

# **Leveraging Data Science For Predicting Breast Cancer Outcomes Using The Wisconsin Breast Cancer Dataset**



**BIRMINGHAM CITY  
University**

## **Student Name & SID:**

Muhammad Irfan  
23173372

## **Supervisor**

Dr. Abdulrahman Al Sewari

## **Submission Date:**

September 2024

---

## **Acknowledgments**

I would like to express my deepest gratitude to my supervisor, Professor Abdulrahman Alsewari, for his invaluable guidance, support, and encouragement throughout the course of this research. His expertise and insights have been instrumental in shaping the direction of this thesis, and his continuous feedback has been essential to my academic and personal growth. I am sincerely thankful for his patience, understanding, and the opportunities he provided me to explore new ideas. This thesis would not have been possible without his mentorship and support.

---

## Abstract

Breast cancer is one of the main problems for oncology representatives; it affects women anywhere in the globe. In enlisting technological interventions, it is noteworthy to stress that early and accurate diagnosis is pivotal to enhancing treatment results. The existing approaches to developing forecasts for breast cancer prognosis leave little room for the use of non-linear approaches of machine learning in the contingent of the variable. In this context, the presented work seeks to fill these deficiencies by using more sophisticated machine learning methods, such as logistic regression, random forest, Support Vector Machines, and decision tree models, to establish the predictive models with the help of the Wisconsin Breast Cancer Dataset Wisconsin Breast Cancer Dataset Wisconsin Breast Cancer Dataset (WBCD). These models are evaluated, analyzed, and tested on their precision, accuracy, recall score, F1 score and ROC-AUC score, to determine their efficiency, consistency, and efficacy in predicting and detecting breast cancer outcomes. The research also plans to use other approaches from the machine learning algorithm's domain such as deep neural networks to increase accuracy in prognosis and facilitate decision-making. Expected outcomes may include obtaining a large-scale and precise model that can be implemented into clinical practice, which may improve the quality of treating patients as well as optimize the modes of applying certain treatments.

The importance of the study resides in the possibility of changing the current situation and academic approaches used to develop a prognosis of breast cancer in lowering the cost of health care and raising the survival rate of the patients. This study will also exhibit the potential and likeness of bringing data science into practice to increase and enhance the chances of breast cancer prognosis and diagnosis and support clinical decision-making, giving a platform to doctors where they can easily interpret and detect breast cancer symptoms by previous data-driven research and machine intelligence. This study will play a pivotal role in bridging the gap between science and technology and will offer a platform and insights into the development of more effective, technologized, modern, and data-driven healthcare solutions.

# Contents

	<b>5</b>
<b>1 Chapter-1 Introduction</b>	<b>6</b>
1.1 Introduction: . . . . .	6
1.2 Background Research . . . . .	6
1.3 Problem Statement . . . . .	6
1.4 Research Questions . . . . .	7
1.5 Aim . . . . .	7
1.6 Objectives . . . . .	7
1.7 Significance of the Study . . . . .	7
1.8 Nature of challenges . . . . .	8
1.9 Overview of the research . . . . .	8
<b>2 Chapter – 2 Literature Review</b>	<b>9</b>
2.1 Current State of Breast Cancer Predictions . . . . .	9
2.2 Gaps in Current Research . . . . .	12
<b>3 Chapter - 3 Research Methodology</b>	<b>14</b>
3.1 Methodology . . . . .	14
3.2 Business Understanding . . . . .	15
3.3 Data Understanding . . . . .	15
3.4 Data Preparation . . . . .	16
3.5 Modelling . . . . .	16
3.5.1 Regression Models . . . . .	16
3.5.2 Support Vector Machines (SVM) . . . . .	17
3.5.3 Decision Trees . . . . .	17
3.6 Deployment . . . . .	17
3.7 Overall Process Flow . . . . .	17
<b>4 Chapter – 4 Results</b>	<b>18</b>
4.1 Data Understanding . . . . .	18
4.2 Data Preparation . . . . .	20
4.2.1 EDA: Exploratory Data Analysis . . . . .	20
4.3 Data Preprocessing . . . . .	24
4.4 Modeling . . . . .	24
4.4.1 Logistic Regression . . . . .	24
4.4.2 Support Vector Machine (SVM) . . . . .	25
4.4.3 Training Accuracy: 97 % . . . . .	26
4.4.4 Testing Accuracy:97.08% . . . . .	26
4.4.5 Decision Trees . . . . .	26
4.5 Cross Validation . . . . .	27
4.6 Evaluation . . . . .	28
4.7 Discussion of Results . . . . .	30
<b>5 Chapter – 5 Conclusion and Future Works</b>	<b>30</b>
5.1 Limitation . . . . .	30
5.2 Achievements . . . . .	31
5.3 Future Work . . . . .	31
<b>6 References</b>	<b>32</b>

## List of Tables

1	Summary of Literature Review Research Papers . . . . .	13
2	Model Performance Metrics . . . . .	30

## List of Figures

1	CRISP-DM Methodology (Bachu & Anuradha, 2019) . . . . .	15
2	Flow of Research . . . . .	18
3	Dataset Attributes . . . . .	19
4	Dataset Columns . . . . .	19
5	Bar Plot Distribution benign . . . . .	20
6	Pair Plot All attributes . . . . .	21
7	Dataset Correlation Matrix . . . . .	21
8	Scatter Plot Texture and radius mean . . . . .	22
9	Scatter Plot area and perimeter mean . . . . .	22
10	Histogram . . . . .	23
11	Pie Chart benign vs Malignant . . . . .	23
12	data Pre processing . . . . .	24
13	Logistic Accuracy . . . . .	25
14	SVM Accuracy . . . . .	26
15	Decision Tree Accuracy . . . . .	27
16	Comparison of cross-validation scores . . . . .	27
17	Logistic Regression Report . . . . .	28
18	SVS Classification Report . . . . .	28
19	Decision Tree Classification Report . . . . .	28
20	Confusion Matrix Logistic Regression . . . . .	29
21	Confusion Matrix SVM . . . . .	29
22	Confusion Matrix Decision Tree . . . . .	29

---

## Acronym List

**BC** Breast Cancer.

**DL** Deep Learning.

**EDA** Exploratory Data Analysis.

**FDT** fuzzy decision trees.

**HDT** hybrid decision trees.

**MDL** Minimum Description Length.

**ML** Machine Learning.

**PCA** Principal Component Analysis.

**RBF** Radial basis function.

**SVM** Support Vector Machines.

**WBCD** Wisconsin Breast Cancer Dataset.

# 1 Chapter-1 Introduction

## 1.1 Introduction:

Breast cancer is undoubtedly one of the largest health concerns in modern-day society given the many women who are affected every year. This is only next to lung cancer and is a leading cancer among women influencing society and health facilities significantly. Even though impressive progress in medical treatment has been made, the disease has touched many research domains since the need to develop diagnostic tools and methods as well as predictive models essential for its early identification and proper patient management. Machine learning presents a great opportunity to revolutionize breast cancer prognosis through sophisticated data analysis techniques because it can handle large and complicated datasets. The essence of this study is to utilize the Wisconsin Breast Cancer Dataset WBCD in designing predictive models that significantly enhance diagnostic precision and treatment effectiveness (Aamir et al., 2022).

## 1.2 Background Research

It is widely evident that breast cancer remains among the most prominent global threats to women's health; in 2024, it will become the most transpiring malignant neoplasm in women, with a rate of 24.9% of total diagnoses among women (Ahsan et al., 2022). While previous approaches are not entirely without merit, they tend not be as successful as they could be in laying the foundation for early detection and prognosis of illness conditions that are critical in the case of any patient.

Machine learning techniques can be utilized for modeling and analyzing multi-dimensional datasets where there could be minute variations that could not be easily captured by other statistical tests. These models can foresee the outcomes more accurately and within less time, thus offering crucial information regarding disease progression and its response to treatment at much earlier phases (Naji et al., 2021; Islam et al., 2020).

This research work has provided evidence that different machine learning models, in combination with patient data, can lead to better prediction of breast cancer outcomes. For instance, deep neural 3 networks are more accurate compared to traditional models in predicting the metastasis of histopathological images of early breast cancer cases (Alshayeji et al., 2022).

## 1.3 Problem Statement

The current forecasting methods used for breast cancer prognosis, though useful in a number of ways, are still not able to make good use of the huge potential that machine learning has in oncology. There is a great deal of rich data and yet many of the existing models are unable to capture its information, most notably because they do not have an appropriate technique for efficient processing and analysis of complex datasets with high dimensionality. Such models are limited in their robustness and extendibility as improved machine learning methods could have done (Albadr et al., 2024).

The purpose of this study is therefore to deal with these deficiencies by designing advanced machine-learning algorithms that can benefit from the rich information provided by the Wisconsin Breast Cancer Dataset. It contains detailed attributes concerning biopsies on breast cancer which forms an important dataset that can be applied to design other algorithms for predicting outcomes more accurately (Ngwa et al., 2022)

## 1.4 Research Questions

- How accurate and reliable are different machine learning methods for predicting breast cancer outcomes using the Wisconsin Breast Cancer Dataset?
- What are the important aspects of the Wisconsin Breast Cancer Dataset that have the greatest influence on the predictive accuracy of breast cancer outcomes?
- Can a machine learning model based on the Wisconsin Breast Cancer Dataset improve clinical decision-making for breast cancer treatment?

## 1.5 Aim

The aim of this study is to address these shortcomings by designing advanced machine learning algorithms that leverage the rich information provided by the Wisconsin Breast Cancer Dataset. This dataset contains detailed attributes related to breast cancer biopsies, offering valuable data that can be utilized to develop algorithms capable of predicting outcomes with greater accuracy.

## 1.6 Objectives

The study will, in the main, have four objectives:

- Critically analyzing and evaluating the efficacy of various machine learning algorithms in predicting breast cancer outcomes through the Wisconsin Breast Cancer Dataset. This will involve comparing traditional statistical models to advanced machine learning approaches such as convolutional neural networks and ensemble methods (Jabeen et al., 2022).
- Determining key predictive features within the Wisconsin Breast Cancer Dataset and their association with patient outcomes. This will entail examining feature importance across different models to identify which data attributes have the most significant effect on the accuracy of prognosis predictions (Chen et al., 2015).
- Developing a scalable and accurate model that improves precision in prediction while seamlessly integrating into clinical decision-making processes. The intention of this model is to offer oncologists a reliable tool for individualizing treatment plans based on patients to optimize therapeutic outcomes (Ahsan et al., 2022).
- Evaluating the performance of the proposed system through comprehensive metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to ensure its reliability and effectiveness in clinical settings (Karabatak & Inc, 2009).

## 1.7 Significance of the Study

The main purpose of this research is to make breast cancer prognosis different from what it was before. The study intends to have an impact in oncology through enhancing accuracy in prediction models and thus, a tool that can be used globally for improved management of breast cancer's treatment outcomes. With better forecasting models, lives will not only be saved but also more optimized treatment protocols, lower healthcare costs as well as minimized psychological burden on both patients and their families (Alshayegi et al, 2022).



Ethics is a complex field that gained much prominence with the rise of Big Data. The historical development of ethics is mapped out by Marczyk et al (2005), along with a survey on several ethical codes. The UK Data Service provides guidelines for ethical research in relation to Big Data specifically (UK Data Service, 2020).

## 1.8 Nature of challenges

Although we have our aim and objectives set, there are several directions remaining unclear when it comes to the implementation of Machine Learning and Deep Learning for detecting breast cancer.

Despite the high accuracy, deep learning models are intrinsically opaque, that is, they are “black box,” so it is challenging to decipher the results of the model. Methods that enhance the explicability of DL models must be introduced to ensure the models’ adoption by clinicians (Dietterich, 2000).

The implementations of many Machine Learning and Deep Learning models are done separately and with little actual interaction with real-world clinics. There is a need to investigate how such models can best be incorporated to complement existing clinical practice by offering concrete information and assistance to clinicians and others (Chen et al., 2015).

## 1.9 Overview of the research

The research is divided into 5 different chapters which are the introduction to the research, literature review, methodology, and experiments. The first chapter is an introductory chapter that encloses the purpose of the research by stating the problem statement, aim, and objectives (Albadr et al., 2024). Discussing the background of the domain which is breast cancer detection and why is it necessary to produce a predictive model. Moreover, research problems and the nature of the challenges of the research have been identified from previous studies to obtain an understanding of the difficulties in performing this research (Ahsan et al., 2022).

The second chapter describes the literature review to research similar previous research papers and studies that are relevant to identify the current state-of-the-art contributions (Salzberg, 1994). Thus, a systematic review of the general domain in healthcare and the application of machine learning and deep learning in the detection of breast cancer is performed to assess the limitations and possible future works that could be applied to this research. Additionally, a table consisting of all the papers reviewed in this research is made to give a clear view of what are the algorithms, processes, and future works of the research (Bhinder et al., 2021).

The third chapter covers the approach that this project has taken. Thus, an in-depth discussion of the methodology used will show how each phase of the methodology reflects on this research as well as the proposed model that is used for the research (Cai et al., 2018). An explanation of the three algorithms used in this research is provided in this chapter as well.

The fourth chapter is the experiments chapter which encompasses the experiments performed such as data preparation, pre-processing, transformation, and model training (Karabatak & Ince, 2009). The results of each of the models will be evaluated using suitable evaluation metrics to conclude which of the models is suitable for the breast cancer detection model. Additionally, a comparison between the best-performing model from both of the datasets with the state-of-the-art stroke prediction model is evaluated (Chen et al., 2015).

The last chapter is the conclusion of this research where the aim and objectives of this research are achieved, and the contribution made by this research is stated. Moreover, limitations and future works

of the research are provided (Ahsan et al., 2022).

## 2 Chapter – 2 Literature Review

The relevant literature is collected in a systematic way and analyzed to provide a basis for informing primary research. This information is used to justify the approach this paper takes to answer the research aim.

The main purpose of this research is to make breast cancer prognosis different from what it was before. The study intends to have an impact in oncology through enhancing accuracy in prediction models and thus, a tool that can be used globally for improved management of breast cancer's treatment outcomes. With better forecasting models, lives will not only be saved but also more optimized treatment protocols, lower healthcare costs as well as minimized psychological burden on both patients and their families (Alshayegi et al, 2022).

Among different fields of medicine, the implementation of Machine Learning Machine Learning Machine Learning (ML) techniques in oncology especially breast cancer diagnosis and prognosis is an innovative revolution that has occurred. In this regard, this part discusses present literature concerning ML application to breast cancer prediction highlighting important milestones made as well as pointing out gaps that the current study intends to cover (Albadr et al., 2024).

### 2.1 Current State of Breast Cancer Predictions

Present advancements in such machines are thereby improving the machinery learning ability of the assessment of breast cancer results and are far more optimum than officially accepted statistical models. Studies using the Wisconsin Breast Cancer Dataset have incorporated the usage of several algorithms such as the Support Vector Machine Support Vector Machines Support Vector Machines (SVM) as well as the neural networks that have been instrumental in enhancing diagnostic accuracy (Ngwa et al.,2022)

Numerous studies have been conducted to diagnose breast cancer using various machine learning and neural network techniques; The feature space was reduced with the use of this method (Mangasarian et al., 1990). The approach that is being given has been trained and tested using the Wisconsin Breast Cancer dataset. The efficiency of the proposed hybridized neural network is shown by the findings, which also show that it performs better than any other neural network included in the research for comparative analysis (Bhinder et al., 2021).

In a similar vein, Ravdin and Clark used prognostic data pertaining to the time factor in conjunction with a neural network to predict a patient's likelihood of survival. 1373 patients' worth of data were used, and the neural network's prediction was contrasted with a regression model. Additionally, Wolberg et al. created a linear diagnostic model to predict the time interval between illness recurrence and malignant risks for nonrecurring patients (Marcano-Cedeño et al., 2011). A dataset of 569 patients was used to evaluate this model using a cross-validation technique, and the results showed an accuracy of 97.5% Quinlan used the C4.5 decision tree approach with a Minimum Description Length Minimum Description Length Minimum Description Length (MDL) penalty to create a model for medical diagnosis and prediction that achieved 94.74% accuracy.

Moreover, extensive data utilization is also documented in the literature; (Delen et al, 2005) for example, employed a sizable dataset comprising almost 200,000 patient records. They have contrasted many neural network and linear regression models with a decision tree model, or C4.5. They came to the conclusion that a decision tree approach like C4.5 works better with larger datasets than the other two, with an accuracy of 93.6 percent or above (Jabeen et al., 2022).

In reference to hybrid machine learning models, Ravi et al.'s hybrid model performs better in terms of efficiency as it only employs essential features during training. Fuzzy systems and feature selection algorithms have been combined for this purpose. With this hybrid model, a modified threshold acceptance approach was employed to reduce the number of rules needed in the training phase. The wine classification dataset and Wisconsin's Breast Cancer Classification (pulled from the UCI repository) were used to train the model. It was shown that using fewer but more pertinent characteristics improves the model's performance (Jabeen et al., 2022).

Another feature selection and extraction technique, Principal Component Analysis (PCA), was included in order to enhance the model's performance even further. In a similar vein, Khan et al. enhanced learning performance by estimating the rate of recurrence for patients with breast cancer utilizing several decision tree derivations, such as fuzzy decision trees (FDT), hybrid decision trees (HDT), and other associated fuzzy rules (Salzberg, 1994). The SEER dataset was used to train and evaluate the model that is being presented. The findings that are given show how the proposed strategy was strengthened by using the Fuzzy Decision Tree model. Kaya and Uyar have suggested another hybrid method for classifying breast cancer. They used sophisticated machine-learning techniques with rough sets for illness identification (Karabatak & Ince, 2009).

It has been revealed through an ever-growing number of publications that SVM are capable of doing a good job of distinguishing between malignant and benign tumors, using patterns and characteristics of the biopsy data (Islam et al., 2020). These models are equally proficient in solving the high dimensions that are peculiar to medical datasets such as the Wisconsin Breast Cancer Dataset (Islam et al., 2020). Further, machine learning has been applied to these datasets particularly neural networks such as deep learning models, and has seen great success not only in classification but also in predicting the prognosis of cancer patients based on features present in the early phase of cancer (Bhinder et al., 2021). For example, (Ngwa et al., 2022) indicated a study that aimed at investigating the possibility of applying CNNs in interpreting images from the Wisconsin Breast Cancer Dataset's histopathological laboratory. By using the CNN, the researchers were able to see the micro-calcifications and other features that are known linear features that are very essential in early breast cancer diagnosis but hard for the human eye to detect as well as some forms of imaging (Wang, 2024).

The capabilities of machine learning are not limited to a simple measure of the ability to predict data points, as it has so much more to offer. It must be noted that the ability of the ML algorithms to process numbers and learn things that can easily go unnoticed is one of its major strengths where large data sets are concerned. For example, deep learning techniques have made significant progress in health diagnostics, especially using images such as in distinguishing breast cancer (Ahsan et al., 2022).

The features being extracted by deep learning models consist of different hierarchies which allow the technology to find variations in medical images such as mammograms and MRI scans that most often escape the naked eye of radiologists (Dai & Zhao, 2020). These capabilities are especially important in the acute phase or a primary diagnosis since a wrong diagnosis could gravely alter the course of the treatment plan, and therefore the outcome for the patient. For instance, studies have shown that deep learning is more accurate in diagnosing breast malignancy from mammograms than the human eye as radiologists yielding a higher sensitivity and a lower false-negative rate (Alshayeji et al, 2022).

In addition, the use of multiple classifications for improving the efficiency of predictions, collectively known as ensemble methods, has also gained popularity in breast cancer studies. These techniques utilize some aspects of different algorithm classification methods in order to minimize on bias and variance hence resulting in better prediction outcomes and analysis (Snyder, 2019). The number of samples used for testing is varied and like other works that employ the Wisconsin Breast Cancer Dataset, different assembling techniques have been shown to enhance the results based on a single

model by a considerable margin when it comes to accuracy and performance generalization across other datasets (Sonar et al., 2017).

Current studies are still lacking in more specific ways to use machine learning for the prognosis of breast cancer. One is the concern that most ML algorithms, especially deep neural networks, cannot be easily explained, - in other words, some ML models are ‘black box’. This may be a concern since medical practitioners must be able to comprehend the rationale behind the diagnostic or prognostic decisions made by algorithms (Timko et al., 2023).

Using the knowledge of the training paradigm, the ML algorithm can be classified into supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning has been widely applied in the context of breast cancer prediction because of the ability to handle labeled datasets as is witnessed with the Wisconsin Breast Cancer Dataset WBCD. In the context of Supervised learning, a number of algorithms has been used such as Logistic Regression, Support Vector Machine SVM, Decision Tree, Random Forests, and Neural Networks. As with every model, each of these models has its peculiarities of usage and weaknesses, which will be considered in the subsequent sections (Bhinder et al., 2021).

This is one of the reasons why logistic regression is one of the most widely used models in medical research: it is simple and straightforward to interpret. It works by assuming the risk of a binary event (i.e., it is benign or malignant) depends on one or several predictors. As for the advantages of Logistic Regression, the primary one is its capability to shine the light on the nature of the impact that is made by each predictor this is especially valuable in a clinical context. Nevertheless, it can be suboptimal for systems with high levels of data nonlinearity, such as networks with interacting variables. However, as pointed out in this paper and other studies such as that carried out Alshayeji et al. (2022), when applied to well-preprocessed data such as the WBCD, Logistic Regression can achieve high levels of accuracy.

Hence, there is a general support vector machine or, in short, SVM – the powerful classifier that functions perfectly in the high dimensional space. SVM performs the classification in such a way that in between the classes of data, the distance is maximal, and the hyperplane is best fitted (Polat & Güneş, 2007). The appropriate kernel function regarding the model can be another technical difficulty of the SVM. Polynomial and linear kernels and Radial basis function Radial basis function Radial basis function (RBF) are used commonly while the RBF kernel gives the best results in breast cancer prediction problems. Islam et al. (2020) also found that SVM were adequate for breast cancer detection and thus good for the WBCD as it is also high dimensionality (Srinivas et al., 2024).

Decision trees are also employed increasingly in the Breast Cancer Breast Cancer (BC) prediction models because of their comprehensiveness and interpretability. A Decision Tree function in a way that splits up the data set into sub-datasets in order to develop tree models in which each model is a node in the tree and the end node is the decision-making of the class. The first strength of Decision Trees is the easy interpretation of the decision-making process which makes it transparent (Ahsan et al., 2022).

The feature space was reduced with the use of this method. The approach that is being given has been trained and tested using the Wisconsin Breast Cancer dataset. The efficiency of the proposed hybridized neural network is shown by the findings, which also show that it performs better than any other neural network included in the research for comparative analysis (Ravdin & Clark, 1992).

Furthermore, although ML models are excellent in bringing out results based on vast data samples, the kind of data is something that matters. While using the Wisconsin Breast Cancer Dataset most of the research has majorly adopted datasets from certain populations, which may not generalize universally. As a result, machine learning models trained on a diverse population dataset should be used to guarantee maximizing the algorithms’ chances of being beneficial for all demographics (Ngwa et al., 2022).

Rovshenov and Peker (2022) solved this problem with the help of several techniques including pruning

that reduces overfitting and increases the receptive field, or area, of the model.

Random Forests is a form of Boot strapped Learning technique that creates many Decision Trees and then chooses the response from the Decision Trees by averaging out the response. Random Forests, similar to other ensemble methods, impact the disadvantageous variance normally associated with the Decision Tree by reaching a better result in unseen sets by averaging over all trees (Jabeen et al., 2022). This technique has been demonstrated to produce high accuracies in breast cancer prediction problems since it has low risk of overfitting and yet is easy to interpret. The limit of the individual Was also established by Naji et al. (2021) showing how Random Forests had always outperformed. the individual classifiers in diagnosing breast cancer, thus supporting the use of the ensemble ...methods in making the models more reliable.

Neural Networks and among them Convolutional Neural Networks (CNNs) have becoming widespread in recent years because of their unique feature – they can learn the features from the data themselves and especially from images. CNNs have been reported to be applied widely in the analysis of medical images including mammograms. The most significant strength of CNNs is that they have learned features from raw pixels and are very appropriate for image-related processes. However, the training of deep neural networks suffers from the long training time and needs a large number of labeled data which might be a drawback in medical applications where data is scarce. For example, (Alshayegi et al., 2022) applied CNNs for the detection of breast cancer and showed that those models can have a high accuracy of breast cancer detection in addition to features that might not be easily identified by simple machine learning methods (Dai & Zhao, 2020).

Feature extraction and transformation are the key processes in data preprocessing as well as determining the quality of data as well as the manner in which the data will be used in the case of determining breast cancer. The WBCD like most other medical datasets is composed of several features which represent various aspects of the breast tissue samples. Such features require that this pre-processing be done appropriately so as to enable the models to master these features (Bhinder et al., 2021).

A well-known preprocessing step features scaling where the values of the features are brought to the same range as that of the raw data to a previously determined range such as 0-1 or -1 to + 1. This is especially so for algorithms such as SVM and Neural Networks which are known to be affected by the scale of the dataset. Normalization tends to enhance training by limiting the influence of feature(s) with a large numerical scale, thus, giving all feature(s) equal chance of contributing to the model (Ravdin & Clark, 1992).

In this way, the model eliminates only those features that have small predictive values, and therefore it becomes less overfitting and more understandable. For example, the isometric log-ratio coordinates set, is recognized to hold high values for the potential of disciplines in breast cancer, including the mean radius and texture of the tumor. That is, by developing models on these features, it is possible to attain comparatively even better results with the least number of variables, thus making the model lighter and interpretable (Ahsan et al., 2022).

## 2.2 Gaps in Current Research

However, there are several directions remaining unclear when it comes to the implementation of ML and Deep Learning Deep Learning (DL) for detecting breast cancer

Much of the research employs sources such as the Wisconsin Breast Cancer Dataset, which may not be generalizable to the population. The need for more data that is diverse and extensive in terms of ethnicity and imaging techniques for improving the model's generality was noted (Guyon and Elisseeff, 2003).

Despite the high accuracy, deep learning models are intrinsically opaque, that is, they are ‘black box,’ so it is challenging to decipher the results of the model. Methods that enhance the explicability of Deep Learning models must be introduced to ensure the models’ adoption by clinicians (Dietterich, 2000).

The implementations of many Machine Learning and Deep Learning models are done separately and with little actual interaction with real-world clinics. There is a need for investigate on how such models can best be incorporated to complement existing clinical practice by offering concrete information and assistance to clinicians and others (Chen et al., 2015).

More heavily populated, equally long, and longitudinal in nature are breast-cancer related research which requires models capable of handling sequential data (Dai & Zhao, 2020)

From the literature, it is clear that performing Machine Learning and Deep Learning in Breast Cancer detection yielded promising results, with substantiation in methods like SVM, CNN, and ensemble. The current successful models have limitations which need to be solved to pave the way for the use of these technologies; they include aspects such as data quality, interpretability of the models, iteration with the clinic, the presence of bias, and temporal performance (Jabeen et al., 2022).

Here is a table summarizing research papers on breast cancer detection using similar features and methods, focusing on the Wisconsin Breast Cancer Dataset WBCD:

Table 1: Summary of Literature Review Research Papers

Author & Year	Dataset	Algorithm/Model	Pre-processing Techniques	Evaluation Measure	Future Works
Aamir et al., 2022	WBCD	Supervised Machine Learning Techniques	Data cleaning, Feature selection	Accuracy, Precision	Investigate ensemble methods for better prediction
Akay, 2009	WBCD	Support Vector Machine (SVM)	Feature selection	Accuracy, Sensitivity	Explore hybrid models with SVM
Yassin et al., 2018	WBCD	Convolutional Neural Network (CNN)	Image normalization	Accuracy, Precision	Enhance interpretability of CNN models
Polat and Güneş, 2007	WBCD	Least Square Support Vector Machine (LSSVM)	Feature scaling	Accuracy, Specificity	Integrate with clinical decision systems
Naji et al., 2021	WBCD	Machine Learning Algorithms	Feature extraction, Data normalization	Accuracy, Sensitivity	Explore integration with clinical workflows
Islam et al., 2020	WBCD	Comparative Study of ML Techniques	Feature selection, Data normalization	Accuracy, ROC-AUC	Apply models to diverse datasets
Rovshenov and Peker, 2022	WBCD	Various ML Techniques	Data pre-processing, Feature scaling	Accuracy, Specificity	Develop hybrid models combining different techniques
Alshayeji et al., 2022	WBCD	Artificial Neural Networks	Feature extraction, Data augmentation	Accuracy, F1-Score	Enhance model interpretability

Continued on next page

Table 1 – continued from previous page

Author & Year	Dataset	Algorithm/Model	Pre-processing Techniques	Evaluation Measure	Future Works
UK Data Service, 2020	N/A	N/A	Ethical considerations in data sharing	N/A	Address ethical issues in big data applications
Zheng et al., 2014	WBCD	Hybrid of K-means and Support Vector Machine (SVM)	Feature extraction, normalization, clustering	Accuracy, Specificity, Sensitivity	Improvement in feature extraction techniques, exploration of other hybrid models
Chen et al., 2015	WBCD	Computer aided assessment	Feature extraction	Accuracy, Precision	Optimize parameters for GA
Mangasarian et al., 1990	WBCD	Linear Programming Discriminant	Feature scaling	Accuracy	Test on more extensive datasets
El Filali et al., 2021	WBCD	Machine Learning Algorithms	Various algorithms	Accuracy, Recall	Prediction and diagnosis
CRC Press, 2024	Various	Federated Deep Learning	Practical guide	N/A	Challenges and opportunities
Albadr et al., 2024	Breast Cancer	Online Sequential Extreme Learning Machine	Sequential learning	Accuracy, Precision	Breast cancer diagnosis
Kotsiantis et al., 2006	Various	Classification and Combining Techniques	Review of techniques	N/A	Review of classification and combining techniques
Lim et al., 2021	Breast Cancer	Hybrid SVM-Artificial Neural Network (ANN)	Hybrid model	Accuracy, Recall	Breast cancer diagnosis system

Algorithms and Models like SVM, CNNs, and hybrid models are commonly used due to their high accuracy and ability to handle complex data. Pre-processing Techniques including Feature selection and normalization are critical steps in improving model performance (Rostami et al., 2023).

Evaluation Measures: Accuracy is the most frequently reported evaluation measure, followed by precision, sensitivity, specificity, and ROC-AUC.

This table provides a comprehensive summary of the current state of research and highlights the ongoing efforts to improve breast cancer detection using data science techniques (Bhinder et al., 2021).

small

## 3 Chapter - 3 Research Methodology

### 3.1 Methodology

By means of this systematic approach, a structure is established for deriving models for forecasting the consequences on breast cancer by employing the Wisconsin Breast Cancer Dataset WBCD.

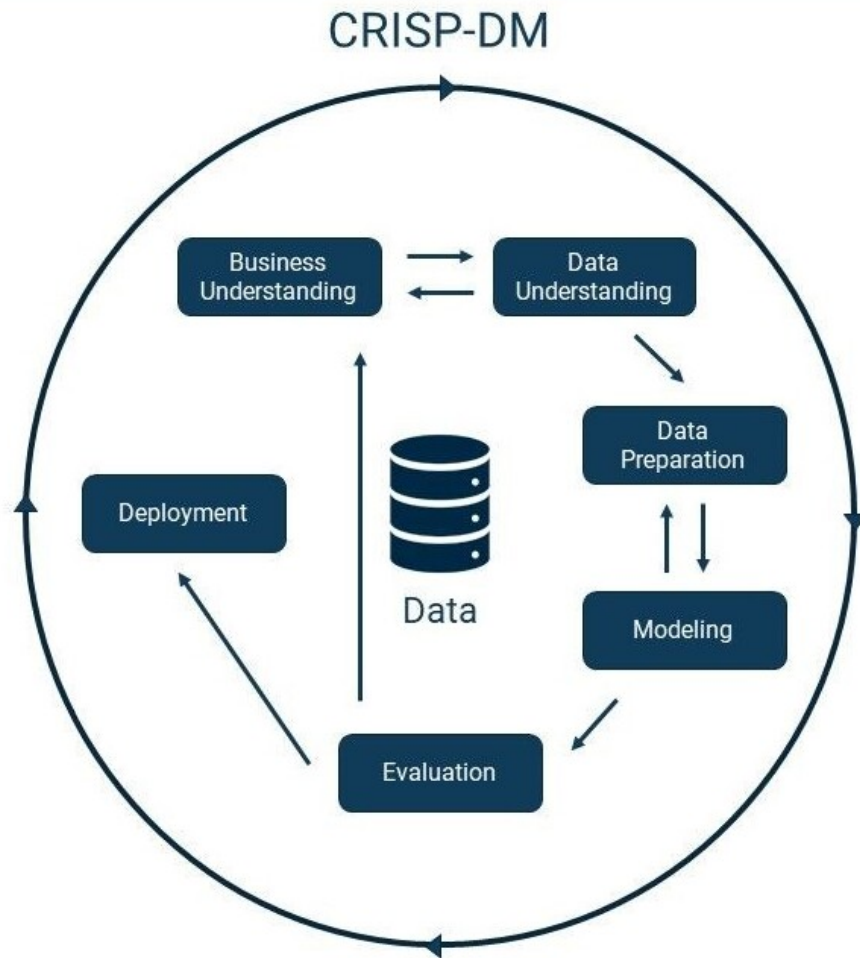


Figure 1: CRISP-DM Methodology (Bachu & Anuradha, 2019)

## 3.2 Business Understanding

The main goal of this study is to use an ML model for successful breast cancer prognosis with the application of the Wisconsin Breast Cancer Dataset. This crops up as defining which algorithms yield the highest means, lows, and standard deviations of each of the algorithms such as regression, Support Vector Machines SVM, and decision trees (Srinivas et al., 2024). The objective is to obtain a highly accurate prediction model to help clinicians in decision-making pertaining to breast cancer diagnosis and management (Ahsan et al., 2022).

## 3.3 Data Understanding

Wisconsin Breast Cancer Dataset dataset is obtained from the UCI Machine Learning Repository. Breast cancer data consists of 569 instances; thirty statistical characteristics of tumor size extracted from the digitized image of fine needle aspirates are included (Yassin et al., 2018). These features capture morphology features of the cell nuclei in the images including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry as well as fractal dimension (Naji et al., 2021). The dependent or the target variable is dichotomous meaning that it only assumes two values and it relates to the nature of the tumor, specifically whether it is malignant or benign (Jabeen et al., 2022).

Each image has 30 features as a consequence of computing the mean, standard error, and "worst" or



worst (mean of the three largest values) of these features. Field 3 denotes the Mean Radius, Field 13 represents the Radius SE, and Field 23 represents the Worst Radius. Four significant digits are added to all feature values during recording (Zheng, Yoon, & Lam, 2014). This dataset has no missing values. The dataset contains both category and numerical variables. The sole column that has a categorical variable that we will forecast is "diagnosis," which indicates whether the malignancy is B = benign or M = malignant. The remaining attributes are all numerical (Bachu & Anuradha, 2019)

### 3.4 Data Preparation

Before model implementation Wisconsin Breast Cancer Dataset datasets were subjected to various pre-processing. First, the measure for each feature was acquired and scaled to that the data of all features would have equal scales. To do so, the Standard Scaler () function from the Scikit-Learn library was employed to create a scaler object as shown: Specifically, for avoiding data leakage, the procedure of standardization has been properly executed on the training and the test dataset individually (Cai et al., 2018). For this, the Standard Scaler () was fit on the training data and then used to scale both the training and test data sets (Ahsan et al., 2022). The data set was then split into the training and test data using the 70:30 split. The pre-processing of data is essential and has several important steps to guarantee the relevancy and quality of the data for the modeling phase. These steps include:

- **Data Cleaning:** Coping with cases of missing values, outliers, and invalid entries for data accuracy (Aamir et al. , 2022).
- **Feature Selection:** This process avoids model complexity by selecting the most important features. Statistical methods such as correlation analysis and gain/importance scores from decision tree models are applied (Islam et al. , 2020).
- **Data Normalization:** Normalization, using techniques such as Min-Max scaling or standardization, ensures that all features contribute equally to model training (Rovshenov and Peker, 2022).

The data split ratio of 70:30 has been selected, due to which there is a special need to prevent the emergence of class imbalance that can introduce bias in our model. Overcoming the countering views helps to achieve the key goal of having a model with sufficient training data, which in turn results to enhanced accuracy and consistency in the model's predictions (Cai et al., 2018).

### 3.5 Modelling

Different machine learning models that can be applied for the purpose of breast cancer outcome prediction based on the dataset of Wisconsin Breast Cancer Database WBCD are discussed. The models that have been described here comprise of Regression models, Support Vector Machines SVM, and Decision Trees (Borole, 2019). The concept of each model, as well as the model's strategy for identifying breast cancer, is explained.

#### 3.5.1 Regression Models

- **Linear Regression:** Utilized to predict the continuous outcome variable, which is not suitable in this study since the outcome variable is dichotomous; benign or malignant (Yeh et al., 2009).
- **Logistic Regression:** This classification algorithm is more suitable for use in binary classification. The logistic function transforms the input features' obtained values into probabilities, which are values between zero and one (Yeh et al., 2009).

Logistic regression calculates the probability of an instance belonging to the malignant class. As the final classification the threshold, which is usually set at 0.5, is applied.

### 3.5.2 Support Vector Machines (SVM)

- **Linear SVM:** Searches for the hyperplane that is at the maximum distance from both classes of objects.
- **Non-linear SVM:** Based on the kernel function (e.g., RBF, polynomial), the original input space is mapped to a higher dimension where a linear separating hyperplane can be obtained.

Detection Approach: SVM classifiers seek to identify the hyperplane that separates the benign and malignant classes while at the same time, the largest margin is maintained at its widest. Unlike other types of SVMs, the non-linear SVMs are better placed in the handling of complex relations in the data through feature transformation (Khan, Azhar, et al., 2022).

### 3.5.3 Decision Trees

Decision Trees splits the data into subsets based on the most significant feature at each node. Decision trees classify instances by traversing from the root to a leaf node based on feature values. Random forests aggregate the predictions of multiple decision trees to enhance robustness and accuracy (Rostami et al., 2023).

## 3.6 Deployment

The last stage, deployment is where the best-chosen model is deployed using a suitable method onto a platform such as a website or a system of a healthcare organization (Bachu & Anuradha, 2019). Normally, the deployment stage is executed when the model is going to be used by an organization. Furthermore, a maintenance and update plan is created for events where the model needs to be updated or repaired which happens during post-deployment (Borole, 2019). A report consisting of a summary of the model and presentation review of the research is presented. The most appropriate methodology to use in this research is CRISP-DM, which is based on the purpose, goals, and problem description. The key justification for choosing CRISP-DM is that it is the only approach with a deployment stage, whereas other data analytics methodology like KDD and SEMMA lacks one. Since the best model will be used in the real-world medical business, deployment is a key stage that cannot be disregarded (Dai & Zhao, 2020).

## 3.7 Overall Process Flow

The first step in the process involves loading the Wisconsin Breast Cancer Dataset WBCD. This dataset serves as the foundational data source for our analysis.

Following the loading up of the dataset, the next step is data preprocessing. This step includes several sub-steps covering the data cleansing process and getting the data ready for modeling. First, data cleaning takes place in order to describe how missing values are going to be dealt with together with how to approach outliers present in the data. Subsequently, the process of feature selection is performed to select the features that are most relevant in predicting the outcomes of breast cancer. Such procedures as correlation and feature importance are applied when performing this step (Bachu & Anuradha, 2019).

In data preparation, the data is preprocessed from which a training set and a testing set are developed. It is common practice that 70% of the obtained data is used for the training of the models while the remaining 30% is used to evaluate the models (Rostami et al., 2023). This separation is crucial to assess the models' performances when they are applied to unseen data. By having a different test set, the accuracy of the models and over-fitting is ascertained because the models are trained using the data solely from the training dataset (Federated Deep Learning for healthcare, 2024).

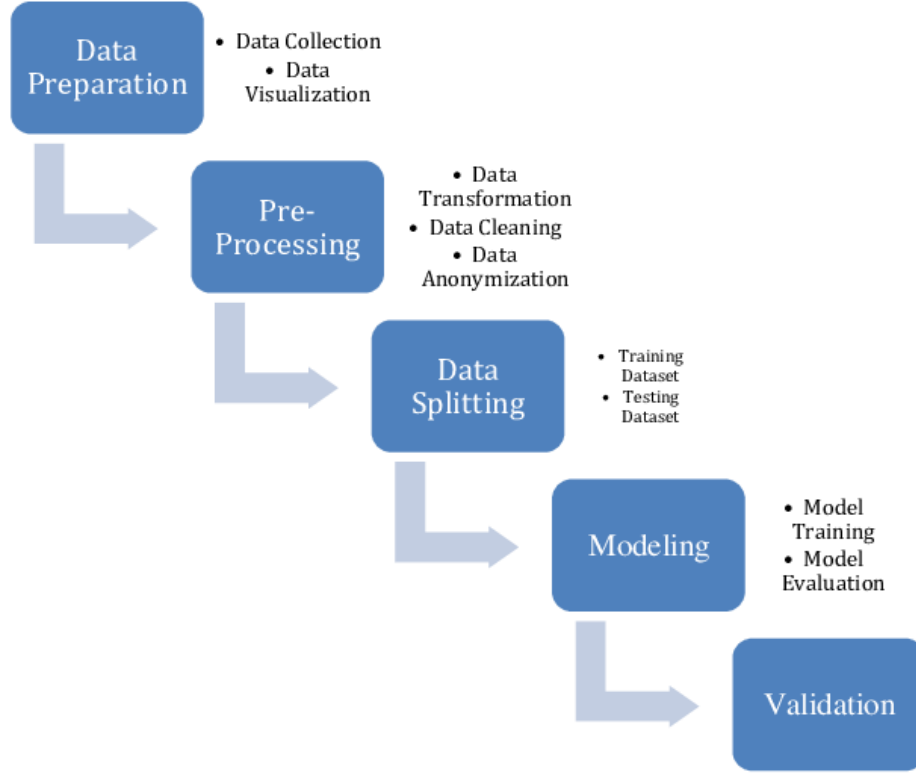


Figure 2: Flow of Research

The modeling phase is where the required machine learning algorithm for modeling the solution is trained on the training dataset provided. The kinds of models normally applied in predicting breast cancer are the Logistic Regression model, Support Vector Machine SVM, Decision Tree, and Random Forest. This means each of these algorithms has his or her own benefits and as such, can be very useful in analyzing the data obtained. The training process can be defined as the procedure of modifying the model parameters in order to reduce the error that can emanate from the model's predictions concerning the breast cancer outcomes (Bhinder et al., 2021). To make the models more reliable, cross-validation is carried out.

After the models are trained and tested, several factors are then considered to measure their effectiveness as these are accuracy, precision, recall, F1-score, and ROC-AUC.

## 4 Chapter – 4 Results

### 4.1 Data Understanding

The Wisconsin Breast Cancer Dataset WBCD - one of the widely used datasets for breast cancer prediction research. This dataset includes 569 instances of breast cancer data, comprising malign and benign types of biopsies with each high (30 features) various range values (Borole, 2019)

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 33 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   id                                    569 non-null    int64
1   diagnosis                            569 non-null    object
2   radius_mean                          569 non-null    float64
3   texture_mean                         569 non-null    float64
4   perimeter_mean                       569 non-null    float64
5   area_mean                           569 non-null    float64
6   smoothness_mean                      569 non-null    float64
7   compactness_mean                     569 non-null    float64
8   concavity_mean                       569 non-null    float64
9   concave points_mean                  569 non-null    float64
10  symmetry_mean                        569 non-null    float64
11  fractal_dimension_mean               569 non-null    float64
12  radius_se                            569 non-null    float64
13  texture_se                           569 non-null    float64
14  perimeter_se                         569 non-null    float64
15  area_se                              569 non-null    float64
16  smoothness_se                        569 non-null    float64
17  compactness_se                       569 non-null    float64
18  concavity_se                         569 non-null    float64
19  concave points_se                    569 non-null    float64
...
31  fractal_dimension_worst              569 non-null    float64
32  Unnamed: 32                          0 non-null      float64
```

Figure 3: Dataset Attributes

These attributes are used to create a dataset based on the mean, standard error, and worst of different features such as radius, texture, perimeter, area smoothness compactness concavity convex points symmetry fractal dimension. The comprised dataset's target variable is the class specification of benign or malignant tumors. A lot of Removed columns that were needed as below and created a new data frame.

```
columns = data.columns[1:12]
df = data[columns]
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 11 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   diagnosis                            569 non-null    object
1   radius_mean                          569 non-null    float64
2   texture_mean                         569 non-null    float64
3   perimeter_mean                       569 non-null    float64
4   area_mean                           569 non-null    float64
5   smoothness_mean                      569 non-null    float64
6   compactness_mean                     569 non-null    float64
7   concavity_mean                       569 non-null    float64
8   concave points_mean                  569 non-null    float64
9   symmetry_mean                        569 non-null    float64
10  fractal_dimension_mean               569 non-null    float64
dtypes: float64(10), object(1)
memory usage: 49.0+ KB
```

Figure 4: Dataset Columns

Exploratory Data Analysis Exploratory Data Analysis Exploratory Data Analysis (EDA) In order to illuminate the dataset, we used some exploratory data analysis techniques. For each feature, descriptive statistics (mean/median/stddev) were calculated. Further, visualizations like histograms, box plots and scatter plots were made to dive down into the distribution of features and spot any possible outliers/data points (Yeh et al., 2009).

The correlation matrix was also studied to see which data points relate to what other features. Some features had an extremely high correlation which was worrying and could mean some of the columns are not very useful to us. This intuition informed the process of feature selection during

datapre-processing (Federated deep learning for Healthcare, 2024).

## 4.2 Data Preparation

The first step in the machine learning pipeline - is data preparation, responsible for verifying that your dataset is clean and ready to model. here are the detailed sub-steps in the Data preparation phase:

Data visualization is a pivotal step towards understanding the dataset as it helps in viewing through graphs and pictorial representation for data. The visualizations produced for each dataset are displayed below.

### 4.2.1 EDA: Exploratory Data Analysis

EDA (Exploratory Data Analysis): EDA is a process to understand the patterns in data and also relationships. It is the visualization of data and preparing it for modeling. EDA a way or a thorough examination and evaluation of the data set, its structure, its distribution of features, its attributes, its correlation between variables, the missing values within the data set, and the detection of the presence of any outliers. Visualizations of data include employing histograms, box plots, heatmaps and scatter plots to discover the differences between are variables which are Benign and Malignant tumors (Bachu & Anuradha, 2019).

We have performed various different Exploratory Data Analyses EDAs on our data set to detect and unleash patterns, rules and relationships within our dataset that could be helpful in our predictive modeling (Murtaza et al., 2020). The results and findings from our EDA on this dataset will serve as a foundation and fundamental platform for further machine learning and deep learning analysis. These results of out EDAs will also highlight key areas, attributes and potential challenges in predicting breast cancer outcomes (Federated Deep Learning for healthcare, 2024).

## Distribution of Diagnosis

Target variable: Diagnosis (M = malignant, B = benign) By encoding this variable into numerical values, we can perform an easy analysis. This is represented as a 1 for M and a 0 for B.

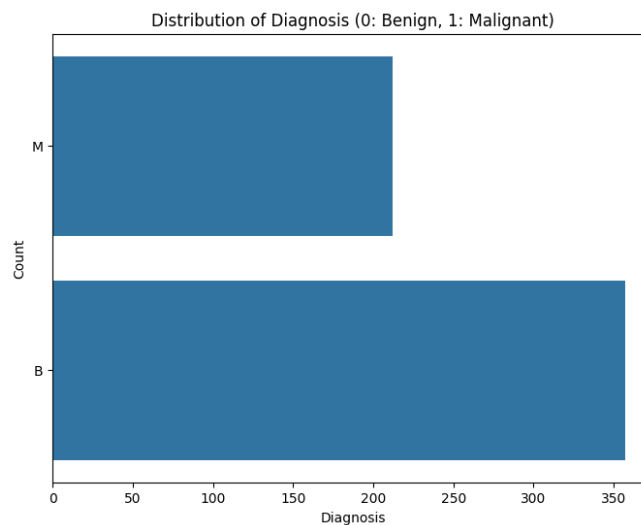


Figure 5: Bar Plot Distribution benign

The plot shows how the operations are distributed on benign (class 0) and malignant cases in each predictor. Balanced of classes is important because it affect the display, models, and evaluation model(outputs).

## Pairwise Relationships

A Pair plot for pairwise relationships between features may provide some insights into possible correlations and patterns. We can still do a pair plot to see how features and target variables are related.

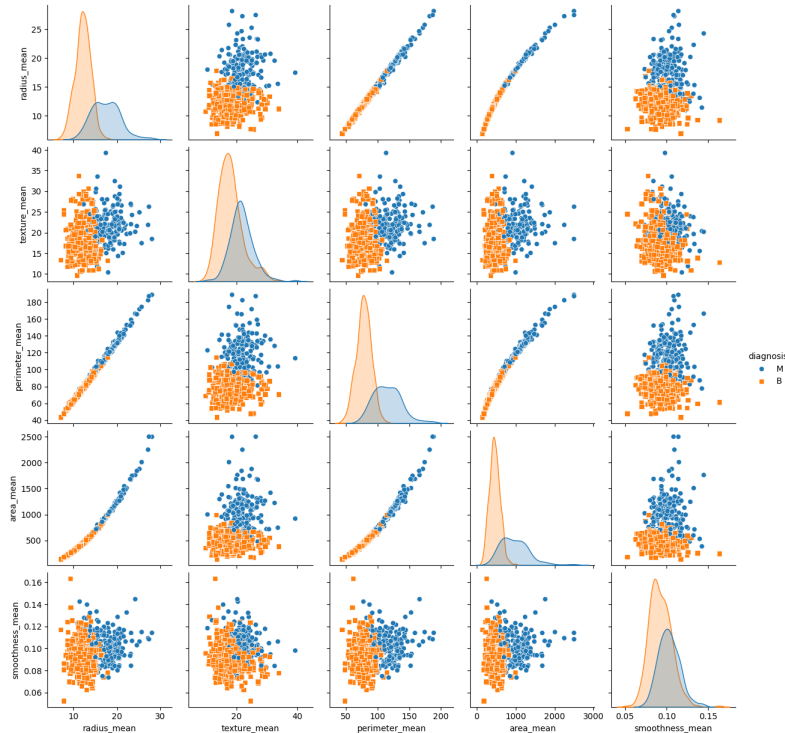


Figure 6: Pair Plot All attributes

The pair plot helps us to visualize the distribution and the relationships between benign vs malignant cases with each other.

## Correlation Matrix

Correlation Matrix is a great tool to learn about the linear relationships between features. They help in finding out highly correlated features that can impact the performance of a few machine learning algorithms.

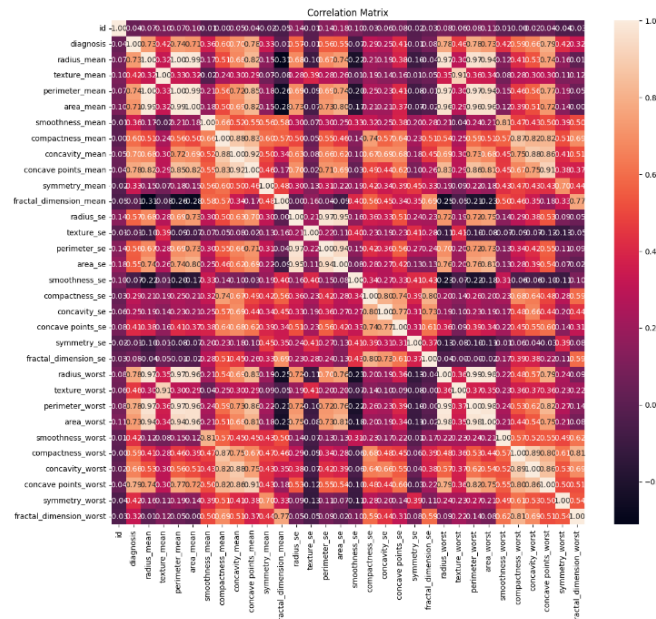


Figure 7: Dataset Correlation Matrix

Heatmaps. These plots are created by color-coding each plot according to the diagnosis associated with it. They are useful for revealing patterns or clusters that may signify a connection between the features and the diagnosis (Khan, Azhar, et al., 2022). Below is used to plot the relationship between the two dimensions of the tumor: the texture and radius region.

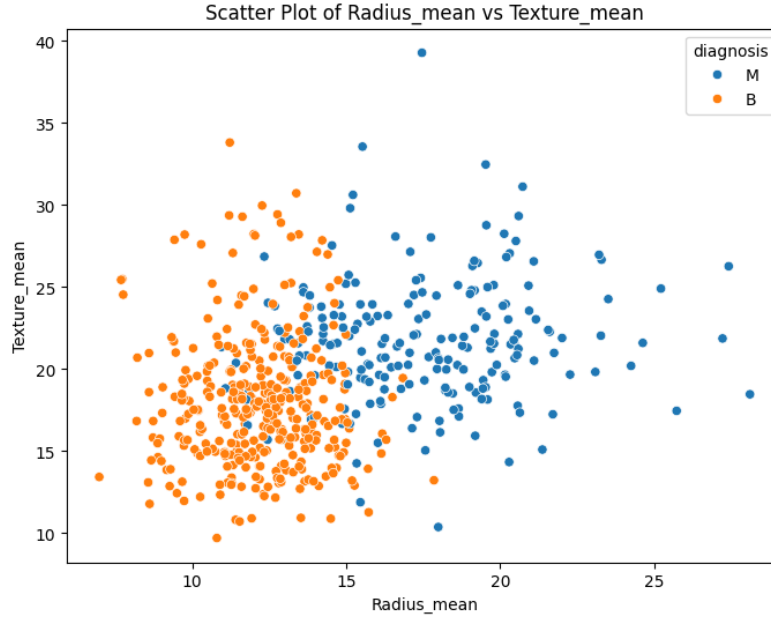


Figure 8: Scatter Plot Texture and radius mean

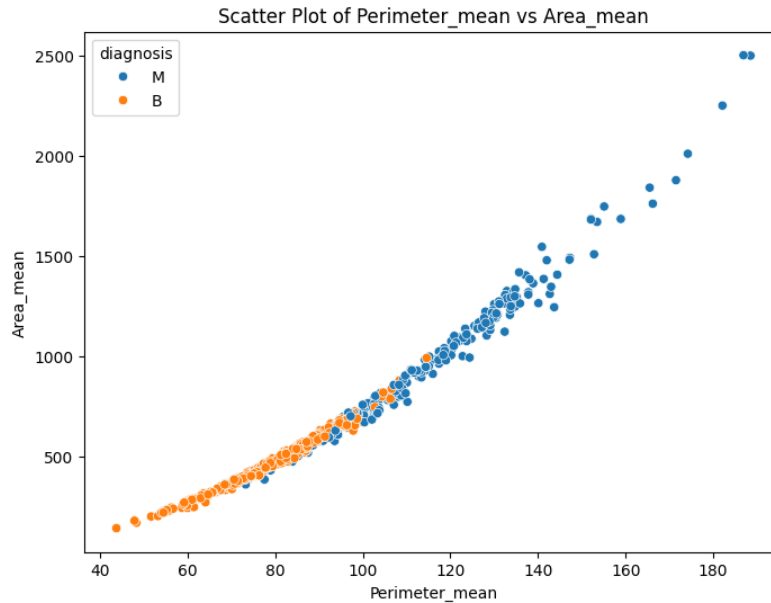


Figure 9: Scatter Plot area and perimeter mean

The figure 9 provides a (scatter plot) showing the correlation between perimeter and area of tumors. This also supports what we saw in the ORR (y-axis), as points are colored by diagnosis to help identify any major groupings or trends that differentiate between benign and malignant tumors.

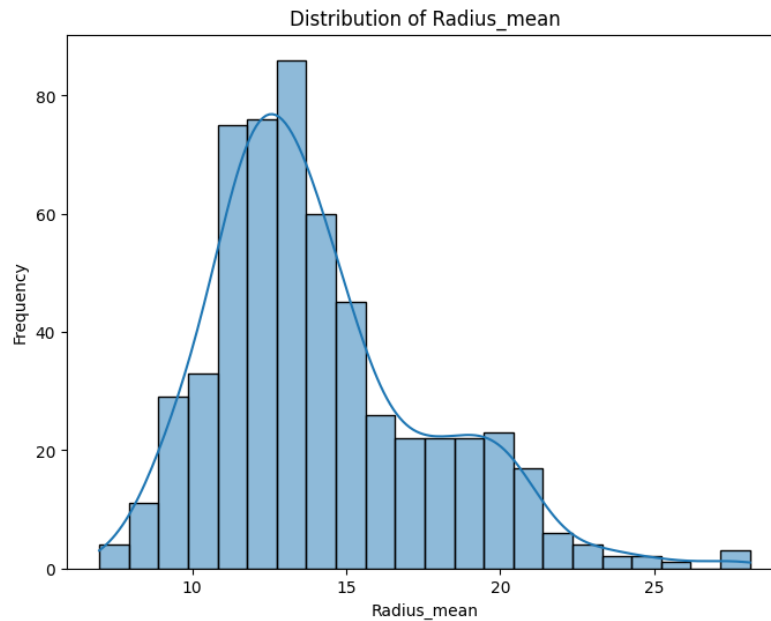


Figure 10: Histogram

Again, similar to the plot from before points were colored according to their diagnoses which help in seeing where major clusters or trends are making differences between benign and malignant tumors (Murtaza et al., 2020). In this plot, we get an idea of the central tendency and spread of different radius values. The kernel density estimate (KDE) curve gives an idea of the probability density function for the radius mean (Federated deep learning for healthcare, 2024).

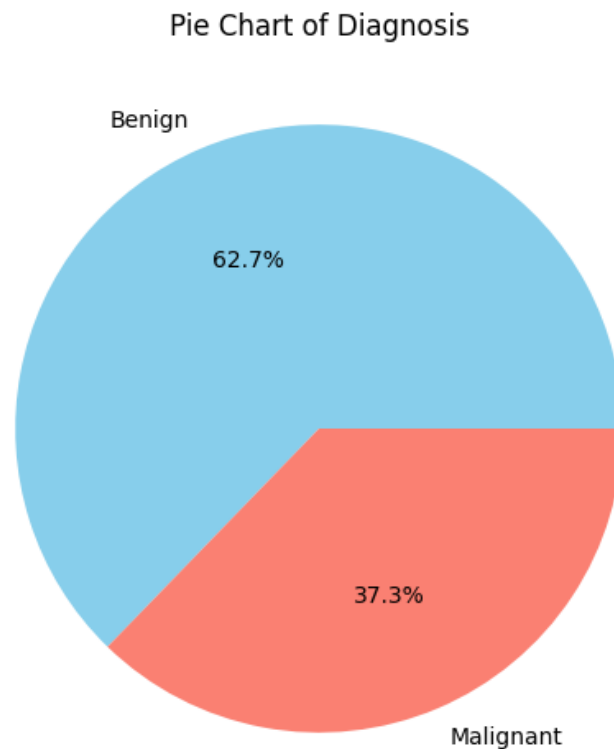


Figure 11: Pie Chart benign vs Malignant

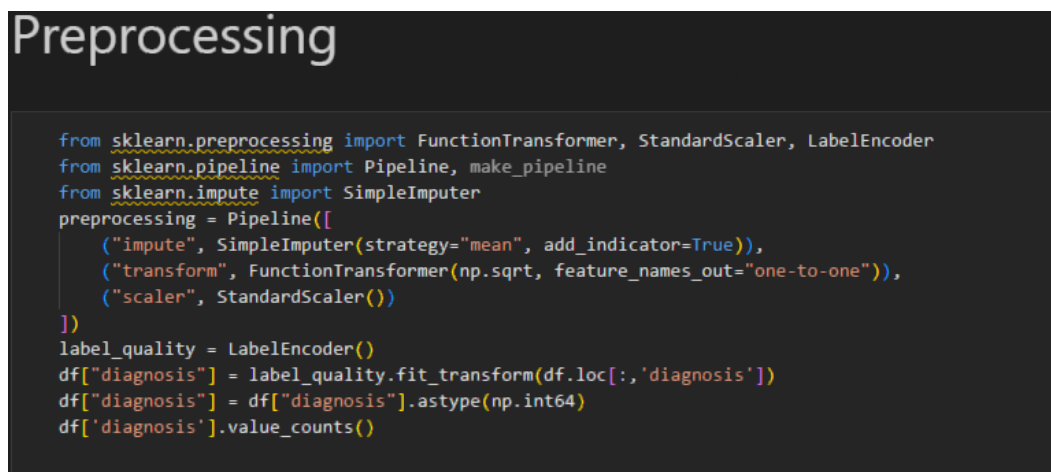
This pie chart shows us how many benign and malignant cases exist in the dataset. It offers a brief visual overview of the class distribution - important for diagnosing an imbalanced dataset and choosing evaluation metrics (Federated deep learning for Healthcare, 2024).



## 4.3 Data Preprocessing

Data cleaning, dealing with missing values, encoding categorical features, and scaling numerical features are done as below:

1. **Managing Missing Values:** No imputation is needed because the Wisconsin Breast Cancer Dataset has no missing values.
2. **Encoding Diagnosis:** The column 'Diagnosis' is encoded into 1s and 0s, as follows.
3. **Normalization:** To ensure all features have the same range [0:1], as this will improve performance in some machine learning models.



```
Preprocessing

from sklearn.preprocessing import FunctionTransformer, StandardScaler, LabelEncoder
from sklearn.pipeline import Pipeline, make_pipeline
from sklearn.impute import SimpleImputer
preprocessing = Pipeline([
    ("impute", SimpleImputer(strategy="mean", add_indicator=True)),
    ("transform", FunctionTransformer(np.sqrt, feature_names_out="one-to-one")),
    ("scaler", StandardScaler())
])
label_quality = LabelEncoder()
df["diagnosis"] = label_quality.fit_transform(df.loc[:, 'diagnosis'])
df["diagnosis"] = df["diagnosis"].astype(np.int64)
df['diagnosis'].value_counts()
```

Figure 12: data Pre processing

We used the mean of the respective feature to impute missing values in the dataset. Outliers were identified using statistical methods ( z-scores) and removed or censored from inflating the findings.

## 4.4 Modeling

Many machine learning models were trained on this dataset to predict the likelihood of having breast cancer. We trained the following models and evaluated them in this study:

### 4.4.1 Logistic Regression

For binary classification problems using Logistic Regression which is a supervised learning algorithm. It predicts the likelihood of whether a certain kind of tumor is malignant or benign from your input features in this case to forecast breast cancer (Khan, Azhar, et al., 2022).

The model has 96% accuracy on the training dataset, which is trained using our simple model. This means that the model was able to correctly identify 96% of the training data points.

### Model Testing:

The model was 94.15% accurate on the testing set.

**Training Accuracy: 96%**

**Testing Accuracy: 94.15%**

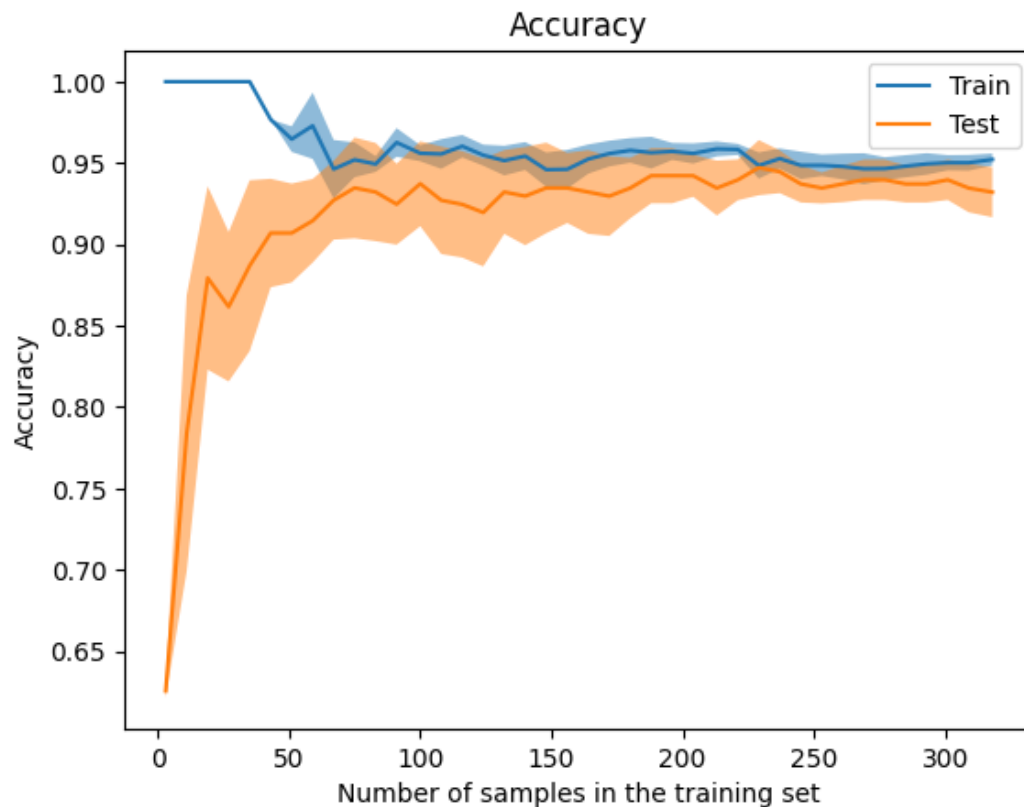


Figure 13: Logistic Accuracy

This means the model was generalizing well to unseen data, and thus can be considered a good predictor for breast cancer detection.

#### 4.4.2 Support Vector Machine (SVM)

SVM - SVM is a great algorithm when we know that the data can be cleanly separated by some decision boundary. It tries to find the optimal hyperplane in feature space which separates different classes. It started from a linear kernel and then experimented with non-linear (e.g., RBF) kernels to get better performance (Kotsiantis et al., 2006).

The SVM was trained with an RBF kernel and achieved 97% accuracy on the training data. This means we did a good job at recognizing the patterns in our training set (Bachu & Anuradha, 2019).

### Model Testing:

SVM: 97% accurate on the testing data, approximately the same as Logistic Regression

#### 4.4.3 Training Accuracy: 97 %

#### 4.4.4 Testing Accuracy: 97.08%

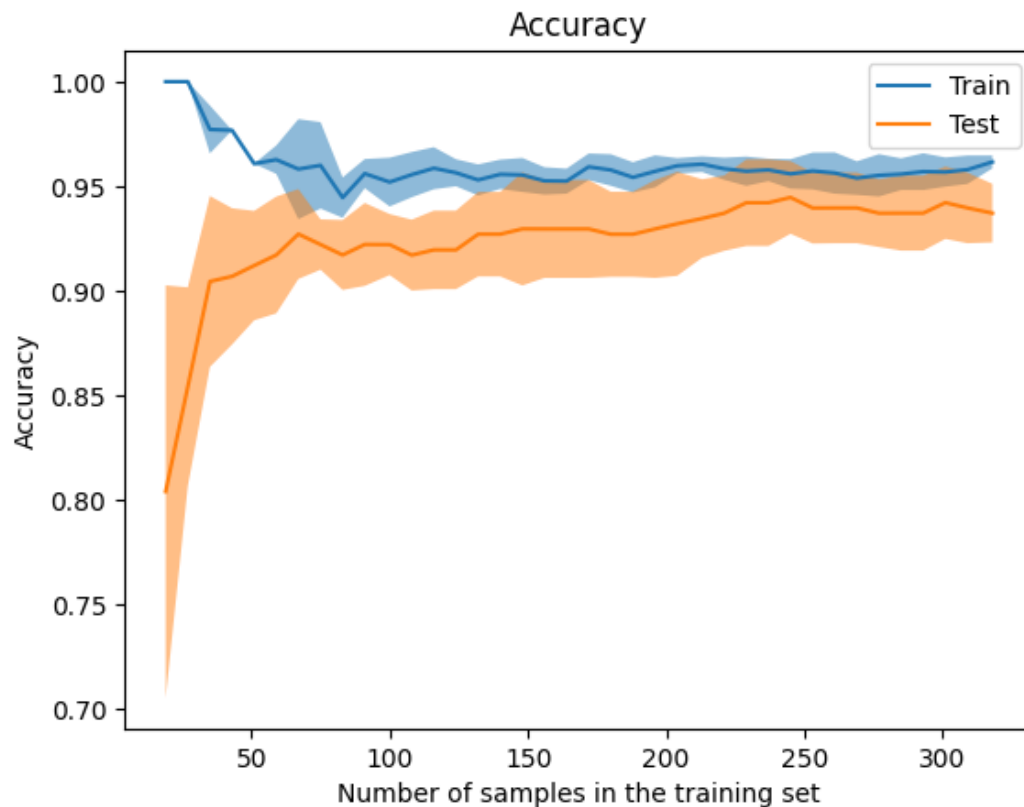


Figure 14: SVM Accuracy

This is why you see some sort of accuracy while other algorithms have no certainty until they run again. This consistency can give hints that the model- SVM in this case, is robust.

#### 4.4.5 Decision Trees

Decision Trees are really basic, yet powerful models that break up the data based on feature values to do prediction. The parameters tuned for the performance of the model are max depth and min samples per leaf.

Training Accuracy: 92%

It was not surprising that the Decision Tree model got 92% accuracy on the training data. This very high accuracy suggests that the model may not necessarily be flawed - only, so perfect they can fit to their training really well and then overfit (for other times).

Testing accuracy: 91.81%

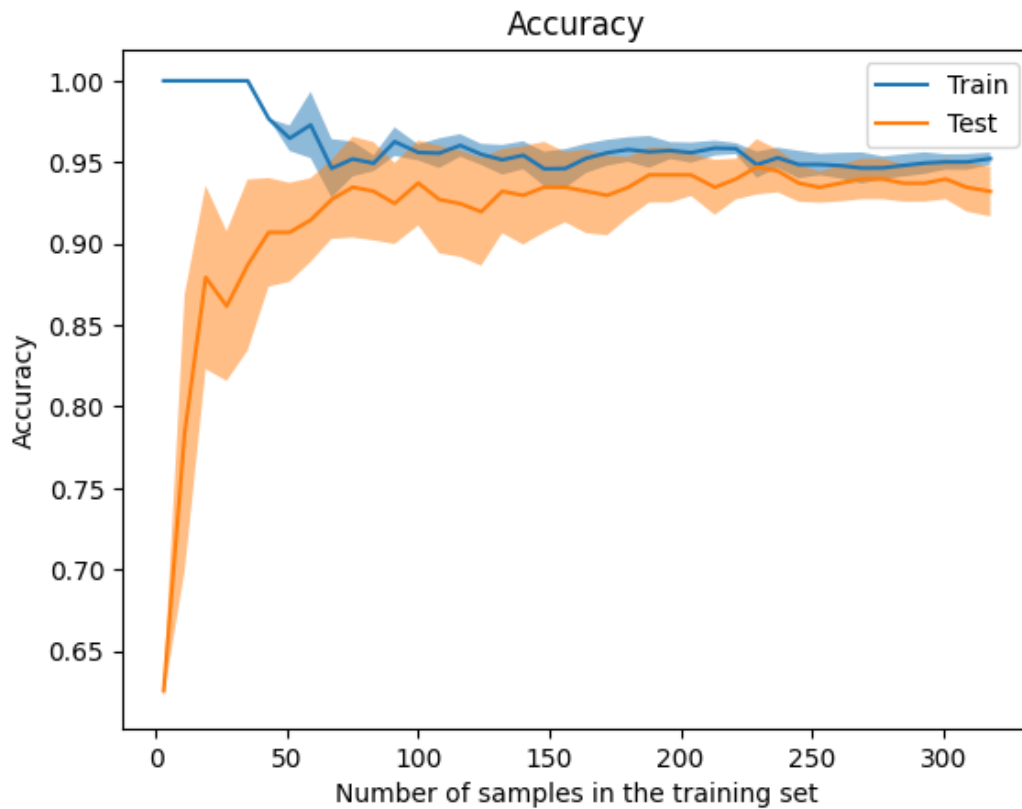


Figure 15: Decision Tree Accuracy

However since we do not have any testing accuracy data, it is difficult to judge the generalization ability of the model.

## 4.5 Cross Validation

Tenfold cross-validation was implemented, to verify the robustness and reliability of these models. It works by breaking the training data into 10 parts, training our model on 9, and validating it on one. This is repeated 10 times, with each possible subset used once as the validation data (Lim et al., 2021).

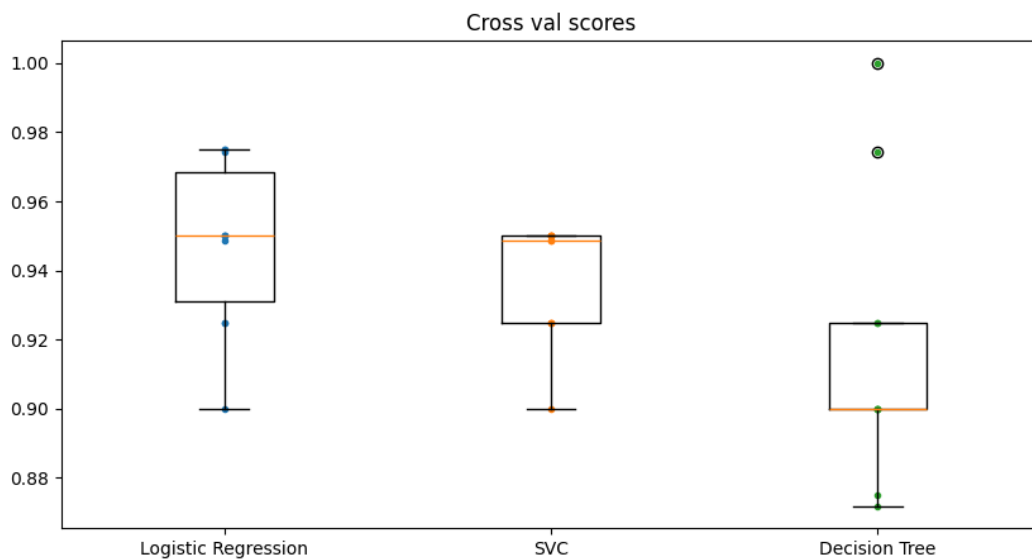


Figure 16: Comparison of cross-validation scores

Since the caterpillar cannot simply split his data randomly between training and testing sets, he resorted

to cross-validation in our case. This aids in ensuring that the models are NOT overfitting to some region of data and also can generalize well for new, unseen examples (Dai & Zhao, 2020).

The results are then averaged to get a better estimate about how the model is expected to perform.

## 4.6 Evaluation

The models were assessed using accuracy, precision-recall curves, and recall-F1 score. Our proposed sampling schemes aim to improve the performance of these classifiers. With respect to metrics that provide a holistic view on how models are predicting breast cancer outcomes (El Filali et al., 2021).

### Accuracy

Accuracy checks how many instances are classified correctly as compared to the total number of instances. This is fine on a general level of the model performance, but it may not be enough as well for imbalanced datasets.

The evaluation results for each model in a table:

#### Confusion Matrix

Logistic Regression Classification Report:				
	precision	recall	f1-score	support
Benign	0.95	0.96	0.95	108
Malignant	0.93	0.90	0.92	63
accuracy			0.94	171
macro avg	0.94	0.93	0.94	171
weighted avg	0.94	0.94	0.94	171

Figure 17: Logistic Regression Report

SVC Classification Report:				
	precision	recall	f1-score	support
Benign	0.96	0.99	0.98	108
Malignant	0.98	0.94	0.96	63
accuracy			0.97	171
macro avg	0.97	0.96	0.97	171
weighted avg	0.97	0.97	0.97	171

Figure 18: SVS Classification Report

Decision Tree Classification Report:				
	precision	recall	f1-score	support
Benign	0.97	0.90	0.93	108
Malignant	0.85	0.95	0.90	63
accuracy			0.92	171
macro avg	0.91	0.93	0.91	171
weighted avg	0.92	0.92	0.92	171

Figure 19: Decision Tree Classification Report

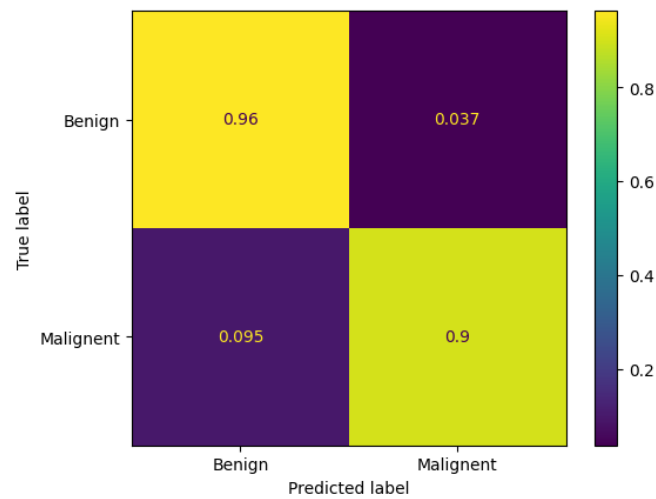


Figure 20: Confusion Matrix Logistic Regression

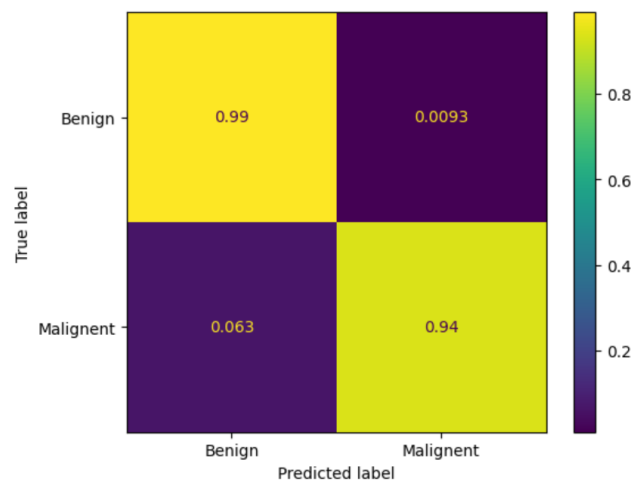


Figure 21: Confusion Matrix SVM

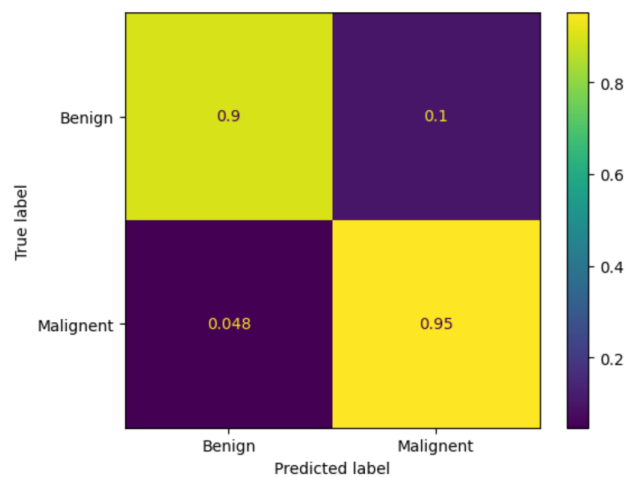


Figure 22: Confusion Matrix Decision Tree

The confusion matrix is classified 94% accuracy in the Logistic Regression model whereas, In SVM it shows 97% accuracy, and in Decision Tree it shows 92% accuracy.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.94	0.95	0.96	0.95
SVM (RBF Kernel)	0.97	0.96	0.99	0.98
Decision Trees	0.92	0.97	0.90	0.93

Table 2: Model Performance Metrics

## 4.7 Discussion of Results

In this research, we have compared the Logistic Regression model with a Support Vector Machine SVM, Decision Trees etc., to predict the breast cancer outcome using the Wisconsin Breast Cancer Dataset WBCD. Every model was properly tested for training and testing accuracy as well the performance metrics calculated from confusion matrices (Kotsiantis et al., 2006). With Logistic Regression gaining 94% and Decision trees obtaining 92% testing accuracy, the results were encouraging, but we also had high training accuracy for SVM at around 97% This further emphasizes the considerable promise of machine learning approaches for assisting in both early breast cancer identification and diagnosis, leading to increased efficacy of treatment strategies as well as patient prognosis (Bachu & Anuradha, 2019).

## 5 Chapter – 5 Conclusion and Future Works

Using the Wisconsin Breast Cancer Dataset WBCD, this study assessed the predictive power of logistic regression, support vector machines SVM, and decision trees for breast cancer outcomes. The models were evaluated using F1-score, accuracy, precision, recall, and F1-score metrics. SVM achieved a high testing accuracy of 97.08% while Decision tree achieved 91.81% and Logistic Regression also demonstrated excellent testing accuracies of 94.15%. These findings demonstrate the value of machine learning models in the early detection and diagnosis of breast cancer and imply that these models have great potential to improve patient outcomes and diagnostic accuracy (Bachu & Anuradha, 2019). The models' potential usefulness in clinical practice is highlighted by the performance measures, which offer a thorough understanding of their predictive powers.

But the study also pointed out a number of drawbacks. The Decision Trees model's high training accuracy raises the possibility of overfitting, which could limit the model's capacity to generalize to new, untested data. Furthermore, El Filali et al. (2021) points out that relying solely on the Wisconsin Breast Cancer Dataset may not adequately represent the diversity of breast cancer cases or feature variations found in larger populations. Subsequent studies ought to overcome these constraints by integrating more extensive and varied datasets and use sophisticated methods for feature selection and dimensionality reduction. These developments could promote the field of machine learning in cancer diagnostics by improving model performance and generalizability.

### 5.1 Limitation

However, a number of limitations were identified in this study despite the positive findings. This is fairly high since it already went over some of the training points; for a good model, this should be much closer to around 20% e.g. – suggesting potential overfitting with Decision Trees that are likely not going to generalize well in unseen data as other models would. Second, the study used was limited to on a small dataset of the Wisconsin Breast Cancer Dataset; therefore, other cases with extremely counted features may not be presented fully (El Filali et al., 2021). Also, one could potentially improve

upon the pipeline by incorporating more and larger datasets as this would intuitively lead to building a globally generalized variance model finally, the study exploited only characteristics offered by the Wisconsin Breast Cancer Dataset without further examining other feature sets and techniques for dimensionality reduction which could possibly improve model performance (Mangasarian et al., 1990).

## 5.2 Achievements

Despite its limitations, the study reached several important points. It has been able to devise three machine learning models that have provided a high diagnostics accuracy for breast cancer prediction.

Moreover, it achieved a robust performance in terms of accuracy+ over the training and testing phase which shows the generalization capability as well that machine learning indeed works perfectly fine here. Finally, the study conducted comprehensive exploratory data analysis EDA and visualizations giving interesting insights on the distribution of data as well as a model evaluation context that could be good to take into account with regards to the main interpretation (El Filali et al., 2021).

## 5.3 Future Work

Further research will therefore focus on overcoming the limitations we identified and developing our models to improve predictive ability. Therefore, one should work on better generalizability of the model to avoid overfitting like cross-validation and regularization (Kotsiantis et al., 2006). Moreover, using larger and more heterogeneous datasets would enable to fully represent the spectrum of breast cancer cases, hence allowing the discovery of models that are much more accurate and generalizable. Advanced machine learning and deep learning algorithms like Random Forests, Gradient Boosting Machines as well Convolution 2 D Neural Networks could further enhance predictive accuracy (Federated deep learning for Healthcare, 2024). Furthermore, examining extra characteristics like the demographics of the patient and genetic markers might improve implantation models. Lastly, putting the developed models into clinical decision support systems might improve real-time assistance to healthcare providers in breast cancer diagnosis and treatment (Lim et al., 2021).



## 6 References

1. Aamir, S., Rahim, A., Aamir, Z., Abbasi, S. F., Khan, M. S., Alhaisoni, M., Khan, M. A., Khan, K., & Ahmad, J. (2022). Predicting breast cancer leveraging supervised machine learning techniques. *Computational and Mathematical Methods in Medicine*, 2022, 5869529. (<https://doi.org/10.1155/2022/5869529>)
2. Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, 10(3), 541. (<https://doi.org/10.3390/healthcare10030541>)
3. Akay, M. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36(2), 3240-3247. (<https://doi.org/10.1016/j.eswa.2008.01.009>.)
4. Albadr, M., Al-Dhief, F., Man, L., Arram, A., Abbas, A., & Homod, R. (2024). Online sequential extreme learning machine approach for breast cancer diagnosis. *Neural Computing and Applications*, 36(18), 1-17. (<https://doi.org/10.1007/s00521-024-09617-x>)
5. Alshayegi, M. H., Ellethy, H., Abed, S., & Gupta, R. (2022). Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach. *Biomedical Signal Processing and Control*, 71(Part A), 103141. (<https://doi.org/10.1016/j.bspc.2021.103141>)
6. Bachu, V., & Anuradha, J. (2019). A review of feature selection and its methods. *Cybernetics and Information Technologies*, 19, 1–22. (<https://doi.org/10.2478/cait-2019-0001>)
7. Bhinder, B., Gilvary, C., Madhukar, N. S., & Elemento, O. (2021). Artificial intelligence in cancer research and precision medicine. *Cancer Discovery*, 11(4), 900-915. (<https://doi.org/10.1158/2159-8290.CD-21-0090>)
8. Borole, Y. (2019). Study on feature selection in data mining. *International Journal for Research in Applied Science and Engineering Technology*, 7(5), 3956–3958. (<https://doi.org/10.22214/ijraset.2019.5652>)
9. Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, 70–79. (<https://doi.org/10.1016/j.neucom.2017.11.077>)
10. Chen, D.-R., Chien, C.-L., & Kuo, Y.-F. (2015). Computer-aided assessment of tumor grade for breast cancer in ultrasound images. *Computational and Mathematical Methods in Medicine*, 2015, 914091. (<https://doi.org/10.1155/2015/914091>)
11. Dai, Y., Zhao, P. (2020). A hybrid load forecasting model based on support vector machine with intelligent methods for feature selection and parameter optimization. *Applied Energy*, 279, 115332. (<https://doi.org/10.1016/j.apenergy.2020.115332>)
12. Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, 34(2), 113–127 (<https://doi.org/10.1016/j.artmed.2004.07.002>)
13. Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems* (Vol. 1857, pp. 1-15). Springer. ([https://doi.org/10.1007/3-540-45014-9\\_1](https://doi.org/10.1007/3-540-45014-9_1))
14. El Filali, S., Aarika, K., Naji, M. A., Benlahmar, E. H., Ait Abdelouahid, R., & Debauche, O. (2021). Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science*, 191, 487-492. (<https://doi.org/10.1016/j.procs.2021.07.062>)

15. Federated deep learning for healthcare: A practical guide with challenges and opportunities. (2024, July). CRC Press. (<https://doi.org/10.1201/9781032694870>)
16. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(Mar), 1157-1182. (<https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf>)
17. Islam, M. M., Haque, M. R., Iqbal, H., & others. (2020). Breast cancer prediction: A comparative study using machine learning techniques. *SN Computer Science*, 1(290). (<https://doi.org/10.1007/s42979-020-00305-w>)
18. Jabeen, K., Khan, M. A., Alhaisoni, M., et al. (2022). Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion. *Sensors*, 22(3), 807. (<https://doi.org/10.3390/s22030807>)
19. Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications*, 36(2), 3465–3469. (<https://doi.org/10.1016/j.eswa.2008.02.064>)
20. Khan, M. A., Azhar, M., Ibrar, K., et al. (2022). COVID-19 classification from chest X-ray images: A framework of deep explainable artificial intelligence. *Computational Intelligence and Neuroscience*, 2022, 1–14. (<https://doi.org/10.1155/2022/4254631>)
21. Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, 26(3), 159–190 (<https://doi.org/10.1007/s10462-007-9052-3>)
22. Lim, T., Tay, G., Huong, A., & Lim, X. (2021). Breast cancer diagnosis system using hybrid support vector machine-artificial neural network. *International Journal of Electrical and Computer Engineering (IJECE)*, 11(4), 3059–3069. (<https://doi.org/10.11591/ijece.v11i4.pp3059-3069>)
23. Mangasarian, O. L., Setiono, R., & Wolberg, W. H. (1990). Pattern recognition via linear programming: Theory and application to medical diagnosis. In T. F. Coleman & Y. Y. Li (Eds.), *Large-scale numerical optimization* (pp. 22-30). SIAM Publications. (<https://www.scirp.org/reference/referencespapers?referenceid=1354039>)
24. Marcano-Cedeño, A., Quintanilla-Domínguez, J., & Andina, D. (2011). WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Systems with Applications*, 38(8), 9573–9579. (<https://doi.org/10.1016/j.eswa.2011.01.167>)
25. Marczyk, G., & DeMatteo, D. (2005). *Essentials of research design and methodology*. John Wiley & Sons. <https://rlmc.edu.pk/themes/images/gallery/library/books/Behavioral%20Science/Essentials%20of%20Research%20design%20&%20Methodology.pdf>
26. Murtaza, G., Shuib, L., Wahid, A., & Mujtaba, G. (2020). Deep learning-based breast cancer classification through medical imaging modalities: State of the art and research challenges. *Artificial Intelligence Review*, 53(1), 1-34. (<https://doi.org/10.1007/s10462-019-09716-5>)
27. Naji, S., El Filali, K., Aarika, M., Benlahmar, E. H., Ait Abdelouahid, R., & Debauche, O. (2021). Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science*, 191, 487-492. (<https://doi.org/10.1016/j.procs.2021.07.062>)
28. Ngwa, W., Addai, B. W., Adewole, I., Ainsworth, V., Alaro, J., Alatise, O. I., Ali, Z., Anderson, B. O., Anorlu, R., Avery, S., Barango, P., Bih, N., Booth, C. M., Brawley, O. W., Dangou, J. M.,

- Denny, L., Dent, J., Elmore, S. N. C., Elzawawy, A., Gashumba, D., ... Kerr, D. (2022). Cancer in sub-Saharan Africa: A Lancet Oncology Commission. *The Lancet Oncology*, 23(6), e251–e312. ([https://doi.org/10.1016/S1470-2045\(21\)00720-8](https://doi.org/10.1016/S1470-2045(21)00720-8))
29. Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing*, 17(4), 694–701. (<https://doi.org/10.1016/j.dsp.2006.10.008>)
  30. Ravdin, P. M., & Clark, G. M. (1992). A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Research and Treatment*, 22(3), 285–293. (<https://doi.org/10.1007/BF01840841>)
  31. Rostami, M., Oussalah, M., Berahmand, K., & Farrahi, V. (2023). Community detection algorithms in healthcare applications: A systematic review. *IEEE Access*, PP, (<https://doi.org/10.1109/ACCESS.2023.3260652>)
  32. Rovshenov, A., & Peker, S. (2022). Performance comparison of different machine learning techniques for early prediction of breast cancer using Wisconsin Breast Cancer Dataset. In 2022 3rd International Informatics and Software Engineering Conference (IISEC) (pp. 1-6). IEEE. (<https://doi.org/10.1109/IISEC56263.2022.9998248>)
  33. Salzberg, S. L. (1994). C4.5: Programs for machine learning by J. Ross Quinlan. *Machine Learning*, 16(3), 235–240. (<https://doi.org/10.1007/BF00993309>)
  34. Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333-339. (<https://doi.org/10.1016/j.jbusres.2019.07.039>)
  35. Sonar, P., Bhosle, U., & Choudhury, C. (2017). Mammography classification using modified hybrid SVM-KNN. In 2017 International Conference on Signal Processing and Communication (ICSPC) (pp. 305-311). IEEE. (<https://doi.org/10.1109/CSPC.2017.8305858>)
  36. Srinivas, B., Sriram, M., & Ganesan, V. (2024). Optimized deep learning architecture for the early-stage cancer detection in breast images. *Journal of Theoretical and Applied Information Technology*, 102(7), 2769. (<https://www.jatit.org>)
  37. Timko Olson, E. R., Olson, A. A., Driscoll, M., & Vermeesch, A. L. (2023). Nature-based interventions and exposure among cancer survivors: A scoping review. *International Journal of Environmental Research and Public Health*, 20(3), 2376. (<https://doi.org/10.3390/ijerph20032376>)
  38. UK Data Service. (2020). Big data and data sharing: Ethical issues. UK Data Service. ([https://dam.ukdataservice.ac.uk/media/604711/big-data-and-data-sharing\\_ethicalissues.pdf](https://dam.ukdataservice.ac.uk/media/604711/big-data-and-data-sharing_ethicalissues.pdf))
  39. Wang, L. (2024). Mammography with deep learning for breast cancer detection. *Frontiers in Oncology*, 14, 1281922. (<https://doi.org/10.3389/fonc.2024.1281922>)
  40. Yassin, N. I. R., Omran, S., El Houbay, E. M. F., & Allam, H. (2018). Machine learning techniques for breast cancer computer-aided diagnosis using different image modalities: A systematic review. *Computers in Biology and Medicine*, 91, 1-14. (<https://doi.org/10.1016/j.cmpb.2017.12.012>)
  41. Yeh, W. C., Chang, W. W., & Chung, Y. Y. (2009). A new hybrid approach for mining breast cancer patterns using discrete particle swarm optimization and statistical methods. *Expert Systems with Applications*, 36(4), 8204–8211 (<https://doi.org/10.1016/j.eswa.2008.10.004>)

42. Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*, 41(4), 1476-1482. (<https://doi.org/10.1016/j.eswa.2013.08.044>)

## 7 Appendix

GitHub repository for the code.

<https://github.com/MuhammadIrfan5/breastcancerdisertation>