role and getting rid of inconsistencies, noise, missing values and incorrect entries in the dataset and makes it easier for interpretation and evaluation. We looked for such noise and inconsistencies in our Wisconsin dataset and found columns that contained incorrect data, outliers and null values. The insignificant data entries were dropped, outliers were removed and further the data was pushed for data exploration data visualization process.

3.3 Exploratory data analysis and data visualizations.

Exploratory data analysis can be defined as the process of investigating data and obtaining significant insights, information, patterns and conclusions from the data. It forms the basis of a dynamic data exploration process to make informed decisions from the gathered insights rather than making assumptions. It also involves understanding the nature of data by extracting mean, averages, maximum and minimum values. In our exploratory data analysis, we have also looked for inconsistencies, wrong data imputations, missing data entries and outliers in the dataset and resolved the errors for further analysis. we have tried to understand the nature of dataset by visualizing each dependent and nondependent variables and finding the correlations between them. We have primarily focused on studying the patterns of highly correlated features and how they are contributing in classification of tumors through box plots, scatter plots and histograms. Hence, our main goal is to thoroughly understand significant patterns in the data and use tools effectively to draw our conclusion and insights

3.4 Feature selection

Feature selection includes taking into account significant features that assist in achieving significant results. Sometimes, the data consists of many features and some of them might not be significant to fetch clear insights related to our problem statement. Training the model with unnecessary attributes might reduce the model's performance hence it becomes important to select necessary attributes that are highly co related to the target variable and help in obtaining required output and conclusions from the data because. The Wisconsin breast cancer data set consists of total 32 attributes which are quite high and might contain several insignificant or irrelevant features. The main question here is how to select significant features that can contribute in obtaining higher performance and hence we have obtained a

correlation plot of all the features in the data and tried to understand and analyze most significant features to implement in our models.

3.5 Model building

The stage of model building is the most crucial part as it involves selecting significant algorithms as per the data and that meet the objectives of our research. We have implemented nine machine learning models and one neural network model on our dataset. The classification algorithms include K-Nearest neighbors, support vector machine, logistic regression, random forest classifier, decision tree classifier, hyper parameter tuning, ada boost classifier, gradient boosting classifier, XGB boost classifier and artificial neural network model was implemented in this research.

1)**K- Nearest Neighbor (KNN)**: The K-Nearest Neighbor is a supervised learning algorithm and is extensively applied in regression and classification problems. The K-Nearest algorithm takes into account significant instances and features relevant to the target variable and further classifies them. Here, K refers to a numerical value and it executes the classification process by taking into account the Euclidean distance to k- nearest neighbor Thus, the K-nearest algorithm helps in interpretation of data.

2)**Random Forest Classifier**: Random Forest classifier can be referred as an ensemble learning model which can be implemented in both regression and classification problems. A random forest classifier is based on feature extraction and these features are clubbed together in K different subsets. It contains many decision trees and aims to deploy accurate and significant classification of data.

3)**Logistic Regression**: Logistic regression have been widely implemented in medical studies since early twentieth century and plays a significant role in predictive analytics as well. A logistic regression algorithm can be implemented when the data contains a target variable that is categorical or binary in nature.it can only be implemented when the target variable is limited to only two classes. The aim of our study is to classify the cancer into benign and malignant and hence logistic regression is implemented as the target variable is categorical.

4)**Support vector machine**: The support vector machine algorithm is a supervised learning algorithm that is extensively applied on the data for regression and classification purposes. The primary objective of this algorithm is to determine a hyperplane for classification of data points in N dimensions. The main work lies in figuring out the plane that maximizes the margin and later diversifies based on the feature numbers. Support vector machine has been extensively used in breast cancer classification and has delivered optimum performance.

5**) Decision tree classifier**: Decision tree algorithm is commonly used supervised machine learning algorithm that helps in regression and classification problems. It basically solves the problem by using a tree representation and the features in this structure are referred as decision nodes while output is referred as leaf nodes. It helps in prediction of target variable by taking into account other variables.

6)**Hyper parameter tuning**: Hyperparameter tuning refers to implementing a learning algorithm by selecting a set of hyperparameters. We can define a hyperparameter as a module argument which contains a certain value and it is assigned before the machine learning process. Hyperparameter tuning is majorly implemented to boost the performance of machine learning models.

7)**ADA boosting classifier**: An ADA boost algorithm is referred as Adaptive boosting algorithm and is basically a boosting technique implemented in machine learning as an ensemble methodology. It is referred as adaptive boosting as it involves re-assigning weights to each attribute such as assigning higher weights to attributes that are classified incorrectly.
8)**Gradient boosting classifier**: A Gradient boosting algorithm is a machine learning algorithm that is extensively used in regression and classification problems. It is basically constructed in a stage wise manner as it incorporates other boosting methods. It also involves optimization of differential loss functions and thus helps in boosting performance of the machine learning model.
9)**Xtreme gradient boosting** :XG boost is a machine learning algorithm extensively used in medical domain for classification and regression purposes. Using XG boost classifier in breast cancer classification can be beneficial for patients as well as medical professional as it will assist them by early cancer diagnosis whereas medics will be able to construct an early treatment plan for the patients.
10)**Artificial neural network model**: An artificial neural network model is inspired by biologically designed computer programs and is structured to stimulate the operations as human brain processes. They basically collect data by investigating several patterns and relationships in the information and contains several input and output layers. They contain as many as 10 hidden layers that assists in modifying the input layer in a way it can be used for output layer.

4 Design Specification

Figures 3 and 4 describe the step-by-step Network architecture of Breast cancer Diagnosis using machine learning and deep learning algorithms that will be carried out in our research. The primary objective of this study is to focus on developing adequate machine learning and deep learning models for early-stage prediction and classification of breast cancer. Our end result is classification of tumor into “malignant” and “benign”. All the process are further explained in detail.

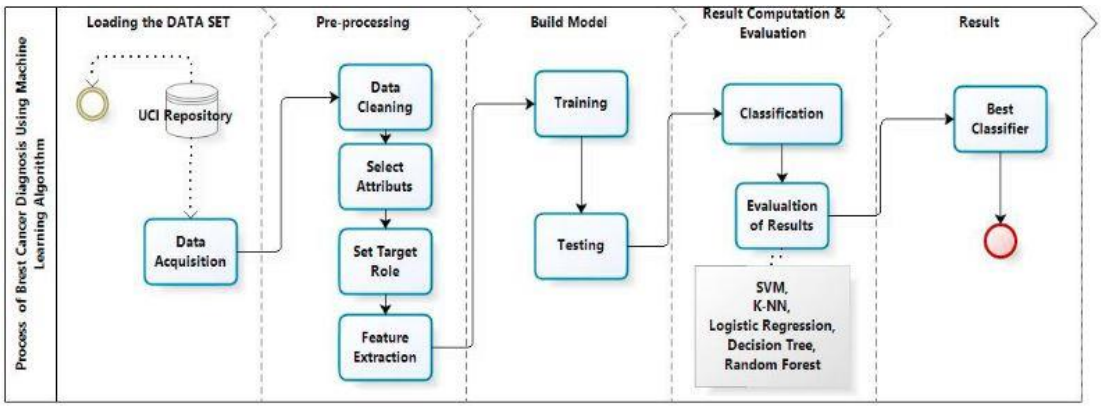

Figure 3: Network architecture of Breast cancer Diagnosis using machine learning algorithm

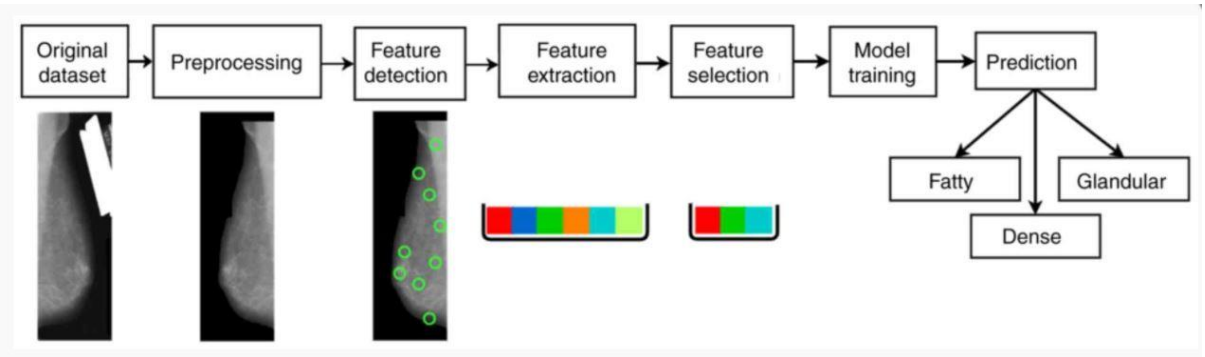

Figure 4: Network architecture of Breast cancer Diagnosis using Artificial neural networks

5 Implementation

5.1 Data Preparation:

Data preprocessing can be defined as the process of transformation of raw data into clean and insightful format. Data preprocessing plays a significant role and getting rid of inconsistencies, noise, missing values and incorrect entries in the dataset and makes it easier for interpretation and evaluation. We looked for such noise and inconsistencies in our Wisconsin dataset and found columns that contained incorrect data, outliers and null values. The insignificant data entries were dropped, outliers were removed and further the data was pushed for data exploration data visualization process. Below are the steps taken:

Basic data cleaning: We have dropped the columns "ID" and "unnamed" as they were not significant for our classification.

Missing values removal: The column "unnamed" had the most null values hence we have dropped the column and now the data is cleaned.

Outlier removal: We have detected the outliers present in the significant features and have removed them.

5.2 Data Transformation:

After data is cleaned, we have moved further in our process with data transformation as it is a necessary step in our module t o prepare the data for further classification operations. As mentioned in the above step, the Wisconsin breast cancer data set is cleaned and does not contain any outliers or missing values. We have transformed our target variable "diagnosis" from categorial to integer type i.e., from B and M to a factor binary number 0 and 1 for benign and malignant tumor. Also, for the artificial neural network model, the factor was transformed to numeric data type.

Feature Scaling: In our feature scaling, we have normalized the range of independent variables/features in the data using a standard scalar.

5.3 Data Exploration:

Data exploration can be defined as the process of investigating data and obtaining significant insights, information, patterns and conclusions from the data. It forms the basis of a dynamic data exploration process to make informed decisions from the gathered insights rather than making assumptions. It also involves understanding the nature of data by extracting mean, averages, maximum and minimum values. We have segmented our data exploration in three steps.

1)**Exploratory data analysis**: In our exploratory data analysis, we have also looked for inconsistencies, wrong data imputations, missing data entries and outliers in the dataset and resolved the errors for further analysis. we have tried to understand the nature of dataset by visualizing each dependent and nondependent variables and finding the correlations between them.

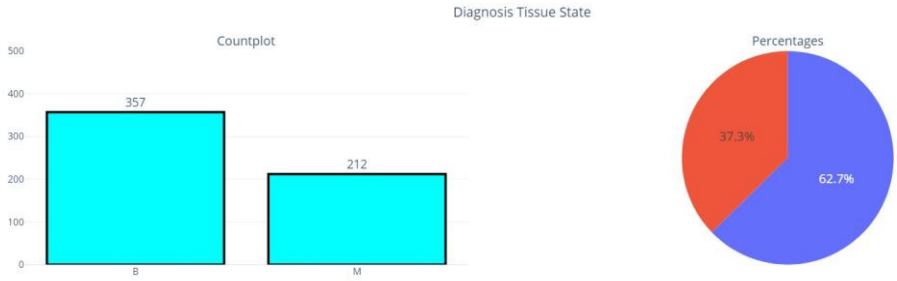

Figure 5: EDA

Exploring the target variable: We can observe our exploratory data analysis of our data in Figure 5 where B and M stands for Benign and Malignant tumors. Benign tumor is basically not cancerous as the cells grow very slowly are usually normal that is does not spread to other parts of the body or invade any nearby tissues. On the other hand, Malignant cancer is extremely cancerous and rapidly spreads to other parts of the body and can be fatal. In our analysis, we found that most of our data represents non-cancerous symptoms. Around 62.7% of our data indicates non-cancerous symptoms that is type B cancer while 37.3% of our data gives cancerous symptoms that is malignant cancer.

2)**Data Visualization**: We have primarily focused on studying the patterns of highly correlated features and how they are contributing in classification of tumors through box plots, scatter plots and histograms. Hence, our main goal is to thoroughly understand significant patterns in the data and use tools effectively to draw our conclusion and insights

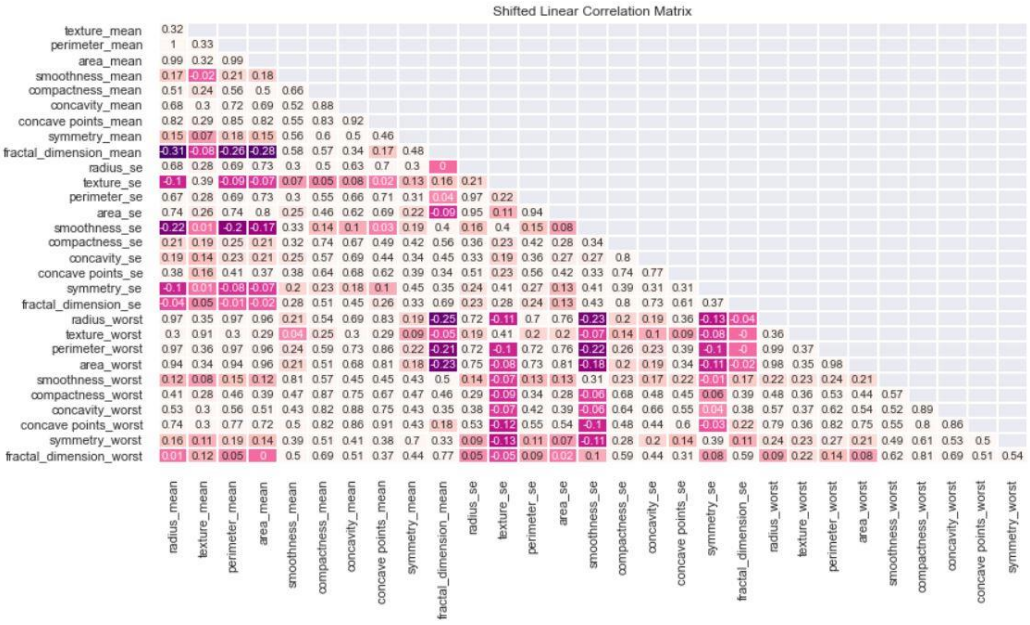

Figure 6: Shifted Linear Correlation matrix

In our shifted co relation matrix we can observe shifted variants with only relevant cases.
We can observe clear distinctions for malignant cancer among larger values of certain parameters and these distinctions further helped us in classification of cancer. We determined that the values that can be used for classification of breast cancer are mean values of cell radius, perimeter, area, compactness, concavity and concave points. We have also determined the attributes that were not successful contributors in our cancer classification and were not showing much preference in the diagnosis. They include the mean values of texture, smoothness, symmetry and fractal dimension.

3)**Group by analysis of both the tumors**:
a) **Analysis of radius and texture mean**: We further extended our analysis to highly co related features and understand their relationship with type M and type B cancer. Figures 7 and 8 indicates our exploratory data analysis of radius and texture mean and its relationship with our target variable. In our analysis with radius mean and target variable, we found that the radius of all cancer tumors in type M cancer is less than 18 and most of them fall under the range of 12-12.99. While on the other hand, the texture mean value of cancer tumors of type M and B exceeded to 19.99. We further concluded that all the tumors of type M showed higher texture and radius mean values than type B tumor and also the texture mean was considerably higher than radius mean for the both the tumors.

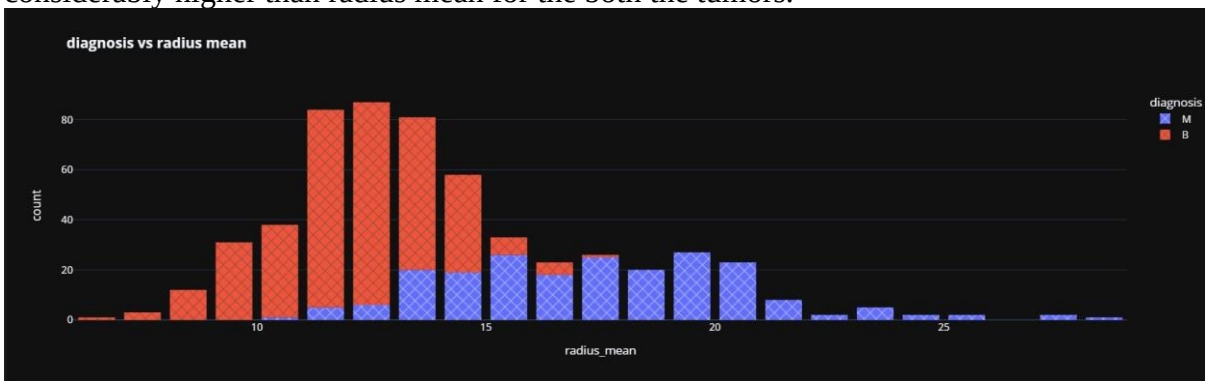

Figure 7 : Analysis of radius and texture mean

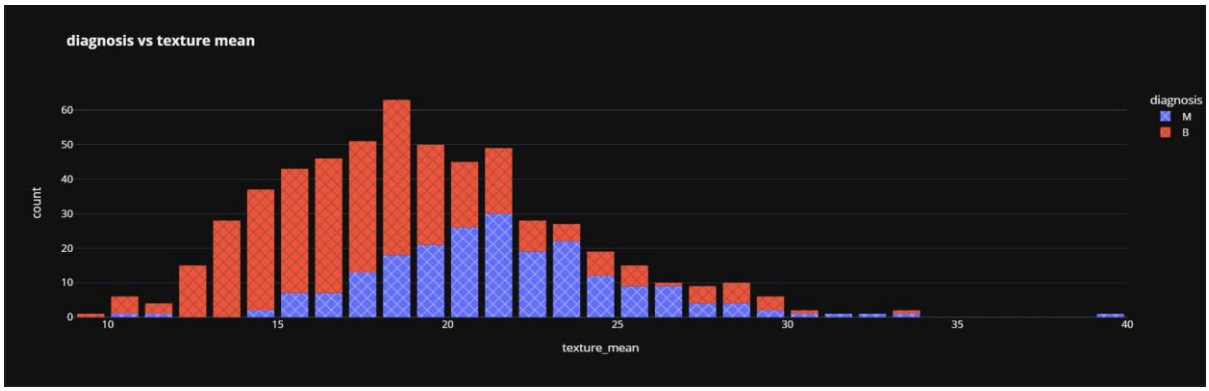

Figure 8 : Analysis of radius and texture mean

b) **Analysis of perimeter and area mean**: We further implemented our exploratory data analysis of perimeter and area mean and its relationship with our target variable. In our analysis with perimeter mean and target variable, we found that the perimeter mean is less than 110-115 for most of the cancer tumors of type M and falls under the range of 75-80.

While on the other hand, the area means value of cancer tumors of type B exceeded to 700 and a similar trend was observed for texture mean value of cancer tumors of type M. We further concluded that all the tumors of type M showed higher perimeter and area mean values than type B tumor and also the perimeter mean was considerably less than area mean for both the tumors.

c)**Group by analysis of both tumors**: In our final analysis of both tumors, we looked for attributes that exceeded with higher mean values from each other. We concluded that the mean values for fractal dimension, texture, smoothness and symmetry were comparatively higher in Tumor B than tumor M

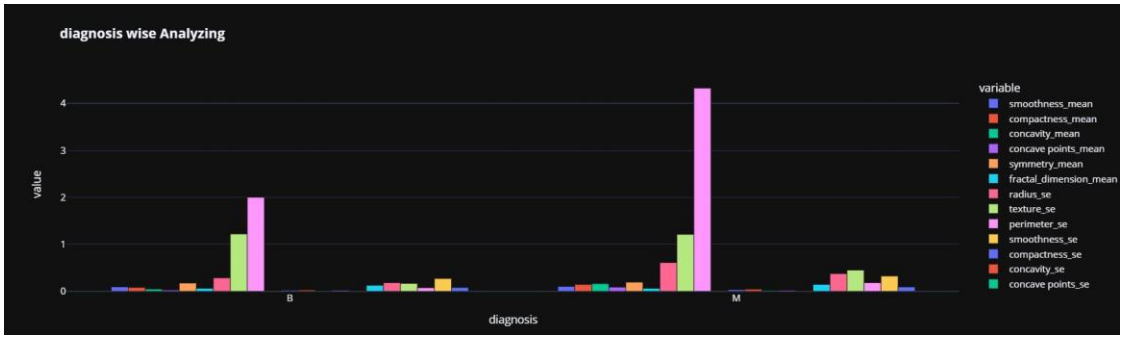

Figure 9 : Groupy analysis of both tumors

5.4 Encoding

We have encoded our target variable to numeric data type using a Label encoder. We have transformed our target variable "diagnosis" from categorial to integer type i.e., from B and M to a factor binary number 0 and 1 for benign and malignant tumor. Also, for the artificial neural network model, the factor was transformed to numeric data type**.**

5.5 Feature selection

The Wisconsin breast cancer data set consists of total 32 attributes which are quite high and might contain several insignificant or irrelevant features. The main question here is how to select significant features that can contribute in obtaining higher performance and hence we have obtained a correlation plot of all the features in the data and tried to understand and analyze most significant features to implement in our models. We have further selected symmetry_mean,symmetry_se,compactness_worst,texture_se,concave.points_se,smoothness _worst,smoothness_se,concavityworst,concavity_se,concave.points_worst as our final features for the classification model.

5.6 Classification Models:

The stage of model building is the most crucial part as it involves selecting significant algorithms as per the data and that meet the objectives of our research. We have implemented nine machine learning models and one neural network model on our dataset. The classification algorithms include K-Nearest neighbors, support vector machine, logistic regression, random forest classifier, decision tree classifier, hyper parameter tuning, ada boost classifier, gradient boosting classifier, XGB boost classifier and artificial neural network model was implemented in this research. We have used python programming language to carry out our analysis.

1)**K- Nearest Neighbor (KNN)**: We have implemented the K-nearest neighbor classification model using the kneighbors classifier and selected our features using the feature selection algorithm. We have trained our model on the training set and further tested its efficiency on the test set to determine its performance in classification of target variable. We have also determined the error rate of the model, obtained the accuracy with accuracy score and also determined the performance of the model based on sensitivity, specificity, false positives and false negatives.
2)**Random Forest Classifier**: We have implemented the random forest classification model using the randomforestclassifier and total 40 n_estimators fed in our forest variable. We have implemented feature extraction in our model and selected key features for training by using feature selection algorithm. We have trained our model on the training set and further tested its efficiency on the test set to determine its performance in classification of target variable. We have also determined the error rate of the model, obtained the accuracy with accuracy score and also determined the performance of the model based on sensitivity, specificity, false positives and false negatives.
3)**Logistic Regression**: We have implemented the Logistic regression model using the logistic regression function and fed our selected features in the logmodel variable. We have implemented our model on the target variable because it suits the requirement of logistic regression as our target variable is binary in nature and has two classes M and B. We have trained our model on the training set and further obtained the accuracy with accuracy score and also determined the performance of the model the pred () function. We have also checked for sensitivity, specificity, false positives and false negatives.
4)**Support vector machine**: Support vector machine has been extensively used in breast cancer classification. We have implemented svm model using the SVC () function and fed our selected features in the svm variable. We have tried to determine a hyperplane for classification of data points in N dimensions. We have worked on figuring the plane that maximizes the margin and later diversifies based on the feature numbers. We have further trained our model, checked for performance of the model on terms of accuracy, sensitivity and specificity.
5) **Decision tree classifier**: We have implemented decision tree classification model using the decision tree classifier and fed our selected features in the predictor variable “dtc”. We have trained our model on the training set and further obtained the accuracy with accuracy score and also determined the performance of the model the pred () function. We have also checked for sensitivity, specificity, false positives and false negatives.
6)**Hyper parameter tuning**: We have defined a hyperparameter as a module argument which contains a certain value assigned it before the machine learning process. We have fitted 5 folds for each 512 candidates and fed them in the grid_search function.We have further determined the best parameters contributing models’ performance also obtained best training score along with accuracy, f1 score, sensitivity and specificity.
7)**ADA boosting classifier**: We have implemented ADA model using the ADA boostclassifier and fed our selected features in the ada variable. We have fitted 10 folds cross validation for each 60 candidates totaling 600 fits and fed them in the grid search function. We further fed our predictor variable with 200 best estimators, learning rate of 0.001 and base estimator “dtc”. We have further determined the best parameters contributing models’ performance and also obtained best training score along with accuracy, f1 score, sensitivity and specificity.
8)**Gradient boosting classifier**: We have implemented gradient boosting model using the gradient boosting classifier and fed our selected features in predictor variable “gd”. We have worked on optimizing the differential loss functions to help in boosting performance of the machine learning model.

9)**Xtreme gradient boosting** : XG boost classifier in breast cancer classification can be beneficial for patients as well as medical professional as it will assist them by early cancer diagnosis whereas medics will be able to construct an early treatment plan for the patients.
We have implemented Xtreme gradient boosting model using the XGBoost classifier and fed our selected features in predictor variable "xgb". We have further determined the best parameters contributing models' performance also obtained best training score along with accuracy, f1 score, sensitivity and specificity.
10)**Artificial neural network model**: We have converted our target variable into numeric data type before implementing our artificial neural network model. We have trained our model using keras package and passed our target variable along with predictor variables by setting the linear output to false. We have trained the model with less number of hidden layers as the model yields optimum performance with lower hidden layers.

6 Evaluation

After implementing the classification techniques, we have evaluated the performance of our models on the basis of different parameters. The primary objective of our study is to provide an overview of patterns that we have obtained from interpretation of our results. Below are the evaluation metrics we have used to evaluate the performance of our classification models.

1)**Accuracy**: Accuracy can be defined as the ratio of number of predictions made by the model to the total number of predictions. Formula to compute accuracy is given below
Accuracy = True positive + True Negative / Total
2)**Sensitivity**: Sensitivity is referred as True positive rates. Sensitivity can be defined as the ratio of correctly classified positive classes in the data to all the positive classes .
Sensitivity = True positive / False Negative + True positive
3)**Specificity**: Specificity is referred as false positive rate. Specificity can be defined as the ratio of wrongly classified negative classes in the data to all negative classes.
Specificity = False positive / False Positive + True Negative

These parameters form the basis of evaluation metrics of our classification models, while we will be judging the best performing models based on accuracy and sensitivity as in breast cancer data it is essential to identify the number of patients that are correctly identified with cancer among all the patients.

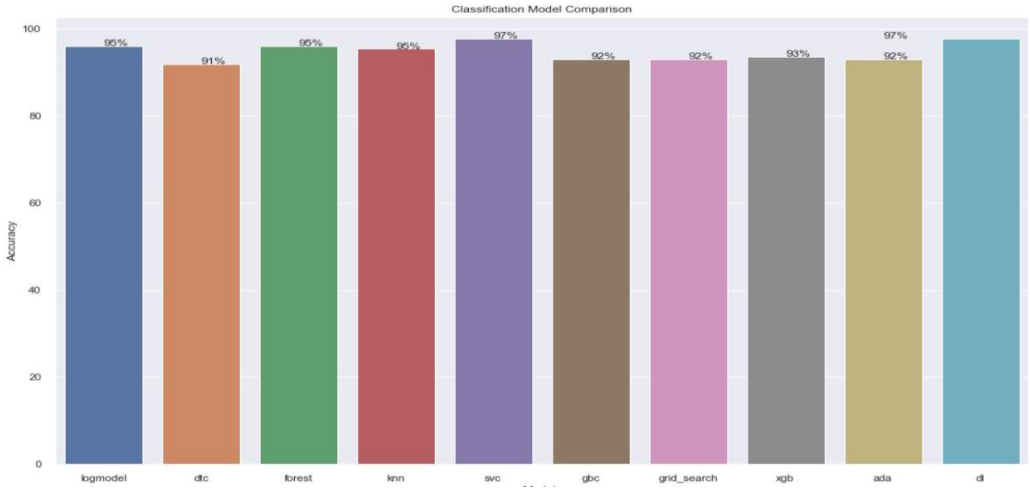

Figure 10 : Accuracy percentage of classification models

Figure 10 shows the comparison of accuracy percentages of KNN, support vector machine, decision tree classifier, random forest classifier, gradient boosting classifier, hyper parameter tuning, extreme gradient boosting classifier, logistic regression, ada boosting classifier and artificial neural network model. As we can evaluate from the comparison plot, the best performing models are support vector machine, gradient boosting classifier and artificial neural network model with the highest accuracy percentage of 97% while the lowest performing classification model was Decision tree classifier with accuracy of 91%.

Models	Sensitivity%	Specificity%
KNN	90.90%	98.09%
LR	95.45%	99.04%
Forest	93.90%	97.14%
DTC	90.90%	92.38%
GBC	87.87%	96.19%
XGB	90.90%	95.23%
ADA	87.87%	96.19%
SVC	95.45%	99.04%
HP	87.87%	96.19%
DL	96.82%	98.14%

Table 3: Sensitivity and specificity percentage of classification models.

Table 3 shows the performance of classification models with respect to their sensitivity and specificity. For sensitivity, the models are judged based on True positive rate that means the model which has correctly classified maximum number of true positives. The model with the highest sensitivity is deep learning model followed by linear regression and support vector machine

6.1 Experiment 1 : Choosing K- values for KNN Algorithm

Our first experiment involves analysing the performance of KNN algorithm for different k values. One of the most important task in determination of performance of classification model is choosing the correct k value. Similarly, in our experiment 1 , we have analysed the accuracy and sensitivity of KNN model with respect to K values 10,15,20,25 and 30.

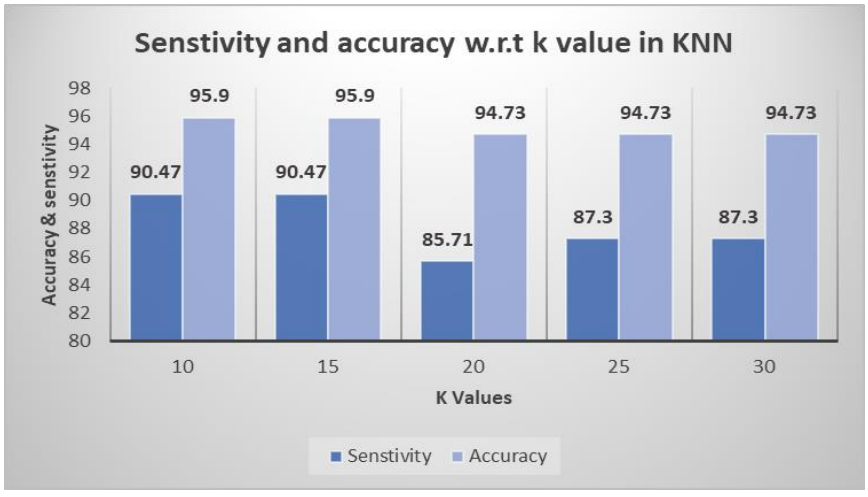

Figure 11 : Sensitivity and accuracy w.r.t k value in KNN

The table for accuracy and sensitivity percentage of KNN Model with respect to following K values is given in figure 11. We can observe from the table that the highest accuracy and sensitivity is obtained at the k value 10 and 15 while the lowest sensitivity and accuracy is obtained at k value 20. Hence, we can consider K value 20 for optimum performance of KNN model.

6.2 Experiment 2: Selecting hidden layers for ANN model

Our second experiment involves analysing the performance of Artificial neural network model with number of hidden layers. An ANN model basically contains an input layer, hidden layer and output layer. We need to specify the number of hidden layers else the model assumes one default layer to resume the process. We know that the performance of the model fluctuates with respect to the number of hidden layers, Hence, in our experiment we will be analysing the performance of the ANN model based on hidden layers 2,3,4,5 and 6 to figure out the best hidden layer for ANN algorithm. The table for sensitivity percentage of ANN Model with respect to following hidden layers is given below. We can observe from the table that the highest sensitivity is obtained at the hidden layer 2 while the lowest sensitivity is obtained at hidden layer 6. Hence, we can consider hidden layer 2 for optimum performance of ANN model.

HIDDEN LAYER	SENSITIVITY %
2	96.82
3	96.1
4	95.4
5	95.4
6	94.11

Table 4: Sensitivity percentage at hidden layers

6.3 Discussion

In our research so far, we have determined the severity of breast cancer by implementing nine machine learning models and one deep learning model using Wisconsin breast cancer dataset. The severity of breast cancer is identified based on the sensitivity percentage of K nearest neighbours, logistic regression model, random forest classifier, decision tree classifier, gradient boosting classifier, extreme gradient boosting classifier, ada boosting classifier, support vector machine, hyperparameter tuning model and deep learning model. We have followed the KDD methodology and extensively performed data exploration, data preparation, data transformation, data analysis, encoding, feature scaling, feature selection of predictor attributes and model implementation on our dataset. We can observe the performance of classification models with respect to sensitivity and deep learning model performed best with 96.82% sensitivity followed by logistic regression model and support vector machine model with 95.45% sensitivity in figure 12. While gradient boosting, ada and hyperparameter tuning prove to be the lowest in performance with respect to sensitivity.

Hence, we can now conclude that deep learning, logistic regression and support vector machine models perform best in determining the severity of breast cancer. We have checked the performance of the model by implementing 10-fold cross validation and noted the evaluation metrics. We have used the k value as 9 here because it was widely used in most research papers. We have further conducted an experiment to determine the performance of our model by analysing the sensitivity percentage at different k values. However, we didn't notice much difference in the functioning of classification models on 10 cross validation and 20 cross validations. Our evaluation metrics includes accuracy, sensitivity and specificity but we have evaluated the performance of our classification models based on the main metric of sensitivity. The main reason for choosing sensitivity as our main evaluation metric is due our medical data, the need of knowing the successful classification model for determining tumour type becomes extremely essential for treating a disease. Sensitivity is basically computed by taking into consideration the true positives correctly identified by the model out of all the true positives. Hence, sensitivity becomes an essential metric in determining the severity of breast cancer

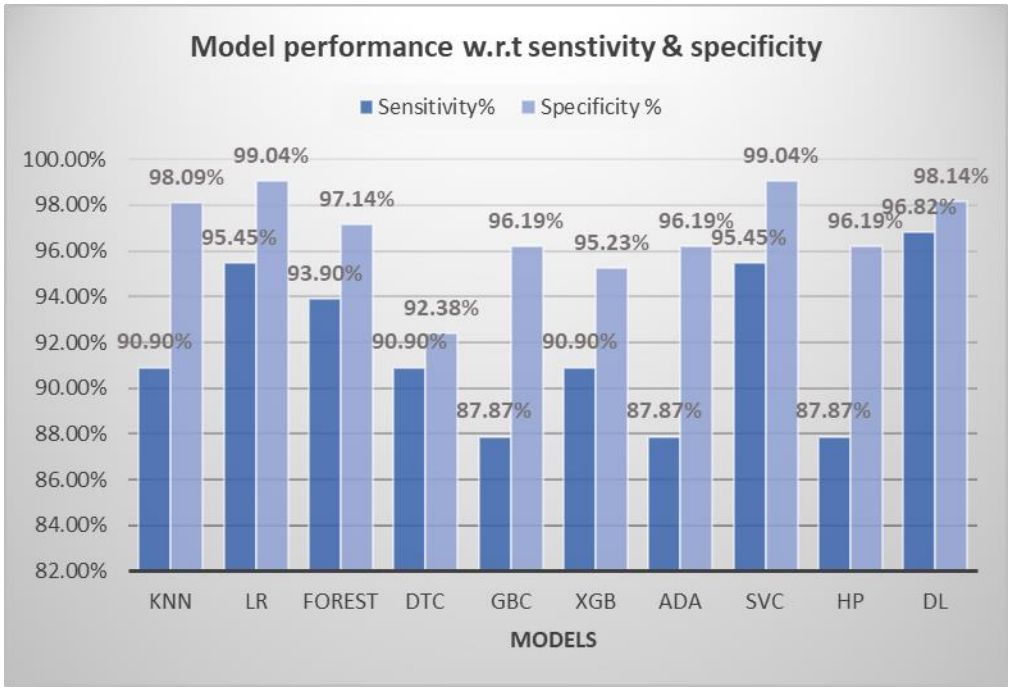

Figure 12: Model performance with respect to sensitivity and specificity

7 Conclusion and future work

The major objective of our study is to obtain the best performing model that can determine the severity of breast cancer in patients. We further implemented nine machine learning models and one deep learning model on Wisconsin breast cancer dataset and analysed the performance of classification models based on their sensitivity score. We also conducted two experiments to analyse the performance change in KNN and ANN models with respect to different k values and hidden layers and were able to fetch significant insights. The results from our evaluation metrics showed highest sensitivity percentage in deep learning model with 96.82% followed by logistic regression and support vector machine model with 95.45% sensitivity. On the basis of the performance results, ANN model can be useful in detection of benign and malignant tumour and assist medical professionals in early-stage

diagnosis of breast cancer. With current scenario of increasing fatalities due to breast cancer, it is to implement a robust classification model that can assist medical professionals in early-stage cancer diagnosis and save many lives. Our Future work will be dependent on taking this research further and developing a breast cancer classification system from a standpoint of community care with an ambition of improving the healthcare system in breast cancer diagnosis. The future work is largely inclusive of bringing more affordability, efficiency and advancement in breast cancer prediction system.

8 Acknowledgement

I hereby express my gratitude towards professor Dr. Bharathi Chakravarthi for his continuous guidance and teachings throughout the research phase.I am grateful to my friends and family for their constant support and motivation in my research journey.Hence, I hereby declare that the information and all the content pertaining to the research project that is not my contribution will be referenced appropriately.

9 References

Avani Ahuja, L. A.-Z. A. K., 2021. *Application of noise-reduction techniques to machine learning algorithms for breast cancer tumor identification,,* s.l.: Computers in Biology and Medicine, volume 135.

Bo-Yang Zhou, L.-F. W. H.-H. Y. T.-F. W. T.-T. R. C. P. D.-X. L. H. S. L.-P. S. C.-K. Z. H.-X. X., November 2021. *Decoding the molecular subtypes of breast cancer seen on multimodal ultrasound images using an assembled convolutional neural network model: A prospective and multicentre study,,* s.l.: EBioMedicine.

Debendra Muduli, R. D. B. M., 7 july 2021. *Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network based approach,* s.l.: Biomedical signal processing and control.

Enas M.F. El Houby, N. I. Y., 28 july 2021. *Malignant and nonmalignant classification of breast lesions in mammograms using convolutional neural networks,* s.l.: Biomedical signal processing and control.

Essam H. Houssein, M. M. E. A. A. A. P. N. S., April 2021,114161. *Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review.* s.l.:Expert Systems with Applications.

Fung Fung Ting, Y. J. T. K. S. S., November 2018. *Convolutional neural network improvement for breast cancer classification,,* s.l.: Expert Systems with Applications.

G. Meenalochini, S. R., 2021. *Survey of machine learning algorithms for breast cancer detection using mammogram images,* s.l.: Materials Today: Proceedings,volume 37,part 2.

Jyoti Parashar, S. M. R., 2020. *Breast cancer images classification by clustering of ROI and mapping of features by CNN with XGBOOST learning,,* s.l.: materials today:proceedings.

Leila Abdelrahman, M. A. G. F. C.-M. M. A.-M., 9 february 2021. *Convolutional neural networks for breast cancer detection in mammography: A survey,* s.l.: computers in biology and medicine.

Mazin Abed Mohammed, B. A.-K. A. N. R. D. A. I. M. K. A. G. S. A. M., August 2018. *Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images,* s.l.: Computers & Electrical Engineering.

Nandita Goyal, M. C. T., 17 december 2020. *Breast cancer classification and identification using machine learning approaches,* s.l.: Materials Today: Proceedings.

Nandita Goyal, M. C. T., 17 December,2020. *Breast cancer classification and identification using machine learning approaches,* s.l.: materials proceedings.

Nosayba Al-Azzam, I. S., 2021. *Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer,* s.l.: Annals of medicine and surgery, volume 62.

Raquel Sánchez-Cauce, J. P.-M. M. L., 16 March 2021. *Multi-input convolutional neural network for breast cancer detection using thermal images and clinical data,,* s.l.: Computer methods and programs in bio medicine.

Saikat Islam Khan, A. S. R. K. M. H. A. R., 17 august 2021. *MultiNet: A deep neural network approach for detecting breast cancer through multi-scale feature fusion,* s.l.: Journal of King Saud University - Computer and Information Sciences.

Sawssen Bacha, O. T., January 2022, 110233. *A novel machine learning approach for breast cancer diagnosis,,* s.l.: Measurement.

Sertan Kaymak, A. H. D. U., 2017. *Breast cancer image classification using artificial neural networks,* s.l.: procedia computer science.

Xin Yu Liew, N. H. J. C., 2021. *An investigation of XGBoost-based algorithm for breast cancer classification,* s.l.: Machine Learning with Applications,volume 6.

Yash Amethiya, P. P. S. P. M. S., 22 october 2021. *Comparative Analysis of Breast Cancer detection using Machine Learning and Biosensors,* s.l.: Intelligent medicine.

Yu-Dong Zhang, S. C. S. D. S. G. J. M. G. S.-H. W., December 2020. *Improved Breast Cancer Classification Through Combining Graph Convolutional Network and Convolutional Neural Network,,* s.l.: Information Processing & Management.