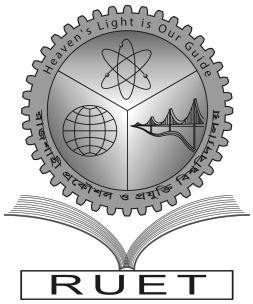
Heaven's Light is Our Guide



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

Breast Cancer Detection Using Different Machine Learning Models

Author

Mahmud Murshed

Roll No. 1703137

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

Supervised by

Dr. Md. Al Mamun

Professor

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

ACKNOWLEDGEMENT

As I consider reaching this important milestone, I am humbled and incredibly appreciative of the unwavering support and inspiration I have received from so many people who have helped me along the way.

I have found comfort and vigor in my faith during times of reflection and effort. I am infinitely grateful to Allah Ta'ala for the opportunity and benefits He has given me.

First and foremost, I want to express my heartfelt gratitude to my respected supervisor **Prof. Dr. Md. Al Mamun**, who oversaw my thesis. His advice, mentoring, and commitment have played a critical role in determining the direction of my study. I owe a debt of gratitude to him for his knowledge and insight, which sparked my enthusiasm for machine learning research and inspired me to explore into the challenging field of machine learning-based breast cancer categorization.

I am incredibly grateful to my family for their unwavering support, words of encouragement, and faith in my abilities. My foundation has been their constant support, which has enabled me to face challenges and celebrate victories. I sincerely thank my friends, whose support and companionship have made even the most trying days more bearable. Your unshakable confidence has motivated me.

I owe my priceless education and direction to the committed faculty members who have shared knowledge, disproved presumptions, and promoted a supportive learning atmosphere.

This path has been a tapestry made of the threads of assistance, guidance, and inspiration from numerous facets of my life. My thesis advisor, my family, my friends, the teachers, and my faith have all made significant contributions to finish this project.

I'll end by saying that the kindness and encouragement I have experienced have served as a reminder of the significant influence that people can have on one another's journeys. I sincerely thank everyone who has contributed to my academic and personal development. This success is evidence of the value of group work and the strength of community.

August 13,2023 Mahmud Murshed

RUET, Rajshahi

Heaven's Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

CERTIFICATE

This is to certify that this thesis report entitled ***"Breast Cancer Detection Using Different Machine Learning Models"*** *submitted by* ***Mahmud Murshed, Roll:1703137*** *in partial fulfillment of the requirement for the award of the degree of Bachelor of Science in Department of Computer Science & Engineering of Rajshahi University of Engineering & Technology, Bangladesh is a record of the candidate own work carried out by him under my supervision.*

Supervisor

_____

Dr. Md. Al Mamun
Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

External Examiner

_____

Md. Farukuzzaman Faruk
Assistant Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

ABSTRACT

The classification of breast cancer has become a crucial field of research, with machine learning approaches for precise and timely diagnosis. Naive Bayes, Logistic Regression, Decision Tree, XGBoost, Random Forest, k-Nearest Neighbors (KNN), and Neural Network are just a few of the machine learning algorithms for breast cancer classification that have been thoroughly examined in this thesis. We carefully designed features based on a correlation value threshold using the Wisconsin Breast Cancer dataset from the UCI Machine Learning Repository, making it easier to choose pertinent variables. Our research provided fascinating insights into how well various algorithms performed, with the XGBoost model emerging as the best at differentiating between benign and malignant breast cancers. This result highlights the effectiveness of ensemble learning in challenges involving breast cancer classification. The Random Forest classifier also demonstrated notable accuracy, taking second-best spot among the examined models. The combination of diligent outcome analysis, thorough algorithmic evaluation, and careful feature selection results in a thorough understanding of the advantages and disadvantages of various machine learning models for breast cancer classification. This study advances the body of information that strives to improve breast cancer early detection and diagnosis for advances in better healthcare outcomes.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Introduction

The most frequent kind of cancer found in women across the world, breast cancer is a serious global health issue [1]. Men are also impacted, but far less frequently. This illness develops when abnormal cells in the breast tissue start to grow and multiply out of control, which causes a tumor to form. These malignant cells may eventually infiltrate the surrounding tissues and, in more advanced stages, travel throughout the body via the circulation or the lymphatic system.

Incidence and Prevalence: Breast cancer incidence and prevalence have a significant effect on general health. Breast cancer is the main cause of cancer-related deaths in women and makes up around 25% of all cancer cases, according to the World Health Organization (WHO)[2]. Breast cancer occurs at different rates across the world, with greater rates in industrialized nations[3]. Although it can affect women of all ages, age is a key risk factor, with the majority of instances happening in women over the age of 50[4].

Risk Factors: Despite the fact that the precise etiology of breast cancer is still unknown, a number of risk factors have been found. Age, gender (women are more susceptible), family history of breast cancer or certain genetic abnormalities (such as BRCA1 and BRCA2), personal history of breast cancer or specific benign breast disorders, early menstrual cycle beginning or late menopause, radiation exposure, and obesity are some of these factors[1]. For prevention, early identification, and customized therapies, it is essential to comprehend these risk variables.

Signs and Symptoms: Breast cancer may not usually show signs in its early stages, underlining the significance of routine breast health examinations and screenings. The presence of a thickening/lump in the breast or underarm, changes in the shape and size, or appearance of

the breast, skin changes on the breast (like dimpling, puckering, redness etc.), nipple changes (inversion, discharge, or scaling), and persistent pain in the breast or nipple are all common signs of tumor growth[5].

Diagnosis and Screening: Early diagnosis of breast cancer is essential for better outcomes and survival rates. Mammography, professional breast examinations, and breast self-examination are all common screening techniques[6]. If an anomaly is found, additional diagnostic procedures, such as ultrasonography, MRI, or biopsy, are carried out to determine whether cancer is present and to characterize it.

Treatment and Management: The management of breast cancer is based on the kind, stage, and condition of the hormone receptors as well as the general health of the patient. Surgery (mastectomy/lumpectomy), chemotherapy, radiation therapy, hormone therapy, targeted therapy, and immunotherapy are examples of common treatment techniques[7]. Treatment results and survivability have significantly improved as a result of developments in medical research.

Breast cancer has major effects on people, families, and society at large. It is a complicated and multidimensional disease[8]. Research efforts are ongoing to comprehend the underlying mechanisms, optimize patient care overall, create specific medicines, and boost early detection techniques. To lessen the impact of breast cancer and enhance patient outcomes, it is imperative to raise awareness, encourage breast health education, and support research initiatives.

1.2 Motivation

Many people die everyday all over the world due to breast cancer. Specially in the low developed countries, people generally get to know about their diseases at adverse states. They are not aware of health. They go to doctors when they can not tolerate the symptoms and pain. At that time diseases already become more dangerous and attacking to the body and also go out of control. Then, it becomes very difficult to rescue the patients. Every year, around 685,000 women die from breast cancer in the world[9], in every 6 cancer deaths in women, 1 of them is for breast cancer[10].

We know that in medical science, cancer detection is one of the most complex diagnostic processes[11]. Complexity and heavy costs are associated with it. In the low developed countries, most of the people can not afford that. In the early stage of the cancer, rural people don't

go to the doctors because they can tolerate the early pain and sorrows. When they go the hospitals, it is too late to defend the disease. For any disease, early diagnosis is crucial for rescue the patients. So late detection is the main cause of such dangerous death rate due to breast cancer[12]. Early, low cost, available process of diagnosis may deduce the high death rate.

Some reasons for early diagnosis are listed here.

More Effective Treatment Possibilities: Early breast cancer detection opens us a larger choice of treatment alternatives, including less drastic but more efficient procedures. Early-stage malignancies are frequently more manageable and confined, making surgery and other focused treatments more effective.

Increased Survival Rate: The likelihood of survival is considerably increased when breast cancer is discovered early. When opposed to tumors that have migrated to other regions of the body and reached later stages, early-stage malignancies often have a better prognosis and longer overall survival.

Decreased Aggression: As they advance to later stages, certain breast cancers may become more aggressive. Early cancer detection and treatment may stop the disease from worsening and perhaps spreading to other organs.

Mental and Physiological Advantages: Early detection of breast cancer may lessen some of the emotional suffering brought on by a diagnosis at a later stage. Early intervention gives patients a feeling of control and empowerment, enabling them to decide on their own course of therapy with knowledge.

Potential Applications of Targeted Treatments: Targeted treatments that concentrate on certain traits of cancer cells have been developed as a result of advancements in medical research. Early detection gives doctors more time to analyze the genetic makeup of the tumor to choose the most appropriate course of action.

Enhancement in Quality of Life: Improved patient quality of life may be a result of early diagnosis. Patients may retain a greater degree of physical and mental health by avoiding the need for expensive therapies and minimizing the disease's effects on everyday life.

Decreased Medical Costs: In the long term, lower healthcare expenses may be achieved by early identification and treatment of breast cancer. More intensive and expensive therapies, such as surgeries, chemotherapy, radiation therapy, and hospital stays, are often used to treat advanced-stage cancer.

Essentially, Less Surgical Techniques: Instead of a complete mastectomy, which involves

the removal of the whole breast, early-stage breast cancers may be dealt with using less invasive techniques including a lumpectomy (removal of the tumor plus a small amount of surrounding tissue). This more fully maintains the breast's form and functionality.

1.3 Problem Statement

Breast Cancer: A Concern for World Health: A powerful shadow is cast by breast cancer, which is becoming a widespread and profoundly important health issue that cuts beyond national boundaries and social structures. Its effects on women's health have a significant and global reach. The idea that early and effective detection of breast cancer is not just a medical need but also a lifeline that may profoundly affect patients' lives is what drives the urgency of the early diagnosis of breas cancer.

To detect cancerous cells in human body or find cancer diseases, people need to go through complex, heavy and costly diagnostic processes[13]. Most of the people, specially in the underdeveloped countries can not afford that.

One of the most prevalent cancers impacting women worldwide is breast cancer[14]. Improved patient outcomes depend on accurate diagnosis and early detection.. The goal of this project is to create and evaluate machine learning models for the automated identification of breast cancer utilizing a variety of data sources, such as clinical data of patients.

1.4 Objectives

The goal of this project is to create, put into practice, and rigorously assess a variety of machine learning models for the aim of identifying breast cancer. The main goal is to assess the performance of these models using a variety of data sources thorough patient clinical records, with a laser emphasis on accuracy and early detection. Finding the most efficient algorithms will progress automated breast cancer screening techniques and increase the likelihood of early intervention and better patient outcomes.

Here are the specific objectives of this research works.

Create and assess several machine learning models for detecting breast cancer: The objective of this study is to construct and assess several machine learning algorithms, including support vector machines (SVMs), random forests, and neural networks, for the purpose of breast

cancer identification. The objective is to identify a machine learning model that can get a high level of accuracy and dependability in the classification of benign and malignant cases.

Choose a model based on machine learning that has breast cancer detection's precision and dependability: The primary goal of this study is to construct a machine learning algorithm with the purpose of enhancing the precision and dependability of breast cancer identification. The proposed model has the potential to serve as an independent tool for breast cancer diagnosis, as well as enhance the precision of conventional techniques like mammography.

This analysis aims to compare the efficacy of machine learning models with conventional approaches, namely mammography, in the context of breast cancer diagnosis: The goal of this study is to conduct a comparative analysis of machine learning models and conventional approaches, such as mammography, for the purpose of breast cancer diagnosis. The objective of this study is to ascertain if machine learning models can surpass conventional approaches in terms of accuracy and dependability.

1.5 Organization of Thesis

Chapter 2 - Background Study

Chapter 3 - Methodologies and Implementation

Chapter 4 - Results & Performance Analysis

Chapter 5 - Conclusion & Future Works

1.6 Conclusion

In general, machine learning has considerable potential as a tool for enhancing the timely identification of breast cancer. Machine learning algorithms have shown the potential to attain a notable level of precision and dependability in the classification of mammograms into categories of either benign or malignant. In addition, they have the potential to enhance the precision of conventional approaches for detecting breast cancer, such as mammography. This has the potential to result in timelier identification and intervention of breast cancer, hence potentially yielding life-saving outcomes.

Chapter 2

Background Study

2.1 Introduction

I have done analysis of different machine learning models for breast cancer classification on different data-sets. Different tools, methods, processes and techniques, knowledge about the corresponding field are gathered from different sources or articles. How these tools can be used in research work, what are the scopes and what are the limitations, scopes of research, mechanism of the models etc. are studies in a broad sense. Specially, previous research works related to this field, their advantages and disadvantages scope of contribution are studied clinically. This information is used to make the project more efficient and effective solution to the problem.

2.2 Machine Learning

Machine learning is a specialized domain within the realm of artificial intelligence (AI) that centers around the creation and refinement of algorithms and models, enabling computers to acquire knowledge and generate forecasts or determinations by leveraging data. It enables computational systems to enhance their performance iteratively through experiential learning, obviating the need for explicit programming for each specific task. Machine learning has emerged as a paradigm-shifting phenomenon in diverse domains and use cases owing to its prowess in processing intricate datasets, automating operations, and yielding novel insights that were hitherto arduous or unattainable through conventional programming approaches[15].

Machine learning fundamentally centers on the notion of training models using data. These

computational models acquire knowledge of patterns, correlations, and relationships inherent in the given dataset, thereby empowering them to generate precise forecasts or categorizations for novel, unobserved data instances. There exist various machine learning methodologies, encompassing:

Supervised Learning: In the realm of supervised learning, models undergo training using labeled data, wherein input data is associated with corresponding target outcomes. The objective is for the model to acquire the ability to comprehend the correlation between input values and output values, thereby enabling it to make forecasts for novel data instances[16].

Semi-Supervised Learning: This approach synergistically integrates components of supervised and unsupervised learning paradigms, wherein a diminutive quantity of annotated data is amalgamated with a substantial quantity of unannotated data to enhance the efficacy of the model[17].

Unsupervised Learning: Unsupervised learning pertains to the handling of unannotated data, wherein the model endeavors to uncover intrinsic patterns, structures, or associations within the data. Clustering and dimensionality reduction are frequently encountered tasks within the domain of unsupervised learning[18].

Reinforcement Learning: In the domain of reinforcement learning, an autonomous agent engages in a dynamic exchange with its surrounding environment, acquiring knowledge on optimal action selection in order to maximize a given reward signal. It is frequently employed in situations where an autonomous entity must engage in a series of sequential decision-making processes[19].

Machine learning is a field that encompasses a wide range of algorithms and techniques, such as neural networks, decision trees, support vector machines, linear regression, and other similar methodologies. The selection of an algorithm is contingent upon the inherent characteristics of the problem at hand, the available data structure, and the intended objective.

2.3 Different Machine Learning Algorithms

There are numerous algorithms in machine learning field. These are Naive Bayes Classifier, Support Vector Machine, Decision tree, Random Forest, Logistic Regression etc. They are explained briefly below:

2.3.1 Naive Bayes Classifier

Naive Bayes is a probabilistic algorithm utilized in machine learning for classification purposes. It is founded upon Bayes' theorem, which enables the computation of the probability of a hypothesis given the available evidence. This is a straightforward yet robust algorithm frequently employed for text classification, spam detection, and analogous tasks[20]. The "naive" assumption in Naive Bayes posits that all features exhibit independence, thereby facilitating computational simplification. However, it is important to acknowledge that this assumption may not be universally valid.

Bayes' Theorem: Bayes' Theorem is a basic concept in probability theory and machine learning, which is expressed as[20]:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

Where,

- $P(H|E)$ is the posterior probability of hypothesis H given evidence E.
- $P(E|H)$ is the likelihood of evidence E given hypothesis H.
- $P(H)$ is the prior probability of hypothesis H.
- $P(E)$ is the probability of evidence E.

Naive Bayes' Classifier: H stands for a class name in the classification context, and E stands for characteristics or attributes connected to the data point. The Naive Bayes classifier determines the likelihood that a data item belongs to each class and allocates it to the class where the likelihood is highest.

The Naive Bayes classifier is a probabilistic machine learning algorithm commonly used for text classification and spam detection. It's based on Bayes' theorem and assumes that all features are independent of each other, which is known as the "naive" assumption.
Consider a binary classification scenario with two classes: C_1 and C_2. The Naive Bayes classifier calculates the probability of a data point X belonging to class C_1 given its features $x_1, x_2, \ldots, x_n$ as[20]:

$$P(C_1|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|C_1) \cdot P(C_1)}{P(x_1, x_2, \ldots, x_n)}$$

The ”naive” assumption implies that features are conditionally independent given the class label, which simplifies the likelihood term:

$$P(x_1, x_2, \ldots, x_n|C_1) = P(x_1|C_1) \cdot P(x_2|C_1) \cdot \ldots \cdot P(x_n|C_1)$$

In practice, the Naive Bayes classifier calculates the probabilities for each class and assigns the data point to the class with the highest probability.

2.3.2 Logistic Regression

Logistic regression is a prevalent statistical technique utilized for binary classification, wherein it estimates the likelihood of a given data point being assigned to a specific class. In spite of its nomenclature, the tool is employed for tasks pertaining to classification rather than regression. Logistic regression is a valuable tool in cases where the dependent variable exhibits categorical characteristics, such as binary outcomes represented by yes/no, true/false, or 0/1[21].

Equations: The logistic transformation, commonly referred to as the sigmoid function, is employed in Logistic Regression for the purpose of converting the linear combination of features and coefficients into a probability value[21]:

$$P(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

Where,

- $P(Y = 1|x)$ is the probability that the outcome Y is 1 given the input features x.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients learned during model training.
- e is the base of the natural logarithm.

This probability value represents the likelihood that the data point belongs to class 1 (positive class) based on the input features.

2.3.3 Decision Tree Classifier

A Decision Tree is an algorithm in the field of machine learning that is capable of performing both classification and regression tasks. It is known for its versatility and interpretability. The decision-making process involves the recursive partitioning of the input feature space into distinct regions that align with various class labels. In the context of this system, it is important

to note that every partition is symbolized by a tree node. Furthermore, the decision rules are defined by the paths that originate from the root node and extend to the leaf nodes[22].

Equations: The computational process of Decision Tree classification necessitates the utilization of mathematical equations to determine impurity measures for the purpose of evaluating splits. Two frequently employed metrics are Gini impurity and entropy, also known as information gain. Here is an abstract representation of the equations:

I. Entropy and Information Gain: Entropy measures the uncertainty or disorder in a set of samples. For a node N with K classes and p_k being the proportion of samples in class k at node N, the entropy H is calculated as[22]:

$$Entropy(N) = -\sum_{k=1}^{K} p_k \log_2(p_k)$$

Information Gain measures the reduction in entropy achieved by partitioning the data based on a particular attribute. If N_{parent} is the number of samples in the parent node, and N_{child} is the number of samples in a child node, the Information Gain IG for a split is given by[22]:

$$IG = Entropy(\text{parent}) - \sum_{\text{child nodes}} \frac{N_{\text{child}}}{N_{\text{parent}}} \cdot Entropy(\text{child})$$

Where:

N_{child} is the number of samples in a child node.

N_{parent} is the number of samples in a child node.

The sum is taken over all child nodes resulting from the split.

II. Gini Impurity: Gini impurity measures the probability of a randomly selected element being misclassified. For a node N with K classes and p_k being the proportion of samples in class k at node N, the Gini impurity $Gini(N)$ is calculated as[22]:

$$Gini(N) = 1 - \sum_{k=1}^{K} (p_k)^2$$

The aforementioned equations are employed in the process of constructing decision trees to assess the split (consisting of a feature and threshold) that will yield the greatest reduction in impurity or entropy.

2.3.4 Random Forest Classifier

Random Forests is an ensemble learning algorithm which leverages the power of combining multiple Decision Trees in order to enhance the performance and generalization capabilities of a singular Decision Tree classifier. It exhibits notable efficacy in classification tasks, manifesting superior accuracy and robustness through the mitigation of overfitting[23].

The Underlying Operational Concept: Random Forests constructs an ensemble of Decision Trees, wherein each tree is trained on a distinct subset of the training data and possibly with a subset of features. The ultimate forecast is generated by consolidating the forecasts of individual trees, typically employing majority voting for classification.

- **Bagging or Bootstrapping:** The training data is partitioned into multiple random subsets using the technique of bootstrapping. Every distinct subset, referred to as a "bootstrap sample," is employed for the purpose of training a distinct Decision Tree[23].
- **Voting Mechanism:** Every tree in a classification job predicts the class label. Out of all the forecasts made by the trees, the ultimate prediction is chosen by majority vote[23].
- **Feature Subset Selection:** Only a portion of characteristics are taken into account for splitting at each node of a decision tree. By doing so, the connection between trees is decreased and a characteristic does not take over the decision-making process[23].

2.3.5 Support Vector Machines (SVM)

Both classification and regression applications use Support Vector Machines (SVM), a powerful machine learning technique. SVMs seek for a hyperplane that maximizes the margin between data points of distinct classes while optimally separating them. It works especially well when the data cannot be separated linearly by moving it into a higher-dimensional space[24].

Working Theory: Support Vector Machines (SVM) operate by identifying the most favorable hyperplane that maximizes the separation between data points belonging to distinct classes. The data points in closest proximity to the hyperplane are referred to as support vectors, and they are responsible for establishing both the location and alignment of the hyperplane. Support Vector Machines (SVMs) have the capability to employ a kernel function, which allows for the implicit mapping of data into a space of higher dimensionality. This enables the SVM to achieve non-linear separation[24].

Equations: Let us engage in a discourse regarding the fundamental equations entailed in Support Vector Machines (SVM).

Equation of The Hyperplane: For a linearly separable dataset, the hyperplane can be represented as[24]:

$$f(x) = \mathbf{w} \cdot \mathbf{x} + b$$

Where:

- $f(x)$ is the decision function that assigns a data point x to a class.
- $\mathbf{x}$ is the input feature vector.
- $\mathbf{w}$ is the weight vector.
- b is the bias term.

Margin Calculation: The distance between the hyperplane and the support vectors is called the margin. The margin is calculated as[24]:

$$\text{Margin} = \frac{2}{\|\mathbf{w}\|}$$

Where $\|\mathbf{w}\|$ is the Euclidean norm of the weight vector.

Optimization Criteria: The Support Vector Machine (SVM) algorithm is designed to optimize the margin between data points while simultaneously ensuring accurate classification. This can be expressed as an optimization problem[24]:

$$\text{Maximize} \left(\frac{2}{\|\mathbf{w}\|} \right) \text{Subject to:} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$$

Where:

- y_i is the class label of data point $\mathbf{x}_i$.
- $\mathbf{w}$ is the weight vector.
- b is the bias term.

Non-linear SVM: For the data that are not linearly separable, SVM can use kernel function $K(x, x')$ to map the data into a higher dimensional space. The set of frequently used kernel functions encompasses linear, radial basis function (RBF), polynomial and sigmoid.

2.3.6 Neural Network

A neural network is a highly potent machine learning model that draws inspiration from the intricate arrangement and operation of interconnected neurons in the human brain. Neural networks are extensively employed for diverse tasks, encompassing image recognition, natural language processing, and additional applications.

Architecture: A neural network is a computational model composed of multiple layers of interconnected computational units known as neurons. The system is structured in a hierarchical manner, comprising of three primary categories of layers[25].

1. **Input Layer:** The layer responsible for receiving the initial data or input features.
2. **Hidden Layer:** There exist one or more intermediary layers positioned between the input and output layers. Every individual neuron within a concealed layer obtains input values from the preceding layer and generates an output value to be transmitted to the subsequent layer.
3. **Output Layer:** The ultimate stratum responsible for generating the network's forecast or result.

Working Principle: Neural networks function by means of a mechanism referred to as forward propagation, wherein inputs are transmitted through the layers in order to generate an output. In the realm of neural networks, it is customary for every inter-neuronal linkage to possess a corresponding weight. Additionally, each individual neuron is equipped with an activation function that governs the computation of its output[25].

Equations: Let us engage in a discourse regarding the fundamental equations entailed within a rudimentary feedforward neural network:

- **Calculation of Weighted Sum:** For each neuron in a layer (except the input layer), the weighted sum of inputs is calculated as[25]:

$$z_j = \sum_{i=1}^{n} w_{ji} \cdot x_i + b_j$$

  Where:

  - z_j is the weighted sum for neuron j in the current layer.

- w_{ji} is the weight of the connection between neuron i in the previous layer and neuron j in the current layer.
- x_i is the output of neuron i in the previous layer.
- b_j is the bias term for neuron j.

- **Activation Function:**

  The weighted sum is passed through an activation function to introduce non-linearity[25]:

$$a_j = \text{activation}(z_j)$$

  Common activation functions include sigmoid, ReLU and tanh.

Training: Neural networks acquire knowledge through the utilization of backpropagation, a technique that involves iteratively modifying the weights and biases in order to minimize the discrepancy between the predicted and observed outputs. The procedure encompasses the computation of gradients and the subsequent modification of weights using optimization algorithms such as gradient descent[25].

2.3.7 K-Nearest Neighbor

The k-Nearest Neighbors (KNN) method is a simple and intuitive classification methodology that is used in both classification and regression applications. The algorithm finds the k training samples in the feature space that are most similar to a given test sample before making a forecast based on the majority class (in classification) or average value (in regression) of those k neighbors[26].

Steps:

- **Choose a value for k:** Please specify the value of k, which represents the number of neighbors to be taken into account for generating predictions for every test sample.
- **Determine distances:** For each test sample, calculate the distance (e.g., Euclidean distance) between the features of the test sample and the features of all training samples.
- **Get the k closest neighbors:** Perform a computation to determine the k training samples that exhibit the lowest distances in relation to the test sample.

- **Classification-based majority voting or regression-based average:** In the context of classification tasks, it is necessary to compute the quantity of neighbors belonging to each class. Subsequently, the test sample should be allocated to the class exhibiting the greatest count. In the context of regression tasks, it is advisable to compute the mean target value of the 'k' nearest neighbors and employ it as the prediction.

Equations in mathematics:

1. **Distance calculation:**

   For two points p and q in an n-dimensional feature space, the Euclidean distance between them can be calculated as[26]:

$$\text{Distance}(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

   Where:

   p_i and q_i are the values of the i-th feature for points p and q, respectively.

2. **Regression-based average:**

   For regression tasks in the k-Nearest Neighbors (KNN) algorithm, the predicted value $\hat{y}$ for a test sample is calculated as the average of the target values of the k nearest neighbors[26]:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i$$

   Where:

   y_i is the target value of the i-th nearest neighbor.

3. **Classification-based majority voting:**

   For classification tasks in the k-Nearest Neighbors (KNN) algorithm, the class label $\hat{y}$ predicted for a test sample is determined by the majority class among the k nearest neighbors[26]:

$$\hat{y} = \arg\max_{y_i} \sum_{i=1}^{k} [y_i = y]$$

Where:

- y_i is the class label of the i-th nearest neighbor.
- y represents each possible class label.

2.3.8 XGBoost (Extreme Gradient Boosting)

A potent machine learning method that belongs to the class of boosting algorithms is called XGBoost (Extreme Gradient Boosting). It is frequently used for both regression and classification tasks, and it has proven to perform better in a variety of tests and practical settings. A powerful predictive model is produced using the ensemble approach known as XGBoost by combining the strengths of several weak learners (usually decision trees)[27].

Steps Of The Algorithm: [27]

1. **Create a new model:** Commence by establishing an initial forecast for every instance in the training dataset. This could be a rudimentary forecast, such as the mean value for regression or the logarithm of the odds for classification class probabilities.

2. **Create a tree:** Design and implement a decision tree algorithm to effectively capture and represent the errors or residuals of the initial predictions. The construction of this tree is designed to minimize a particular loss function, commonly the Mean Squared Error (MSE) for regression tasks or the Log Loss for classification tasks.

3. **Update predictions:** Aggregate the forecasts generated by the original model with the forecasts generated by the recently incorporated tree. The predictions undergo updates according to a learning rate (shrinkage) parameter, which regulates the impact of each tree's contribution.

4. **Repeat steps 2 to 3:** Construct additional trees to rectify the inaccuracies present in the preceding trees. Every individual tree within the model strives to enhance its predictive capabilities by prioritizing instances that exhibit significant errors.

5. **Final prediction:** Aggregate the forecasts generated by each individual tree to derive the ultimate forecast of the XGBoost model.

Equations:

1. **Loss Function:** In the context of regression tasks, it is customary to employ the Mean Squared Error (MSE) as the designated loss function for the purpose of minimization[27].

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where:

- y_i is the true target value for the i-th instance.
- $\hat{y}_i$ is the predicted target value for the i-th instance.

In the realm of classification tasks, it is common practice to employ the Log Loss (also known as cross-entropy) as the designated loss function.

$$\text{Log Loss} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$$

Where:

- y_i is the true class label (0 or 1) for the i-th instance.
- $\hat{y}_i$ is the predicted probability of the positive class for the i-th instance.

2. **Update Prediction:** The prediction for a sample is updated by using the learning rate η and the output of a recently appended tree's forecast. $f_t(x)$ [27]:

$$\text{Updated Prediction} = \text{Previous Prediction} + \eta \cdot f_t(x)$$

Where:

- Previous Prediction is the prediction from the previous step.
- η is the learning rate, controlling the contribution of the new tree's prediction.
- $f_t(x)$ is the output of the t-th tree for the instance x.

The efficacy of XGBoost arises from its ability to proficiently model complex relationships through iterative procedures involving multiple trees. Moreover, XGBoost utilizes regularization methodologies such as maximum depth regulation and column subsampling in order to alleviate the potential problem of overfitting.

It is of utmost significance to acknowledge that XGBoost incorporates a multitude of auxiliary parameters and methodologies, encompassing gradient boosting, regularization, and early stopping, thereby augmenting its performance and adaptability. The previously mentioned description provides a thorough overview, along with the essential mathematical symbols, for understanding the mechanics of XGBoost.

2.4 Biological Concept Of Neural Network

The fundamental cognitive apparatus within the human body that facilitates the acquisition and assimilation of knowledge is the brain. The neural network consists of approximately 10^{10} interconnected neurons. A neuron acquires information from other neurons via its synapses. Once the cumulative total of the inputs surpasses a specific threshold, the neuron initiates an electrical spike transmission to other neurons via the axon[28].

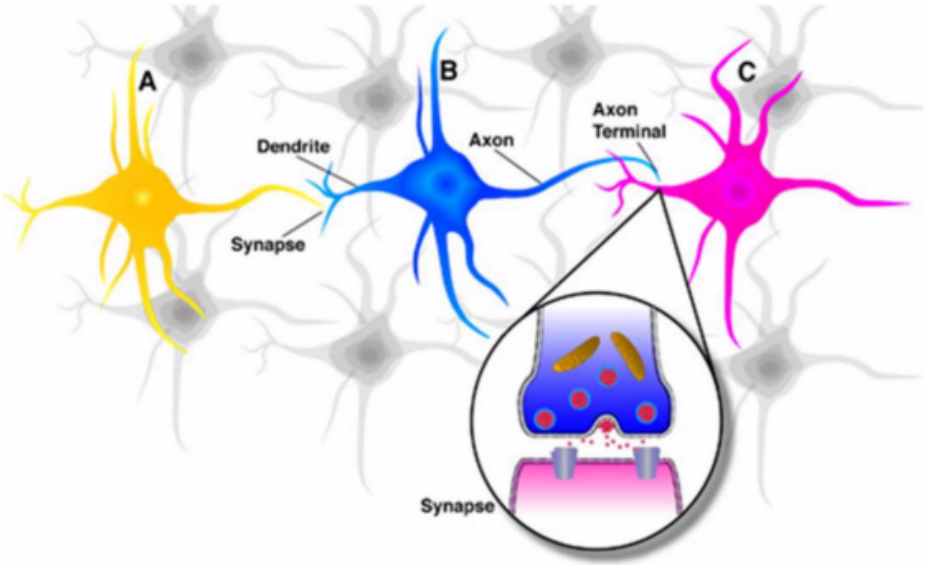


Figure 2.1: Biological Neural Network[29]

2.5 Perceptron Learning

The perceptron algorithm is a commonly employed machine learning technique utilized for the purpose of binary classification.It is a rudimentary implementation of an Artificial Neural Network (ANN).In the realm of binary classifications, the learning process involves the utilization of linear separation within the feature space, akin to the methodology employed in logistic

regression[30].

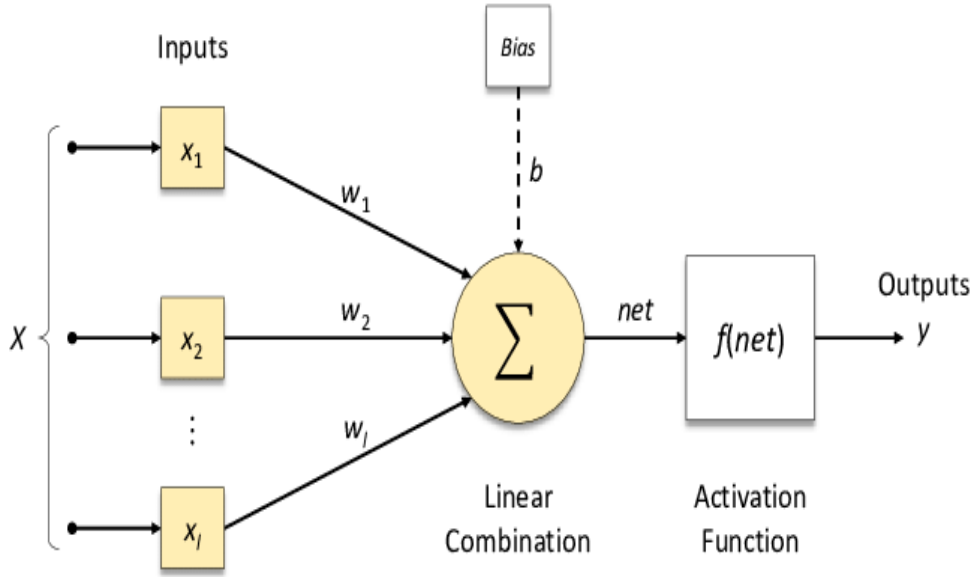


Figure 2.2: Perceptron[31]

In the other words:

$$f(x) = \begin{cases} 1, & W^T X + w_0 \geq 0 \\ 0, & W^T X + w_0 \leq 0 \end{cases}$$

Here,

$$W = \begin{bmatrix} w_1 \\ w_2 \\ . \\ . \\ w_m \end{bmatrix} \qquad X = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_m \end{bmatrix} \tag{2.1}$$

Artificial neural networks can be described as a type of multi-layer perceptron.This is commonly referred to as a feedforward neural network.This comprises the constituent elements:

- Input layer
- Hidden Layers
- Bias unit

- Weights
- Activation function
- Output layer

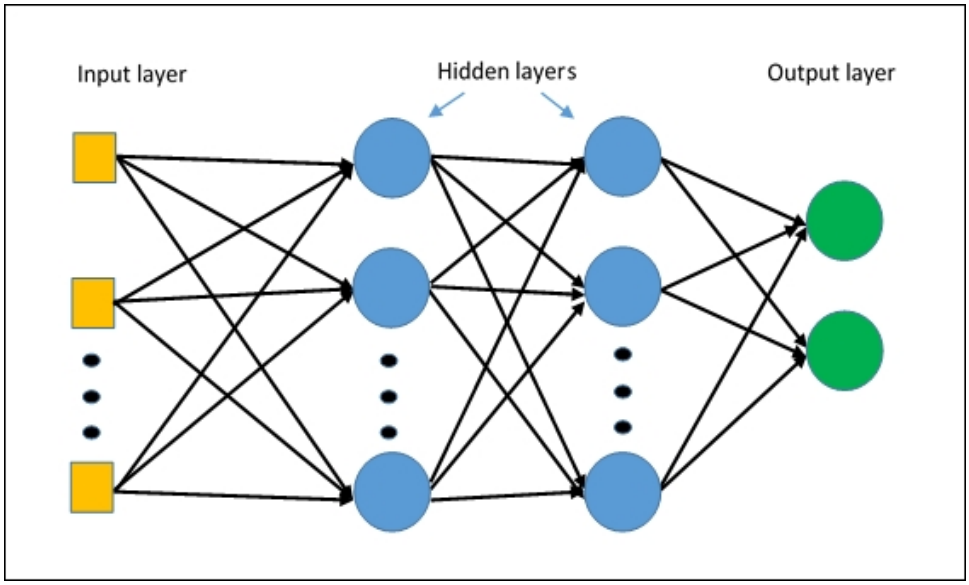


Figure 2.3: Neural Network Layer[29]

2.6 Feed Forward Neural Nets

A neural network is employed to compute a function f that yields a collection of outputs given a collection of inputs.The nomenclature ”feed forward network” is derived from the absence of input updates affecting the output[32].

$$net_j = \sum_{i=1}^{d} x_i w_i j + w_j 0 = \sum_{i=0}^{d} x_i w_{ij}$$

$$\Rightarrow y_j = f(net_j) \tag{2.2}$$

Here f is activation function.So,

$$net_k = \sum_{j=1}^{nH} y_j w_{kj} + w_{k0} = w_k^T y \tag{2.3}$$

Let nH denote the count of perceptrons existing in the hidden layers, while w_0 represents the bias units.

Given the input neurons and their corresponding weights, we can calculate the output neurons z.:

$$z_k = f(net_k) = sgn(net_k)$$

$$\Rightarrow z_k = f(\sum_{j=1}^{nH} w_{kj} f(\sum_{i=1}^{d} x_i w_{ji} + w_{j0}) + w_{k0}) \tag{2.4}$$

2.7 Artificial Neural Net Learning Process

Each prediction made by a neural network is associated with a cost function. The aforementioned cost function is commonly denoted as a loss function in computer science. In relation to this matter, it is imperative to optimize the aforementioned loss function based on the hidden weights and inputs. In order to attain the optimal cost function, it is necessary to compute the partial derivative of the cost function with respect to each weight and subsequently update the weights. Please be advised that in the context of neural networks, it is possible to repeatedly pass a single input sample through the network in order to calculate and optimize the weights associated with that particular input.This task necessitates a significant amount of computational resources.

The partial derivative with respect to each weight can be computed using back-propagation in a single backward pass. In accordance with our current understanding, it is possible to assert[33]:

$$J(w) = \frac{1}{2}\sum_{k=1}^{c}(t_k - z_k)^2 = \frac{1}{2}||t - z||^2 \tag{2.5}$$

here J(w) is the loss function, c = total output & t=output and z =predicted output

$$\Delta w = -\eta \frac{\delta J}{\delta w}$$

$$\Rightarrow \Delta w_{pq} = -\eta \frac{\delta J}{\delta w_{pq}}$$

$$\Rightarrow \Delta w_{kj} = -\eta \frac{\delta J}{\delta w_{kj}} = -\eta \frac{\delta J}{\delta net_k} * \frac{\delta net_k}{\delta w_{kj}} \&\& where, net_k = \sum_{j=1}^{nH} y_i w_{kj} + w_{k0}$$

$$\Rightarrow \Delta w_{kj} = -\eta(\frac{\delta J}{\delta z_k} * \frac{\delta z_k}{\delta net_k}) * y_j \&\& where, z_k = f(net_k)$$

$$\Rightarrow \Delta w_{kj} = \eta(t_k - z_k) * f(net_k) * y_j \tag{2.6}$$

here η is learning rate. As with the other weights,

$$\Rightarrow \Delta w_{ji} = -\eta \frac{\delta J}{\delta w_{ij}} = -\eta \frac{\delta J}{\delta net_j} * \frac{\delta net_J}{\delta w_{ji}}$$

$$\Rightarrow \Delta w_{ji} = -\eta (\frac{\delta J}{\delta y_j} * \frac{\delta y_j}{\delta net_j}) * \frac{\delta net_j}{\delta w_{ji}}$$

$$\Rightarrow \Delta w_{ji} = -\eta ((\frac{\delta J}{\delta z_k} * \frac{\delta z_k}{\delta net_k} * \frac{\delta net_k}{\delta y_j}) * \frac{\delta y_i}{\delta net_j}) * \frac{\delta net_j}{\delta w_{ij}}$$

$$\Rightarrow \Delta w_{ji} = -\eta ((-\sum_{k=1}^{c} (t_k - z_k) * f'(net_k) * w_{kj}) * f'(net_j)) * x_i \tag{2.7}$$

In the context of the backward pass and error function, the weights undergo modification.The weight update process:

$$w(m+1) = w(m) + \Delta w \tag{2.8}$$

Through the process of back-propagation, utilizing stochastic gradient descent or its various adaptations, the weights are calculated.

2.8 Forward-Propagation

Forward propagation is a fundamental concept in the realm of artificial neural networks, serving as the underlying principle governing the flow of information within the network when engaging in the task of making predictions or classifications. The process entails the computation of consecutive layers of interconnected nodes, referred to as neurons, wherein each neuron executes a weighted summation of its inputs and applies an activation function to generate an output.

Commencing from the input layer, wherein the unprocessed data or characteristics are introduced, the computed values are subsequently propagated through the concealed layers of the network. Within every concealed layer, the weights that are allocated to the connections linking neurons ascertain the magnitude of impact that one neuron exerts upon another. The weighted

inputs are aggregated, and the activation function is applied to introduce non-linearity, enabling the neural network to effectively capture intricate relationships within the data.

In the typical architecture of a network, the ultimate layer is responsible for generating the output. This output can take the form of a probability score for classification tasks or a continuous value for regression tasks. During the process of forward propagation, as data is passed through the layers of the network, the network undergoes a learning process where it generates progressively more abstract representations of the initial input. This leads to the emergence of features that are pertinent to the given task.

In essence, forward propagation serves as the primary stage in the processing of a neural network. During this stage, input data undergoes a series of transformations as it passes through interconnected neurons arranged in layers. Each neuron applies weights and activation functions to the input data, culminating in the generation of an output that serves as the network's prediction. This procedure establishes the framework for subsequent actions such as loss calculation and backward propagation, which collectively empower the network to acquire knowledge and enhance its predictive abilities through training[34].

2.9 Back-Propagation

Backward propagation, colloquially referred to as backpropagation, assumes a pivotal role during the training process of artificial neural networks. It performs a crucial function in modifying the weights of the network in order to minimize the discrepancy between predicted outputs and the desired target values. Through the process of backpropagation, the gradient of a selected loss function is computed with respect to the weights of the network. This enables the determination of the individual contributions of each weight to the overall error.

Commencing from the output layer and proceeding in a reverse manner through the layers of the network, the gradients are calculated layer by layer by employing the chain rule of calculus. The gradients serve as indicators of both the direction and magnitude of adjustments required to minimize the error in predictions. By utilizing optimization algorithms, such as gradient descent, the weights of the network are subsequently modified in the direction opposite to the gradients. This process gradually guides the weights towards values that yield enhanced predictions.

Backpropagation is an essential mechanism that facilitates the learning and refinement of in-

ternal representations in neural networks. It performs synaptic weight adjustments, optimizing their respective influences on the global prediction mechanism. By repetitively iterating over training examples and propagating them through the network, computing gradients, and adjusting weights, the network progressively improves its capability to capture complex patterns and correlations within the data, thereby enhancing its proficiency in accurately predicting outcomes for novel, unseen examples[35].

2.10 Activation Function

Activation functions play a crucial role in artificial neural networks (ANNs) by introducing non-linearity into the computations performed by the network. The output of individual neurons or nodes is determined by computing the weighted sum of their inputs. Activation functions are of utmost importance in facilitating artificial neural networks (ANNs) to acquire knowledge and represent intricate associations within datasets. Here is a concise overview of several prevalent activation functions[36]:

1. **ReLU (Rectified Linear Unit):** The Rectified Linear Unit (ReLU) is extensively employed in the field of computer science due to its straightforwardness and efficacy in mitigating the issue of the vanishing gradient. The program returns the input value if it is positive, and returns zero if it is negative. There exist alternative forms of the Rectified Linear Unit (ReLU) activation function, such as Leaky ReLU, which produces a small negative output for negative inputs, and Parametric ReLU, which enables the learning of the negative slope[36].

2. **Sigmoid Function:** The sigmoid activation function is a mathematical function that maps input values to a bounded range of values between 0 and 1. It is frequently employed during the initial stages of neural network development, yet it exhibits certain constraints such as the occurrence of vanishing gradients and output saturation[36].

3. **Softmax:** The softmax function is commonly employed in the output layer of classification networks. Its purpose is to transform a vector of raw scores into a probability distribution. It guarantees that the total sum of probabilities equals 1, thereby facilitating the process of class selection[36].

4. **Swish:** An emerging activation function that amalgamates attributes from Rectified

Linear Unit (ReLU) and sigmoid. It permits the passage of certain negative values while delivering more continuous gradients compared to the Rectified Linear Unit (ReLU)[36].

5. **Tanh (Hyperbolic Tangent):** Analogous to the sigmoid function, this function maps input values to a range spanning from -1 to 1. It aids in the mitigation of the vanishing gradient problem associated with the sigmoid function, while simultaneously providing more robust gradients[36].

2.11 Loss Function

The loss function quantifies the disparity between the desired outcome variable and the neural network's generated output. There exist three distinct types[37].

- **Regressive Loss:** Regressive loss functions are utilized in cases where the target variable is of a continuous nature. The Mean Square Error (MSE) is commonly employed. The concepts of absolute error and smooth absolute error are additional alternatives.
- **Classification Loss Function:**A classification loss function is utilized in cases where the output variable represents the probability of a specific class. Most classification loss functions exhibit a propensity to increase the margin. Several widely recognized terms in the field include margin classifier, categorical cross-entropy, and negative log-likelihood.
- **Embedding Loss Function:** Embedding loss functions are utilized for the purpose of evaluating the degree of similarity between multiple inputs. The L1 hinge error and cosine error are two frequently utilized loss functions for embedding.

2.12 Literature Survey

Machine learning is a computational approach that enables the identification of optimal solutions for a given problem without the need for explicit programming by a human programmer or experimenter [38]. There exists a plethora of machine learning algorithms that can be utilized for the purpose of predicting and diagnosing breast cancer. Several machine learning algorithms commonly used in the field include Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors (KNN Network), among others.

Many researchers have undertaken breast cancer research by utilizing multiple datasets, including the SEER dataset, mammogram images as a dataset, the Wisconsin Dataset, and datasets obtained from various hospitals. By leveraging the dataset, the authors employ various techniques to extract and select relevant features, thereby facilitating the completion of their research. These findings exhibit considerable significance in the realm of research. In the paper by Sudarshan Nayak [39], a comparison is made between different supervised machine learning algorithms for classifying breast cancer using 3D images. The results indicate that Support Vector Machines (SVM) outperforms other algorithms in terms of overall performance. The Breast Cancer Wisconsin dataset, as documented by Wolberg in 2019 [40], is a publicly available and extensively utilized dataset within the field of machine learning research.

According to the research conducted by B.M. Gayathri [41], a comparative analysis was performed on the Relevance Vector Machine (RVM) in the context of breast cancer detection. The study revealed that RVM offers a significant advantage in terms of computational efficiency when compared to other machine learning techniques. Furthermore, the research highlights the superiority of RVM over alternative machine learning algorithms in accurately diagnosing breast cancer, even when the number of variables is reduced. The achieved accuracy rate was reported to be 97%.

S. Sidhu conducted a study [42] to evaluate the effectiveness of Support Vector Machine, Artificial Neural Network, and Naïve Bayes algorithms in analyzing the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset. The study also involved integrating these machine learning techniques with feature selection and feature extraction methods to identify the most appropriate approach. Based on the simulation outcomes, it was observed that SVM-LDA outperformed all other approaches in terms of computational efficiency.

Vikas Chaurasia and Saurabh Pal conducted a comparative analysis of the performance of different classifiers in the domain of supervised learning, as documented in the reference [40]. The findings indicate that the Support Vector Machine – Radial Basis Function (SVM-RBF) kernel demonstrates the highest level of accuracy among the classifiers, achieving an accuracy rate of 96.84%.

In the study conducted by Hiba Asri [43], it was shown that the Support Vector Machine (SVM) algorithm demonstrates high efficiency in the prediction and diagnosis of Breast Cancer. The SVM algorithm achieves superior performance in terms of precision and low error rate, with an accuracy rate of 97.13%. In recent literature, a study conducted by Youness Khoudfi and

Mohamed Bahaj [44] presents a comparative analysis of various machine learning algorithms. The authors discovered that the Support Vector Machine (SVM) outperformed other classifiers such as K-Nearest Neighbors (K-NN), Random Forest (RF), and Naive Bayes (NB). The SVM achieved an impressive accuracy rate of 97.9%. The authors employed a Multilayer Perceptron (MLP) model with five layers and performed ten-fold cross-validation using MLP.

In their research paper titled ”Breast Cancer Detection: A Comparative Study of Artificial Neural Networks (ANN) and Support Vector Machines (SVM)”, Kalyani Wadkar, Prashant Pathak, and Nikhil Wagh conducted an analysis on the performance of different classifiers including Convolutional Neural Networks (CNN), K-Nearest Neighbors (KNN), and Inception V3. The objective was to enhance the processing of the WBCD dataset by integrating these classifiers. The paper is referenced as [45]. Based on the experimental results and performance analysis, it can be concluded that Artificial Neural Network (ANN) outperformed Support Vector Machine (SVM) as a classifier. ANN demonstrated a higher efficiency rate compared to SVM.

The researcher Latchoumiet TP [46] discovered a classification accuracy of 98.4% by suggesting an optimization technique called Weighted Particle Swarm Optimization (WPSO) that relies on the Support Vector Machine (SVM) with Structural Support Vector Machines (SSVM) for classification purposes.

S. Aruna and L. V. Nandakishore have conducted a comparative analysis and determined that the Support Vector Machine (SVM) classifier achieved the highest accuracy of 96.99% [47].

Ahmed Hamza Osman [48] put forth a proposition for the diagnosis of Wisconsin breast cancer (WBCD) utilizing an SVM algorithm. The solution achieved a prediction accuracy of 99.10% by amalgamating a clustering algorithm with a proficient probabilistic vector support machine. Our research is centered on evaluating machine learning algorithms and methodologies to determine the optimal approach for predicting and diagnosing breast cancer.

2.13 Conclusion

In the pursuit of precise and dependable breast cancer categorization, the thorough investigation of various machine learning algorithms serves as a pivotal cornerstone for this research endeavor. By conducting a thorough examination of algorithms such as Logistic Regression,

Support Vector Machines (SVM), Naive Bayes, Decision Trees, k-Nearest Neighbors (KNN), Neural Networks, XGBoost, and Random Forest, we have acquired a comprehensive comprehension of their fundamental mechanisms and capabilities. Furthermore, the meticulous examination of the current body of literature has yielded invaluable revelations regarding the collaborative endeavors of the research community in implementing said algorithms within the intricate realm of breast cancer classification. Through an extensive exploration of diverse research works sourced from various publishers, we have discerned a vast array of methodologies, feature engineering techniques, and evaluation metrics utilized in studies pertaining to the classification of breast cancer. This preliminary investigation acts as a foundational step, establishing the basis for our endeavor to make additional contributions to this domain by amalgamating these algorithms with customized methodologies and inventive tactics. The fusion of theoretical knowledge and practical insights acquired from algorithmic comprehension and literature review shall steer our quest to enhance the precision and effectiveness of breast cancer detection using machine learning models.

Chapter 3

Materials & Methodologies

3.1 Introduction

This chapter describes the methodologies and techniques used in our experiment or research work, choosed data-set and it's description, data pre-processing etc. The overall architecture of the whole project containing different concepts and techniques is illustrated here.

3.2 Dataset Acquisition

The dataset known as the Breast Cancer Wisconsin (Diagnostic) dataset, or WBCD dataset for short, was created and presented by Dr. William H. Wolberg in conjunction with fellow health-care experts[49]. The dataset was generated with the objective of facilitating the identification of breast cancer through the utilization of characteristics extracted from fine-needle aspirates of breast tissue.

The data acquisition process entails acquiring cytological samples via fine-needle aspiration, a procedure with minimal invasiveness that employs a thin, hollow needle to extract tissue samples from potentially malignant breast lesions[50]. The provided specimens are subsequently subjected to microscopic analysis in order to ascertain the cellular characteristics and discern their classification as either malignant (possessing cancerous properties) or benign (devoid of cancerous attributes).

The WBCD dataset involved the extraction of a diverse range of features from microscopic images. The aforementioned features encompass quantifiable attributes of cell nuclei, including but not limited to radius, texture, smoothness, compactness, concavity, and symmetry. These

attributes are quantified in order to obtain numerical values that can be used for the purpose of analysis and modeling.

It is of utmost significance to acknowledge that the dataset does not encompass the unprocessed images per se, but rather the computed features derived from said images. The dataset was compiled through the process of aggregating feature information from multiple instances of fine-needle aspiration cases. This methodology enables researchers and practitioners in the field of machine learning to manipulate measurable data that encompasses essential attributes that are suggestive of either malignancy or benignity.

Breast Cancer Wisconsin Dataset(WBCD) is publicly available different online area such as UCI (University of California, Irvine) machine learning repository or Kaggle repository. The data set for this research work is collected from UCI machine learning repository.

3.3 Dataset Description

The dataset known as the Breast Cancer Wisconsin dataset has gained significant popularity in the fields of machine learning and medical research. It serves as a standard reference for the creation and assessment of predictive models aimed at aiding in the detection and diagnosis of breast cancer. It empowers researchers to investigate the potential of machine learning algorithms in assisting medical professionals in formulating precise diagnoses using a collection of precisely defined characteristics derived from microscopic samples.

The dataset consists of 569 instances, where each instance corresponds to a breast tissue sample. Each instance has 32 attributes and 2 additional attributes(instance ID and class Name). The Data set contains binary class. One if Malignant and another is Benign. Malignant means cancerous. Among 569 instances, no of malignant case is 212 and no of benign case is 357[50].

Samples	Benign	Malignant	Attributes	Additional attributes
569	357	212	30	2

Table 3.1: Description of the Data Set.

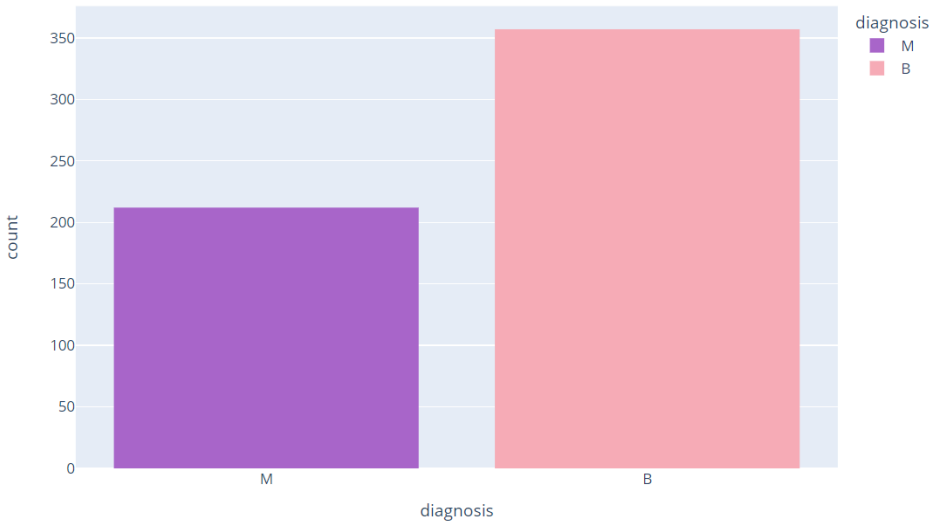


Figure 3.1: Dataset Overview Using Barchart

3.4 Pre-processing

Data preprocessing is an essential phase within the machine learning pipeline wherein raw data is cleaned, transformed, and organized into a suitable format for the purpose of training and evaluating machine learning models. Efficient data preprocessing is imperative for enhancing model performance and guaranteeing the dependability of outcomes[51].

Some techniques of data preprocessing are data cleaning, data transformation, data reduction, normalization and standardization, handling outliers, handling imbalanced data, data splitting etc.

In the experiment, we have almost a perfect data set. We just need to do just two operations on the data set:

- **Data Reduction:** Additional two columns, instance ID and Diagnosis(Benign or Malignant) are dropped to form training set of data.. Some other attributes are removed based upon the correlation threshold value.
- **Data Splitting:** Data splitting is done for building the models.
- **Data Transformation:** Data transformation is used to bring all in a same scale, that is making standard deviation 1 for every attribute and this centers the data around 0(zero). Machine learning algorithm run on this type of data very well and give better results and precision[52].

3.5 Data Splitting

Data splitting is an essential process in the field of machine learning, wherein a dataset is partitioned into separate subsets to enable the training, tuning, and evaluation of models. The prevalent divisions typically involve the segregation of data into training, validation, and testing subsets, albeit the specific partitioning ratio may fluctuate contingent upon the magnitude and characteristics of the dataset[53].

Here, in this experiment, Total samples are splitted into two category, training and testing.

Total samples	Training samples	Testing samples
569	456	113

Table 3.2: Data splitting

- Training samples: 456 (almost 80% of total samples).
- Testing samples: 113 (almost 20% of total samples).

They are randomly selected.

3.6 Hyper Parameter Tuning

3.6.1 Batch Size

The concept of batch size refers to the aggregate count of samples contained within a solitary batch. It is a hyperparameter that governs the batch size, which specifies the number of samples to be processed simultaneously. The complete dataset is partitioned into multiple batches to enhance and optimize the training process. The batch size utilized in this context is 32 for both classification tasks[54].

3.6.2 Epoch

In the context of computational processes, epochs typically refer to the sequential iterations or cycles. An epoch is completed when the model undergoes a full pass of the entire dataset in both the forward and backward directions. A singular epoch may consist of one or multiple batches. The occurrence of a single epoch initiates a situation where the curve in the graph begins to

exhibit underfitting. As the value of the epoch variable increases, the weight parameters in the neural network undergo a greater number of adjustments. Consequently, the curve representing the model's performance transitions from an underfitting state to an optimal state and eventually to an overfitting state[54].

3.6.3 Learning rate

The learning rate is a parameter that determines the speed at which the model will iterate.In our computational model, the learning rate was set to 0.001.This is commonly referred to as the model's learning rate.The hyperparameter governs the learning rate or the velocity at which the model acquires knowledge. The error allocation regulation mechanism determines the extent to which the model's weights are updated after each update event, such as at the completion of a batch of training instances[55].

The model shall acquire the optimal ability to approximate the function, taking into account the resources at hand - specifically, the quantity of layers and nodes within each layer, along with the designated number of training epochs. This shall be achieved under the condition that the learning rate is impeccably calibrated.An optimal learning rate is one that is sufficiently small to ensure convergence of the network towards a valuable solution, while also being sufficiently large to facilitate training within a reasonable timeframe.

Reduced learning rates require an increased number of training epochs due to the diminished magnitude of modifications. However, higher values of learning rates lead to accelerated modifications.

3.7 Environment and Platform

We have used Kaggle as our Platform to implement this project. Kaggle is an extensively acknowledged and prevalent online platform that functions as a central repository for individuals passionate about data science, machine learning, and artificial intelligence. Kaggle, which was procured by Google in the year 2017, provides a highly interactive and community-oriented platform where individuals with expertise in data science, researchers, and practitioners hailing from various geographical locations engage in collaborative efforts to tackle intricate problems[56]. This platform also facilitates participation in competitions and the exchange of valuable insights and knowledge.

Language: Python programming language is used to write the necessary program for the project. Python is very popular, effective and efficient language for machine learning and data science field[57]. Python has experienced significant adoption and is widely recognized as a highly efficient programming language in the domain of machine learning owing to various pivotal factors:

1. **Large Community Support:** The extensive user base of Python has resulted in a dynamic and lively community. The provision of community support guarantees an abundance of resources, documentation, tutorials, and online forums that facilitate user assistance and knowledge sharing.

2. **Rich Ecosystem of Libraries:** Python is equipped with a vast array of libraries and frameworks that have been meticulously designed to cater to the specific needs of machine learning and data analysis. The aforementioned libraries, namely NumPy, pandas, scikit-learn, TensorFlow, and PyTorch, offer a plethora of robust tools that facilitate various tasks such as data manipulation, preprocessing, modeling, and deep learning.

3. **Ease of Use and Readability:** The syntax of Python exhibits clarity and conciseness, bearing resemblance to constructs found in natural language. The enhanced readability of this programming paradigm facilitates the ease of code composition, debugging, and codebase maintenance for developers, researchers, and data scientists. Consequently, it mitigates the initial challenges faced by novices when acquiring proficiency in this domain.

4. **Open Source Philosophy:** A multitude of fundamental libraries and frameworks employed in Python's machine learning ecosystem adhere to the principles of open source. This facilitates cooperation, novelty, and development driven by the community.

5. **Flexibility and Versatility:** The versatility of Python empowers developers to effortlessly incorporate machine learning models into web applications, databases, APIs, and various other technologies. The inherent adaptability of this characteristic proves to be particularly beneficial when implementing models in practical, real-world scenarios.

RAM capacity: Kaggle provides me 13 GB RAM(Random access memory)

ROM/HD: Kaggle provides 73 GB Disk(ROM - Read only memory).

3.8 The machine learning models used in this project:

We have used different types of machine learning algorithms in this project. Almost of of them are included in the python libraby 'sklearn', which is abbreviated form of 'scikit-learn'.

Scikit-learn: Scikit-learn, frequently shortened as sklearn, represents a widely used open-source machine learning library designed for the Python programming language. It is constructed upon existing libraries such as NumPy, SciPy, and Matplotlib, and furnishes an extensive assortment of utilities for diverse machine learning undertakings, thereby establishing itself as an essential asset for data scientists, researchers, and machine learning professionals.

- **Logistic Regression:** We have imported this from the Scikit-learn library and run the algorithm against the preprocessed data.
- **Decision Tree:** We have imported this from the Scikit-learn library and run the algorithm against the preprocessed data.
- **K Nearest Neighbor:** We have imported this from the Scikit-learn library and run the algorithm against the preprocessed data.
- **Random Forest Classifier:** We have imported this from the Scikit-learn library and run the algorithm against the preprocessed data.
- **Naive Bayes:** We have imported this from the Scikit-learn library and run the algorithm against the preprocessed data.
- **Neural Network:** At first, we have built this algorithm. Then training dataset is applied to the algorithm and testing data set is applied to test the model.
  - **Epoch:** An epoch is defined as the point at which the model has iterated through and processed each instance in the training dataset, thereby acquiring knowledge from all examples. We have used total 140 epoch
  - **The number of Layers:** There are total 3 hidden layers and one input layer and one output layer.

Illustration Of Overall Processes: Here is an illustration from data acquisition to model building and evaluation:

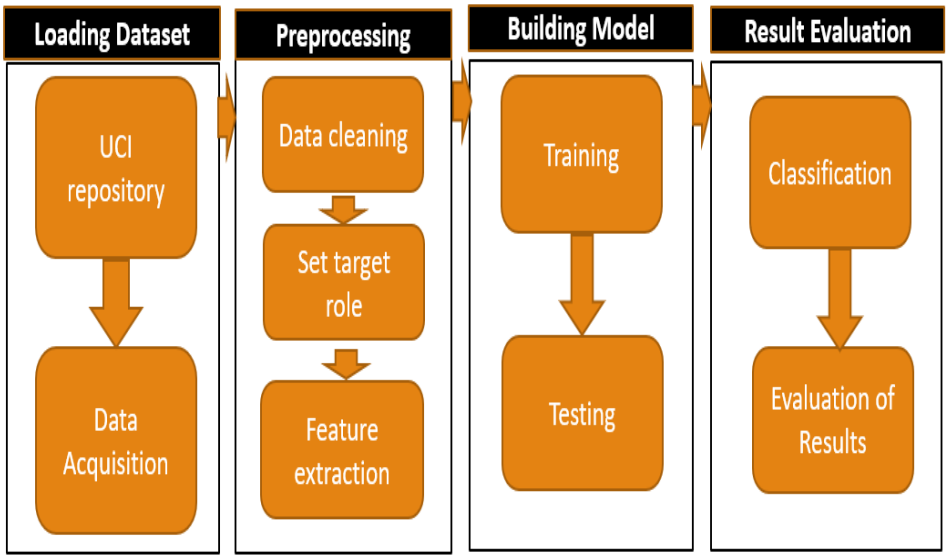


Figure 3.2: Illustration of overall processes

Here, we have done data acquisition from UCI(University of California, Irvine) machine learning repository, which is publicly available source. Then some pre-processing on the dataset has been done to increase effectivity and efficiency on the result of applying different machine learning algorithm. To build the models, training and testing has been done. Then we have done evaluation of the results. We can say that this is the blue-print or architecture of the research work.

3.9 Conclusion

In this chapter, we have about the main methodologies and processed used to implement the project successfully. Methodologies and materials are one of the most essential part of a research work. The is called the foundation of the experiment. Here, it has also been described that how the parameters of the machine learning algorithms are initalized and measured.

Chapter 4

Results and Performance Analysis

4.1 Introduction

Different assessment criteria for the effectiveness of machine learning models are explained in depth in this chapter. The outcome analysis of the machine learning models is then provided based on these performance measurement components. Thus, the complete experimental analysis of this study project is covered in this chapter.

4.2 Evaluation Metrics

Within the realm of machine learning, evaluation metrics, also known as evaluation measures, are numeric quantities employed to appraise the efficacy and caliber of a machine learning model[58]. These metrics offer valuable insights regarding the performance of the model on a particular task, such as classification, regression, clustering, or any other form of predictive analysis.An epoch is defined as the point in the training process where the model has iterated through and processed each example in the training dataset exactly once, thereby acquiring knowledge from each instance.

Evaluation metrics facilitate the quantification of various facets of model performance, thereby enabling us to make well-informed decisions regarding model selection, parameter tuning, and algorithm comparison. The selection of an evaluation metric is contingent upon the inherent characteristics of the problem at hand and the objectives of the analysis.

To evaluate the proposed Convolutional Neural Network (CNN) model, we employ metrics such as Precision (PR), Recall (RC), Accuracy (AC), Specificity (SP), F1-score (F1), and the confusion matrix. Prior to elucidating all evaluation metrics, it is imperative to explicate the concepts of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP (True Positive) is the result in which both the ground value and the outcome are positive, while TN (True Negative) represents the scenario where both of them are negative. False Positive (FP) is defined as the occurrence of a positive result when the actual ground truth is negative. On the other hand, False Negative (FN) refers to the occurrence of a negative result when the actual ground truth is positive.

4.2.1 Confusion Matrix

The confusion matrix holds significant importance within the realm of machine learning, specifically in the assessment of classification model efficacy. The provided analysis offers a precise and exhaustive elucidation of the model's prognostications by juxtaposing them with the factual ground truth values. The utilization of the confusion matrix is a common practice in the computation of diverse evaluation metrics for classification endeavors[59]. These metrics encompass accuracy, precision, recall, and F1-score.

Let us deconstruct the fundamental terminologies and constituents of a confusion matrix:

- **True Positives (TP):** The count of instances accurately classified as positive by the algorithm.
- **True Negatives (TN):** The count of instances accurately classified as negative by the algorithm.
- **False Positives (FP):** The count of instances that were erroneously classified as positive by the model despite being negative. Also referred to as a "Type I Error" in the field of computer science.
- **False Negatives (FN):** The count of instances that were erroneously classified as negative by the model despite being positive. Also referred to as a "Type II Error."

Through the examination of the confusion matrix and its corresponding metrics, we can acquire a more profound comprehension of the performance of the algorithm and detect domains in

which it may exhibit proficiency or encounter challenges in generating precise predictions.

4.2.2 Accuracy

The accuracy is determined by performing a division operation on the count of correctly predicted sample and the count of total predicted sample, encompassing both accurate and inaccurate predictions[60]. The mathematical expression used to compute accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

4.2.3 Precision

To compute the precision value, we perform the division of the total count of accurately predicted positive examples by the total count of predicted positive examples. A higher level of precision implies that an object that is explicitly defined as positive will definitely be classified as positive[61].

$$Precision(PR) = \frac{TP}{TP + FP} \tag{4.2}$$

4.2.4 Recall

The percentage of positive occurrences that the model properly recognized as being positive in relation to the overall number of positive instances is known as recall. The proportion of accurately anticipated positive samples to all positive samples is known as the recall measure. When the recall rate is high, the class has been correctly identified. The value should be increased to the maximum possible level[61].

$$Recall(RC) = \frac{TP}{TP + FN} \tag{4.3}$$

4.2.5 Specificity

The specificity (SP) metric quantifies the proportion of correctly predicted negative items out of all the negative data. The formula for calculating the specificity value is as follows:

$$Specificity(SP) = \frac{TN}{TN + FP} \tag{4.4}$$

4.2.6 F1-Score

Distinguishing between two models with high precision and low recall, or vice versa, can be a challenging task. To achieve relativity, the F1-Score is implemented. The F1-score facilitates the computation of Precision and Recall simultaneously. The Harmonic Mean is employed in lieu of the Arithmetic Mean by rectifying the absolute values[62].

$$F1 - Score = \frac{PR * RC}{PR + RC} * 2 \tag{4.5}$$

4.3 Performance Analysis of Different Models

The criteria of evaluation of performance are previously described in detail with necessary mathematical equation and reasoning. Here, Using these criteria, performance for each model is analyzed.

4.3.1 Logistic Regression

The confusion matrix is given here for Logistic Regression model.

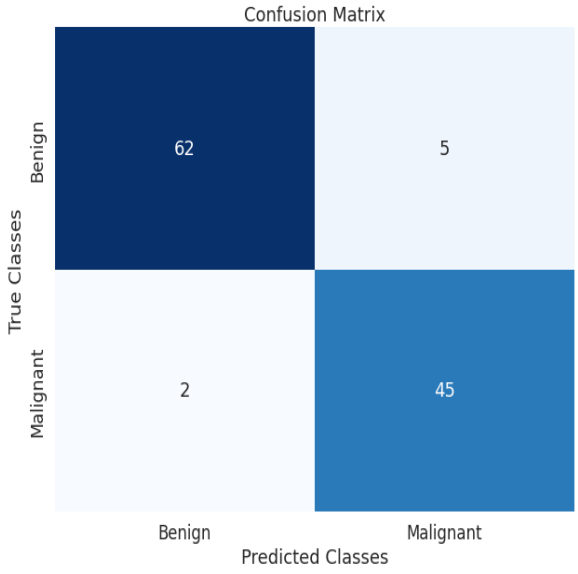


Figure 4.1: Confusion Matrix for Logistic Regression Model

We know, we can derive other measurement criteria from this confusion matrix to analyze the model's performance. These are given below:

- **Accuracy = 0.938596**
- **Precision = 0.900000**
- **F1-Score = 0.927835**
- **Recall = 0.957447**

Tabular Illustration:

Table 4.1: Performance for the Logistic Regression

Model Name	Accuracy	Precision	F1-Score	Recall
Logistic Regression	0.938596	0.900000	0.927835	0.957447

4.3.2 Decision Tree

The confusion matrix is given here for Decision Tree model.

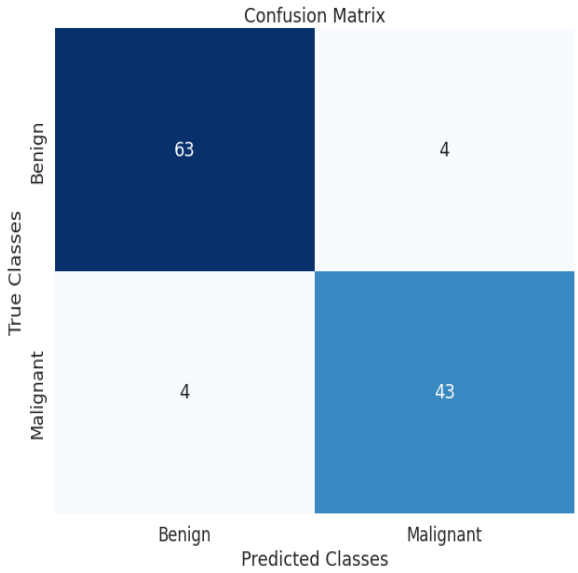


Figure 4.2: Confusion Matrix for Decision Tree Model

We know, we can derive other measurement criteria from this confusion matrix to analyze the model's performance. These are given below:

- **Accuracy = 0.929825**
- **Precision = 0.914894**
- **F1-Score = 0.914894**

- **Recall = 0.914894**

Tabular Illustration:

Table 4.2: Performance for the Decision Tree

Model Name	Accuracy	Precision	F1-Score	Recall
Decision Tree	0.929825	0.914894	0.914894	0.914894

4.3.3 K-Nearest Neighbors

The confusion matrix is given here for Decision Tree model.

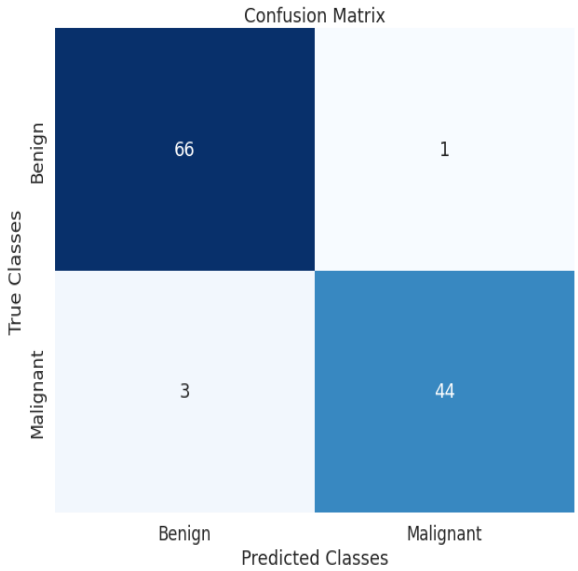


Figure 4.3: Confusion Matrix for K-Nearest Neighbors Model

We know, we can derive other measurement criteria from this confusion matrix to analyze the model's performance. These are given below:

- **Accuracy = 0.964912**
- **Precision = 0.977778**
- **F1-Score = 0.956522**
- **Recall = 0.936170**

Tabular Illustration:

Table 4.3: Performance for the K-Nearest Neighbors

Model Name	Accuracy	Precision	F1-Score	Recall
K-Nearest Neighbors	0.964912	0.977778	0.956522	0.936170

4.3.4 Random Forest Classifer

The confusion matrix is given here for Random Forest Classifer.

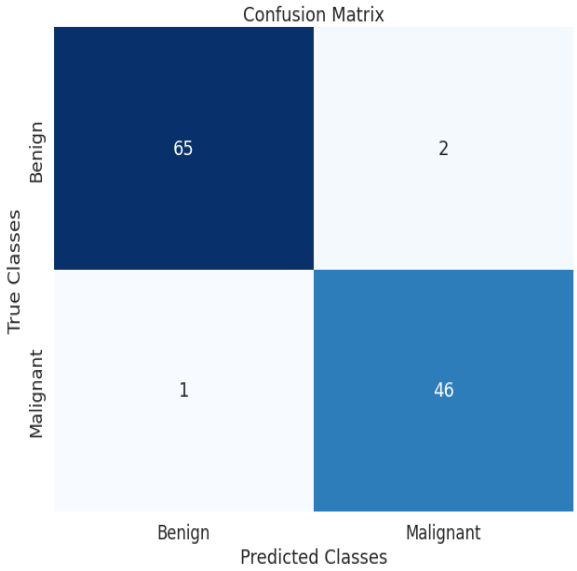


Figure 4.4: Confusion Matrix for Random Forest Classifier Model

We know, we can derive other measurement criteria from this confusion matrix to analyze the model's performance. These are given below:

- **Accuracy = 0.973684**
- **Precision = 0.958333**
- **F1-Score = 0.968421**
- **Recall = 0.978723**

Tabular Illustration:

Table 4.4: Performance for the Random Forest Classifer

Model Name	Accuracy	Precision	F1-Score	Recall
Random Forest Classifer	0.973684	0.958333	0.968421	0.978723

4.3.5 XGBoost

The confusion matrix is given here for XGBoost model.

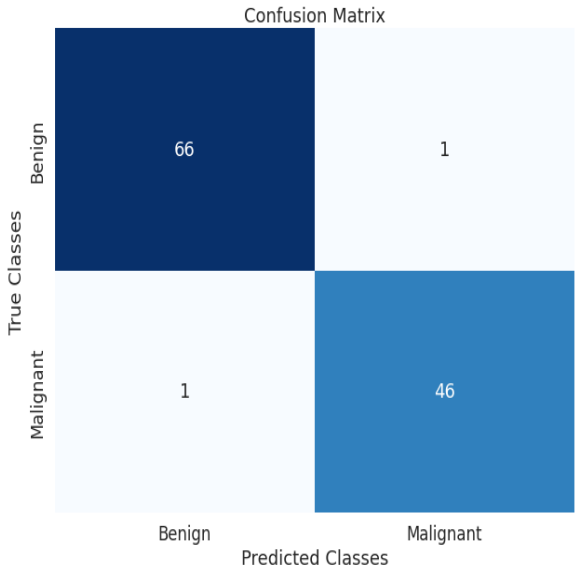


Figure 4.5: Confusion Matrix for XGBoost Model

We know, we can derive other measurement criteria from this confusion matrix to analyze the model's performance. These are given below:

- **Accuracy = 0.982456**
- **Precision = 0.978723**
- **F1-Score = 0.978723**
- **Recall = 0.978723**

Tabular Illustration:

Table 4.5: Performance for the XGBoost

Model Name	Accuracy	Precision	F1-Score	Recall
XGBoost	0.982456	0.978723	0.978723	0.978723

4.3.6 Naive Bayes

The confusion matrix is given here for Naive Bayes model.

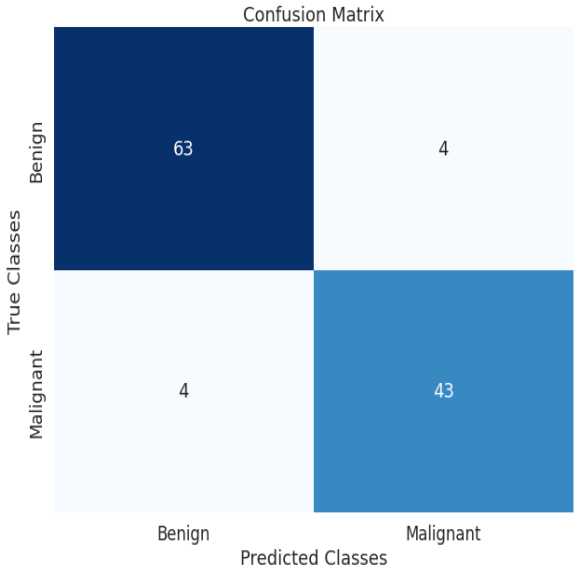


Figure 4.6: Confusion Matrix for Naive Bayes

We know, we can derive other measurement criteria from this confusion matrix to analyze the model's performance. These are given below:

- **Accuracy = 0.929825**
- **Precision = 0.914894**
- **F1-Score = 0.914894**
- **Recall = 0.914894**

Tabular Illustration:

Table 4.6: Performance for the Naive Bayes

Model Name	Accuracy	Precision	F1-Score	Recall
Naive Bayes	0.929825	0.914894	0.914894	0.914894

4.3.7 Neural Network

The confusion matrix is given here for Neural Network model.

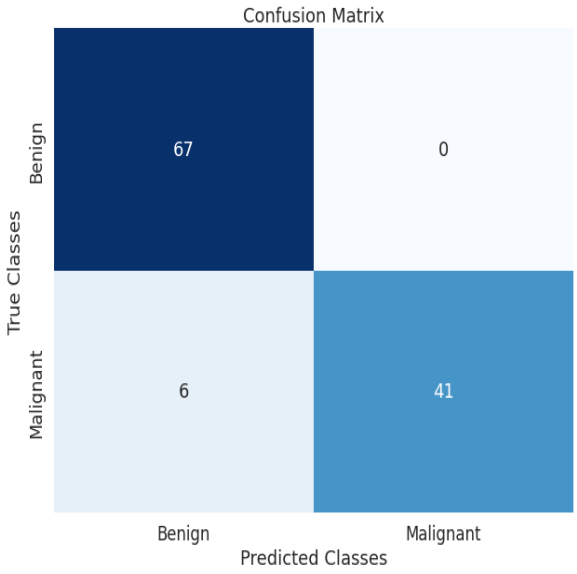


Figure 4.7: Confusion Matrix for Neural Network Model

We know, we can derive other measurement criteria from this confusion matrix to analyze the model's performance. These are given below:

- **Accuracy = 0.947368**
- **Precision = 1.000000**
- **F1-Score = 0.931818**
- **Recall = 0.872340**

Tabular Illustration:

Table 4.7: Performance for the Neural Network Model

Model Name	Accuracy	Precision	F1-Score	Recall
Neural Network	0.947368	1.000000	0.931818	0.872340

4.4 Comparative Performance Analysis Among the models

The above measurement criteria for all the models are listed here according to the descending order based on accuracy:

Table 4.8: Comparison of performance among all the used models

Model Name	Accuracy	Precision	F1-Score	Recall
XGBoost	0.982456	0.978723	0.978723	0.978723
Random Forest Classifer	0.973684	0.958333	0.968421	0.978723
K-Nearest Neighbors	0.964912	0.977778	0.956522	0.936170
Neural Network	0.947368	1.000000	0.931818	0.872340
Logistic Regression	0.938596	0.900000	0.927835	0.957447
Decision Tree	0.929825	0.914894	0.914894	0.914894
Naive Bayes	0.929825	0.914894	0.914894	0.914894
Naive Bayes	0.929825	0.914894	0.914894	0.914894

Here, we can see that among these models XGBoost model is giving the best accuracy Based on the Wisconsin Breast Cancer Dataset(WBCD) among all the models. Although, First 2 models' accuracy in the table are nearly the same almost.

In the performance analysis, accuracy is the main measurement. Because accuracy says the correctly classification rate of a machine learning model.

So, we can say that the most accurate machine learning model based on the used dataset is XGBoost model.

4.5 Conclusion

This chapter provides a comprehensive overview of accuracy, confusion matrix, and other evaluation matrices.The chapter provides a detailed analysis of the correct and incorrect predictions of the samples, utilizing a confusion matrix. The primary evaluation metrics for machine learning models, such as accuracy, precision, recall, etc., are presented in tabular format. Finally, the evaluation of the employed machine learning models has been delineated.

Chapter 5

Conclusion and Future Works

5.1 Introduction

This chapter presents a concise overview of the comprehensive study, delineating the problem domain, our distinctive contribution, the empirical analysis conducted, the findings derived from the study, and the inherent limitations. We have additionally provided a concise analysis concerning the prospective avenues for this research.

5.2 Thesis Summary

Breast cancer continues to be a highly prevalent and significant health issue that impacts women on a global scale. In this thesis, we addressed the formidable task of breast cancer classification utilizing a wide array of machine learning methodologies. The investigation centered around the widely recognized Wisconsin Breast Cancer dataset obtained from the UCI Machine Learning Repository. It utilized a carefully selected set of features based on a correlation threshold of 0.2.

The main goal of the research was to create efficient classification models that can accurately differentiate between malignant and benign breast cancer cases. The examined machine learning algorithms comprised of Naive Bayes, Logistic Regression, Decision Tree, XGBoost, Random Forest, k-Nearest Neighbors (KNN), and Neural Network.

By conducting extensive experimentation and analysis, it has been determined that the XGBoost model exhibits superior performance in terms of classification accuracy. The ensemble learning approach employed exhibited exceptional predictive capabilities, thereby yielding the

highest accuracy compared to all evaluated models. The aforementioned result highlights the effectiveness of XGBoost in managing intricate medical datasets and capturing intricate patterns that are indicative of instances of breast cancer.

Moreover, the Random Forest classifier demonstrated exceptional performance, emerging as the second-most proficient model in terms of accuracy. The methodology is based on an ensemble approach, where multiple decision trees are combined. This approach has been proven to effectively leverage the collective intelligence of diverse trees in order to generate predictions that are robust. The aforementioned outcome highlights the significance of ensemble techniques in augmenting predictive precision within the context of medical diagnosis scenarios.

The correlation-based feature selection process was instrumental in determining the input data utilized by the algorithms. By selecting characteristics that displayed notable correlations with the target variable, the investigation sought to capture the utmost pertinent data for precise categorization. The aforementioned feature selection strategy exhibited enhancements in computational efficiency and demonstrated competitive performance across all models.

In summary, this thesis highlights the fundamental significance of machine learning in the classification of breast cancer. The results not only confirm the efficacy of XGBoost as a top selection for medical diagnostic tasks, but also emphasize the capabilities of Random Forest classifiers. The examination of feature selection strategies serves to underscore the significance of domain knowledge in shaping efficacious predictive models. In the future, the knowledge acquired from this research possesses the capability to enhance breast cancer diagnosis and patient care by implementing resilient machine learning algorithms.

5.3 Contribution

The aforementioned thesis presents notable contributions in the domain of breast cancer classification through the utilization of machine learning methodologies. By means of rigorous experimentation and meticulous analysis, a thorough evaluation of diverse machine learning algorithms was performed on the Wisconsin Breast Cancer dataset. The principal contributions of this work can be succinctly summarized as follows:

- **Algorithmic Assessment:** Seven well-known machine learning algorithms were subjected to a thorough analysis, including Naive Bayes, Logistic Regression, Decision Tree, XGBoost, Random Forest, KNN, and Neural Network. The accuracy, precision, recall,

F1-score, and balanced accuracy of each algorithm's performance were all carefully assessed. This evaluation offers insightful information on the advantages and disadvantages of each approach in the context of breast cancer classification.

- **Model Compare:** This thesis provides a fair assessment of the predicting capacities of various algorithms by comparing their performance. The considerable performance differences between the XGBoost and Random Forest models emphasize the potential superiority of ensemble-based techniques in medical diagnosis tasks, offering practitioners and researchers invaluable assistance when choosing the best algorithm.

- **Strategy of Feature Selection:** A unique addition that simplifies the feature space by keeping just the most useful qualities is the use of a correlation-based feature selection technique. This method not only increases model effectiveness but also makes the outcomes easier to understand. Future research may be guided by the proven efficacy of this feature selection strategy in selecting pertinent characteristics for medical classification tasks.

- **Relevant Applications:** The research findings possess practical implications for medical professionals and decision-makers engaged in the diagnosis of breast cancer. The recognition of XGBoost and Random Forest as the most auspicious models can provide insights for the creation of precise and dependable diagnostic tools, potentially assisting healthcare professionals in making well-informed decisions and enhancing patient outcomes.

- **Promoting Research:** Through the systematic exploration of various machine learning algorithms and feature selection techniques, this thesis makes a significant contribution to the progress of breast cancer classification research. The empirical evidence presented in this study establishes a standard for future inquiries, promoting the utilization of strong methodologies to tackle the obstacles linked to early cancer detection and diagnosis.

5.4 Limitations

Although this thesis has made significant contributions to the domain of breast cancer classification, it is important to acknowledge and address certain limitations that exist. These limitations present opportunities for future research and enhancement:

- **Small Dataset:** The Wisconsin Breast Cancer dataset, although extensively utilized for breast cancer classification, possesses inherent limitations attributable to its dimensions and extent. The limited size of the dataset in terms of instances and features could potentially affect the ability to apply the findings to larger and more diverse patient populations. Future research should investigate the application of the proposed models on datasets of greater magnitude in order to authenticate their efficacy in various contexts.

- **The Sensitivity of Feature Selection:** The feature selection strategy utilized in this study is correlation-based and depends on a designated correlation threshold (0.2) for ascertaining the significance of features. Further investigation is warranted to assess the sensitivity of the results to the specified threshold and to evaluate the potential impact of alternative methods for selecting features. By systematically varying threshold values and employing domain-specific knowledge-driven feature selection techniques, it is possible to augment the resilience of the models.

- **Outside Validation:** The suggested models' generalizability and reliability across various medical institutions and patient groups must be evaluated using external, independent datasets. To validate the results seen on the Wisconsin Breast Cancer dataset in practical settings, external validation is essential.

- **Insufficient Clinical Data:** The dataset utilized in this investigation includes elements generated from fine-needle aspiration biopsies, although it is deficient in information about patients and more extensive clinical context. By obtaining more thorough and unique insights, the integration of new clinical features, genetic data, and patient history may improve the predicted accuracy of the models.

- **Imbalanced Data:** The class distribution observed in medical datasets, such as the Wisconsin Breast Cancer dataset, frequently exhibits an imbalance, wherein the prevalence of benign cases surpasses that of malignant cases. The presence of imbalanced data can have a detrimental impact on the model's capacity to effectively classify minority classes. Future research should aim to rectify this imbalance by employing methodologies such as oversampling, undersampling, or sophisticated ensemble techniques specifically designed for imbalanced datasets.

5.5 Future works

The present thesis aims to provide a comprehensive analysis of potential avenues for future research in the field of breast cancer classification and diagnostic techniques. The following domains require further investigation in order to enhance the knowledge obtained from this study:

- **Ensemble Methods:** Exploring the synergistic potential of ensemble models through the fusion of XGBoost and Random Forest methodologies has the potential to yield improved predictive accuracy. By delving into ensemble fusion techniques, such as stacking or blending, one can effectively leverage the synergistic qualities of diverse models to attain enhanced classification performance.

- **External Validation:** It is of utmost importance to extend the evaluation of the developed models to external datasets originating from various medical institutions and populations. The importance of robust external validation cannot be overstated as it serves to showcase the generalizability and practicality of the proposed methodologies in real-world scenarios.

- **Higher-Level Hyperparameter Optimization:** Performing an exhaustive hyperparameter optimization procedure customized for each algorithm has the potential to unleash their complete capabilities and further augment their predictive prowess. Methods such as Bayesian optimization or evolutionary algorithms possess the capability to effectively explore the hyperparameter space.

- **Medical Integration:** A more complete knowledge of breast cancer cases could be possible with the incorporation of clinical characteristics, genetic markers, and patient history into the categorization models. Through better patient outcomes and individualized treatment strategies, this integration may produce more precise and individualized diagnostic predictions.

5.6 Conclusion

The present thesis explores the domain of breast cancer classification through the utilization of diverse machine learning models. Through the evaluation of various algorithms such as Naive

Bayes, Logistic Regression, Decision Tree, XGBoost, Random Forest, KNN, and Neural Network, our objective was to improve the accuracy of early detection. XGBoost has emerged as a leading contender, demonstrating exceptional proficiency in capturing complex patterns to achieve accurate predictions. In the interim, Random Forest demonstrated the efficacy of ensemble methodologies. The feature selection approach driven by correlation effectively optimized data and upheld competitiveness. While constraints underscore intricacies, forthcoming investigations present opportunities for optimization.

This work provides valuable insights for both researchers and healthcare practitioners. By leveraging the potential of machine learning, we can enhance the efficiency of breast cancer diagnosis, thereby resulting in enhanced patient care and improved outcomes. As the dynamic environment undergoes continuous transformation, the methodologies employed in this study serve as a catalyst for future progress, potentially facilitating the integration of machine learning capabilities with clinical practices.

REFERENCES

[1] "Breast cancer." https://www.who.int/news-room/fact-sheets/detail/breast-cancer, 2023.

[2] F. Taleghani, Z. P. Yekta, A. N. Nasrabadi, and S. Käppeli, "Adjustment process in iranian women with breast cancer," *Cancer Nursing*, vol. 31, no. 3, pp. E32–E41, 2008.

[3] C. Mettlin, "Global breast cancer mortality statistics," *CA: a cancer journal for clinicians*, vol. 49, no. 3, pp. 138–144, 1999.

[4] D. R. Youlden, S. M. Cramb, C. H. Yip, and P. D. Baade, "Incidence and mortality of female breast cancer in the asia-pacific region," *Cancer biology & medicine*, vol. 11, no. 2, p. 101, 2014.

[5] "Breast cancer: Does stress fuel its spread?." https://www.medicalnewstoday.com/articles/322832, 2023.

[6] S. Ferro, A. Caroli, O. Nanni, A. Biggeri, and A. Gambi, "A cross sectional survey on breast self examination practice, utilization of breast professional examination, mammography and associated factors in romagna, italy," *Tumori Journal*, vol. 78, no. 2, pp. 98–105, 1992.

[7] M. A. Keitel and M. Kopala, *Counseling women with breast cancer*, vol. 5. Sage, 2000.

[8] K. D. Stein, S. C. Martin, D. M. Hann, and P. B. Jacobsen, "A multidimensional measure of fatigue for use with cancer patients," *Cancer practice*, vol. 6, no. 3, pp. 143–152, 1998.

[9] L. Wilkinson and T. Gathani, "Understanding breast cancer as a global health concern," *The British Journal of Radiology*, vol. 95, no. 1130, p. 20211033, 2022.

[10] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality world-

wide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[11] E. M. Drew and N. E. Schoenberg, "Deconstructing fatalism: ethnographic perspectives on women's decision making about cancer prevention and treatment," *Medical anthropology quarterly*, vol. 25, no. 2, pp. 164–182, 2011.

[12] I. Andersson, K. Aspegren, L. Janzon, T. Landberg, K. Lindholm, F. Linell, O. Ljungberg, J. Ranstam, and B. Sigfusson, "Mammographic screening and mortality from breast cancer: the malmö mammographic screening trial.," *British Medical Journal*, vol. 297, no. 6654, pp. 943–948, 1988.

[13] B. Kaur, S. Kumar, and B. K. Kaushik, "Recent advancements in optical biosensors for cancer detection," *Biosensors and Bioelectronics*, vol. 197, p. 113805, 2022.

[14] M. Arbyn, X. Castellsagué, S. de Sanjosé, L. Bruni, M. Saraiya, F. Bray, and J. Ferlay, "Worldwide burden of cervical cancer in 2008," *Annals of oncology*, vol. 22, no. 12, pp. 2675–2686, 2011.

[15] Z.-H. Zhou, *Machine learning*. Springer Nature, 2021.

[16] P. Cunningham, M. Cord, and S. J. Delany, "Supervised learning," in *Machine learning techniques for multimedia: case studies on organization and retrieval*, pp. 21–49, Springer, 2008.

[17] M. F. A. Hady and F. Schwenker, "Semi-supervised learning," *Handbook on Neural Information Processing*, pp. 215–239, 2013.

[18] H. B. Barlow, "Unsupervised learning," *Neural computation*, vol. 1, no. 3, pp. 295–311, 1989.

[19] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[20] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, pp. 41–46, 2001.

[21] J. M. Hilbe, *Logistic regression models*. CRC press, 2009.

[22] S. Tangirala, "Evaluating the impact of gini index and information gain on classification using decision tree classifier algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 612–619, 2020.

[23] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[24] W. S. Noble, "What is a support vector machine?," *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[25] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, 2018.

[26] O. Kramer and O. Kramer, "K-nearest neighbors," *Dimensionality reduction with unsupervised nearest neighbors*, pp. 13–23, 2013.

[27] L. Wang, C. Wu, L. Tang, W. Zhang, S. Lacasse, H. Liu, and L. Gao, "Efficient reliability analysis of earth dam slope stability using extreme gradient boosting method," *Acta Geotechnica*, vol. 15, pp. 3135–3150, 2020.

[28] S. Chung and L. Abbott, "Neural population geometry: An approach for understanding biological and artificial neural networks," *Current opinion in neurobiology*, vol. 70, pp. 137–144, 2021.

[29] "Multilayer perceptron definition." `https://deepai.org/machine-learning-glossary-and-terms/multilayer-perceptron`.

[30] N. Gupta *et al.*, "Artificial neural network," *Network and Complex Systems*, vol. 3, no. 1, pp. 24–28, 2013.

[31] `https://www.researchgate.net/figure/Structure-of-Perceptron_fig2_330742498`.

[32] G. Bebis and M. Georgiopoulos, "Feed-forward neural networks," *Ieee Potentials*, vol. 13, no. 4, pp. 27–31, 1994.

[33] "Architecture and learning process in neural network." `https://www.geeksforgeeks.org/ml-architecture-and-learning-process-in-neural-network/`.

[34] K. Hirasawa, M. Ohbayashi, M. Koga, and M. Harada, "Forward propagation universal learning network," in *Proceedings of international conference on neural networks (ICNN'96)*, vol. 1, pp. 353–358, IEEE, 1996.

[35] A. T. Goh, "Back-propagation neural networks for modeling complex systems," *Artificial intelligence in engineering*, vol. 9, no. 3, pp. 143–151, 1995.

[36] S. Sharma, S. Sharma, and A. Athaiya, "Activation functions in neural networks," *Towards Data Sci*, vol. 6, no. 12, pp. 310–316, 2017.

[37] P. Christoffersen and K. Jacobs, "The importance of the loss function in option valuation," *Journal of Financial Economics*, vol. 72, no. 2, pp. 291–318, 2004.

[38] M. Walter, S. Alizadeh, H. Jamalabadi, U. Lueken, U. Dannlowski, H. Walter, S. Olbrich, L. Colic, J. Kambeitz, N. Koutsouleris, *et al.*, "Translational machine learning for psychiatric neuroimaging," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 91, pp. 113–121, 2019.

[39] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for rf-based breast cancer detection," in *2017 Computing and Electromagnetics International Workshop (CEM)*, pp. 13–14, IEEE, 2017.

[40] W. Wolberg and O. Mangasarian, "Breast cancer wisconsin (original) dataset (1992)," *Accessed September*, 2019.

[41] N. Preethi and W. Jaisingh, "Analysis of fine needle aspiration images by using hybrid feature selection and various machine learning classifiers," in *Data Science and Security: Proceedings of IDSCS 2022*, pp. 383–392, Springer, 2022.

[42] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine learning classification techniques for breast cancer diagnosis," in *IOP Conference Series: Materials Science and Engineering*, vol. 495, p. 012033, IOP Publishing, 2019.

[43] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.

[44] Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," in *2018 International conference on electronics, control, optimization and computer science (ICECOCS)*, pp. 1–5, IEEE, 2018.

[45] K. Wadkar, P. Pathak, and N. Wagh, "Breast cancer detection using ann network and performance analysis with svm," *International journal of computer engineering and technology*, vol. 10, no. 3, pp. 75–86, 2019.

[46] T. Latchoumi and L. Parthiban, "Abnormality detection using weighed particle swarm optimization and smooth support vector machine," *Biomedical Research*, vol. 28, no. 11, pp. 4749–4751, 2017.

[47] V. Chaurasia and S. Pal, "Data mining techniques: to predict and resolve breast cancer survivability," *International Journal of Computer Science and Mobile Computing IJCSMC*, vol. 3, no. 1, pp. 10–22, 2014.

[48] A. H. Osman, "An enhanced breast cancer diagnosis scheme based on two-step-svm technique," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 4, 2017.

[49] H. William, W. N. Street, and O. L. Mangasarian, "Breast cancer wisconsin (diagnostic) data set," *UCI Machine Learning Repository*, 1995.

[50] D. Dua and C. Graff, "Breast cancer wisconsin (diagnostic) data set." `https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic`, 2023.

[51] A. Famili, W.-M. Shen, R. Weber, and E. Simoudis, "Data preprocessing and intelligent data analysis," *Intelligent data analysis*, vol. 1, no. 1, pp. 3–23, 1997.

[52] S. Manikandan, "Data transformation," *Journal of Pharmacology and Pharmacotherapeutics*, vol. 1, no. 2, p. 126, 2010.

[53] R. R. Picard and K. N. Berk, "Data splitting," *The American Statistician*, vol. 44, no. 2, pp. 140–147, 1990.

[54] J. Brownlee, "What is the difference between a batch and an epoch in a neural network," *Machine Learning Mastery*, vol. 20, 2018.

[55] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.

[56] "Kaggle." `https://www.kaggle.com/`.

[57] M. Lutz, *Programming python*. " O'Reilly Media, Inc.", 2001.

[58] Ž. Vujović *et al.*, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599–606, 2021.

[59] R. Susmaga, "Confusion matrix visualization," in *Intelligent Information Processing and Web Mining: Proceedings of the International IIS: IIPWM '04 Conference held in Zakopane, Poland, May 17–20, 2004*, pp. 107–116, Springer, 2004.

[60] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–12, 2019.

[61] R. Yacouby and D. Axman, "Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models," in *Proceedings of the first workshop on evaluation and comparison of NLP systems*, pp. 79–91, 2020.

[62] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.