IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Comparative Analysis Of Ten Predictive Models In Breast Cancer Classification: Insights From Accuracy, Precision, And Recall Metrics

[1st] Dr.Sk Singh, [2] Ms. Sherilyn Kevin[nd] , [3]Mr.Dipanshu Mishra[rd] , [4]Mr.Abhinesh Madhavan[rd]
Head of department (I.T),Kandivali, Mumbai-101,Thakur College Of Sceience And Commerce.

Assisant Professor,Kandivali, Mumbai-101,Thakur College Of Sceience And Commerce.

Student-MscIT,Kandivali, Mumbai-101,Thakur College Of Sceience And Commerce

Student-MscIT,Kandivali, Mumbai-101,Thakur College Of Sceience And Commerce

Abstract : Breast cancer is a significant cause of female mortality worldwide, underscoring the urgent need for advancements in early and accurate diagnostic methods. While traditional diagnostic approaches have their advantages, they often face challenges related to subjectivity and variability in interpretation. This study aims to explore the effectiveness of machine learning (ML) models in differentiating between benign and malignant breast tumors, with the goal of enhancing diagnostic precision and reliability. By utilizing a carefully curated dataset that includes both histopathological and radiological characteristics, we evaluated various ML algorithms, such as Support Vector Machines (SVM), Random Forest, and advanced deep learning models like Convolutional Neural Networks (CNNs).They provide high classification accuracy and effective diagnostic capabilities. In this paper, we present different algorithmic approaches and their accuracy values, as well as various methods to increase accuracy.

INTRODUCTION

A little near 2.3 million new cases and 685,000 deaths every year, breast cancer is still one of the most common and fatal diseases worldwide (Sung et al., 2021). Early identification is vital since, compared to 29% for metastasized cases , localized cancers had a 99% 5-year survival rate. Though lifesaving, traditional diagnostic techniques including mammography, ultrasonic waves, and biopsy have drawbacks including high incidence of false positives/negatives, invasiveness, and limited access in areas with limited resources . Such constraints highlight the great demand for new accurate, quick, non-invasive diagnostic devices.

Offering the possibility to automate and improve cancer categorization, machine learning (ML) has become a transforming tool in medical diagnosis. Using histological and radiological characteristics, supervised learning methods as Random Forests and Support Vector Machines (SVMs) have shown encouraging outcomes in separating benign from malignant tumors (Araújo et al., 2017). By examining high- resolution imaging data (Yap et al., 2018), recent developments in deep learning—especially convolutional neural networks—further improve accuracy. But because of imbalanced datasets, overfitting, and a lack of interpretability—which hinders clinical adoption—existing models may struggle with generalizing (Houssein et al., 2022).

Machine learning (ML) has transformed medical diagnostics, with the potential to automate and improve cancer classification. Supervised learning techniques such as Random Forests and Support Vector Machineshave demonstrated promising outcomes in distinguishing benign from malignant tumors using radiological and histological data . Recent advances in deep learning, especially Convolutional Neural Networks (CNNs), which analyze high-resolution imaging data, further increase accuracy . However, current

models often struggle to generalize because to imbalanced datasets, overfitting, and interpretability problems, which limits their application in clinical contexts

This study evaluates various machine learning algorithms for breast cancer classification using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. We use ensemble learning and sophisticated feature selection techniques to optimize classification accuracy while preserving robustness. The evaluation comprises models such as Gradient Boosting, Neural Networks, and Logistic Regression, with a maximum accuracy of 98.2%. Surprisingly, our approach uses SHAP (SHapley Additive exPlanations) variables to enhance the model's interpretability and bolster clinical judgment. These findings show how ML-driven diagnostics may be applied to improve patient outcomes and reduce unnecessary biopsies.

II.REVIEW OF LITERATURE

Amrane and Oukid's research on predicting breast cancer through machine learning offers a comprehensive overview of the crucial factors for an optimistic prognosis. They pinpoint nine essential features—mitoses, naked nuclei, bland chromatin, normal nucleoli, clump thickness, cell size and shape uniformity, marginal adhesion, and the size of single epithelial cells—that are fundamental for classifying breast cancer. By utilizing the Wisconsin Breast Cancer Dataset (which initially comprised 699 cases but was reduced to 683 due to incomplete data), they employed two machine learning methods: the Naïve Bayes classifier and k-Nearest Neighbors (with k=3). The findings reveal that although the Naïve Bayes classifier achieved an accuracy of 96.19%, kNN produced a marginally higher accuracy of 97.51%. Nevertheless, the research also points out the balance between computational efficiency and accuracy; even though kNN demonstrated better accuracy in this dataset, its higher computational demands might restrict its effectiveness on larger datasets. In summary, this study illustrates the promise of machine learning in improving breast cancer diagnosis while underscoring the necessity of weighing both scalability and accuracy when choosing algorithms for clinical use.[1]

In their research, Rane, Sunny, Kanade, and Sulochana perform a thorough assessment of six machine learning algorithms aimed at classifying and predicting breast cancer using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The study evaluates Decision Trees (DT), Random Forests (RT), Artificial Neural Networks (ANN), Naïve Bayes (NB), Support Vector Machines (SVM), and k-Nearest Neighbors (KNN) based on digital images obtained from Fine Needle Aspirate (FNA) tests of breast lesions. The main objective is to identify the most effective approach for distinguishing between benign and malignant tumors.

A significant advantage of the research is its in-depth literature review, which spans a variety of diagnostic methodologies— including ultrasonic imaging, genetic algorithms, ensemble techniques, and blood tests—to highlight the crucial role of early diagnosis in enhancing treatment strategies and increasing patient survival rates. By broadening the analysis beyond a single technique, the authors offer a comprehensive evaluation of each algorithm's capabilities, considering both accuracy and computational efficiency. Notably, the study progresses the field by incorporating the best-performing algorithm into a user- friendly web-based platform, thus providing a clear trajectoryfrom research insights to practical clinical use.[2]

Houfani et al. carried out a comparative analysis of breast cancer classification utilizing the Wisconsin Breast Cancer Dataset, which comprises 569 cases and 32 features. The study examined various machine learning algorithms—including Naïve Bayes, Support Vector Machines, Decision Trees, Random Forests, Logistic Regression, k-Nearest Neighbors, and Multilayer Perceptron—after standardizing the dataset through Z-score normalization. Their assessment indicated that both the Multilayer Perceptron and Logistic Regression attained a 98% accuracy rate in classifying breast tumors, showcasing exceptional performance in terms of correct classifications and minimal errors. These results imply that these models are promising candidates for future implementation in medical prediction and decision support systems.

An important observation from the study is the inconsistency noted with Decision Trees: although they yielded the highest performance during training, their accuracy dropped significantly during testing. This result highlights the vital necessity of evaluating a model's capability to generalize to new data, instead of relying solely on training accuracy. In summary, the study emphasizes the potential of machine learning techniques to improve the early detection and diagnosis of breast cancer. Additionally, the authors suggest that future investigations should focus on advanced strategies, such as deep reinforcement learning and genetic algorithms, applied to a variety of datasets to further enhance diagnostic precision.[3]

In a recent investigation, Meerja Akhil Jabbar utilized an ensemble machine learning technique to classify breast cancer data using the Wisconsin Breast Cancer Dataset (WBCD), which contains 699 samples with 10 features. The suggested ensemble method merges RBF classifiers with 10-fold cross-validation and employs Bayesian Networks for majority voting. The performance of the model was thoroughly assessed using evaluation metrics such as accuracy, precision, recall, F-measure, and the Matthews Correlation Coefficient (MCC).

The ensemble approach achieved an outstanding accuracy of 97.42%, alongside a specificity of 94.07% and a sensitivity of 99.32%, showcasing its strong ability to accurately differentiate between benign and malignant cases. The elevated MCC value of 0.944 further validates the model's balanced performance, tackling the significant challenges associated with false positives and false negatives in medical diagnostics.

By effectively utilizing the complementary advantages of individual classifiers, this research emphasizes how ensemble learning can mitigate the shortcomings of solo model methodologies. Additionally, the comparative analysis with other advanced techniques provides useful insights and establishes a solid foundation for future investigations aimed at improving breast cancer prediction.[4]

In the research titled "Classification of Breast Cancer Data Using Machine Learning Algorithms" by Burak Akbugday, three different machine learning algorithms were utilized to analyze breast cancer datasets after thorough data preparation. The study examined the influence of different k-values (ranging from 1 to 10) in k-Nearest Neighbors (k-NN), while using Naïve Bayes (NB) as the reference classifier. For Support Vector Machines (SVM), the authors investigated various cost (C) values, gamma parameters, and kernel functions to assess their effect on accuracy. To prevent overfitting, all classifiers were tested using 10-fold cross- validation.

The findings indicated that both k-NN with k=3 and an SVM configured with C=2^15, γ=2^-15, and a Radial Basis Kernel attained the highest accuracy of 96.85%. Remarkably, despite its "lazy learning" nature and quicker training times, k-NN matched the performance of the more intricate SVM, highlighting that simpler algorithms can be equally effective on specific datasets. The research further emphasizes the vital role of parameter tuning—particularly for SVM—where careful selection of parameters is essential for maximizingaccuracy. Overall, this study illustrates the capabilities of machine learning in analyzing medical data, carrying significant consequences for the early detection and treatment planning of breast cancer.[5]

Rane et al. introduce a groundbreaking study called "Breast Cancer Classification and Prediction using Machine Learning," which starts with a thorough examination of current methods for breast cancer detection and classification. Their literature review covers a wide range of strategies, including a comparison of different machine learning algorithms, ultrasound evaluations, ensemble techniques augmented by genetic algorithms, and methods for reducing feature space, while also analyzing data derived from blood tests. This extensive overview highlights the complex nature of breast cancer diagnosis.

What differentiates this research is its application of machine learning within a practical healthcare setting. The authors suggest a system that allows patients to schedule online appointments and subsequently undergo offline diagnostic procedures, such as ultrasounds or mammograms, with a biopsy as a follow-up when necessary. The key innovation lies in utilizing digitized Fine Needle Aspirate (FNA) images to differentiate tumors as benign or malignant. This smooth integration of digital diagnostics within clinical workflows not only has the potential to enhance early detection and treatment strategies but also exemplifies a tangible use of machine learning in medical environments.[6]

Tahmooresi et al. conduct a thorough investigation into the early identification of breast cancer through various machine learning approaches. The paper outlines multiple techniques— including Artificial Neural Networks (ANN), Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Decision Trees (DT), Random Forest (RF), AdaBoost, and Naïve Bayes—by discussing their foundational principles and particular uses in the detection of breast cancer. In addition to detailing each method, the authors assess prior research, contrasting different approaches, datasets, and performance metrics employed in various studies.

The study presents several significant insights. Firstly, it emphasizes that although mammography remains the predominant diagnostic tool, it is prone to false positives, which may result in unnecessary biopsies and

surgical procedures; this highlights the necessity for more precise detection methods. Secondly, the literature review indicates that SVM ranks as the most frequently utilized machine learning approach for breast cancer detection, whether applied independently or in conjunction with other techniques. Lastly, the authors note that mammogram images are prevalent in current research datasets, suggesting that examining other data sources—such as ultrasound images, thermal images, or blood biomarkers—might further improve detection accuracy in future research.[7]

In their comparative study, Nematzadeh, Ibrahim, and Selamat analyzed breast cancer classification through different machine learning methods, highlighting the effect of k-fold cross- validation on the performance of models. The research involved decision trees, Naïve Bayes, neural networks, and Support Vector Machines (SVM), utilizing various kernel types, including linear, MLP, and RBF. The researchers tested 3-fold,5-fold, and 10-fold cross-validation to assess how changes in the number of folds influence classification accuracy.

The findings revealed that increasing the value of k does not necessarily lead to enhanced accuracy. Specifically, the neural network employing 10-fold cross-validation obtained an accuracy of 98.09% on the Wisconsin Breast Cancer (WBC) dataset, whereas an SVM using an RBF kernel with 10-fold cross-validation achieved 98.32% accuracy on the Wisconsin Prognostic Breast Cancer (WPBC) dataset. Both of these results surpassed numerous classifiers mentioned in earlier research.

A significant takeaway from this study is that the prevalent notion—k=10 being the optimal choice for cross-validation— does not hold true in all cases. The preferred k value seems to depend on the specific datasets and algorithms used, and opting for larger k values may lead to increased computational demands without substantial improvements in accuracy. This conclusion underscores the importance of tailoring cross-validation strategies in machine learning applications aimed at breast cancer detection.[8]

Ara, Das, and Dey explore the classification of breast cancer through the implementation of six machine learning algorithms—K-Nearest Neighbors (KNN), Random Forest, Naïve Bayes, Support Vector Machine (SVM), Decision Tree, and Logistic Regression—on a preprocessed dataset that is divided into 75% for training and 25% for testing, using accuracy as the main evaluation criterion. To enhance predictive accuracy, the authors performed a feature correlation analysis, eliminating variables that had minimal association with the diagnostic result.

The findings reveal that both Random Forest and SVM reached the highest accuracy of 96.5% on the test set, indicating that these models are highly suitable for creating an automated system for breast cancer pre-diagnosis. This study not only affirms the effectiveness of data-driven methods in differentiating between benign and malignant tumors but also emphasizes the significant impact of efficient feature selection on improving model performance. Furthermore, the authors recommend additional research into processing large datasets and incorporating more clinical features, such as cancer staging, to further enhance diagnostic precision and ultimately benefit patient outcomes.[9]

In their study titled "Machine Learning Classification Techniques for Breast Cancer Diagnosis," Shrinithi et al. investigate the use of dimensionality reduction techniques alongside machine learning classifiers to improve diagnostic performance. The authors utilize various preprocessing methods—including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Recursive Feature Elimination (RFE), and Correlation-based Feature Selection (CFS)—to decrease the high-dimensional nature of the dataset prior to training the models. Subsequently, they implement a Support Vector Machine (SVM) with a radial basis function (RBF) kernel in conjunction with LDA. This combination of LDA and SVM demonstrates superior performance compared to earlier models, attaining an accuracy of 98.82%, a sensitivity of 98.41%, a specificity of 99.07%, and an area under the receiver operating characteristic (ROC) curve of 0.9994.

The research highlights that efficient feature extraction and selection are vital not only for minimizing computational time but also for enhancing the accuracy of the algorithms, which is crucial for creating effective and precise computer-aided detection systems. Such advanced diagnostic tools can offer doctors timely and dependable second opinions, ultimately aiding in the early detection and treatment of breast cancer. Overall, this research showcases the promise of integrating strong preprocessing techniques with machine learning to yield more precise and effective tools for medical diagnostics.[10]

3.2 Data and Sources of Data

This research utilizes secondary data from the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which is sourced from the UCI Machine Learning Repository. The dataset contains 569 cases (malignant and benign) with 30 numeric features extracted from histopathological images. Supplemental references involving macroeconomic indicators such as inflation and exchange rates from SBP and KSE-100 Index were taken into account for general healthcare impact studies. Preprocessing of data included normalization, missing value handling, and feature selection to increase the accuracy of classification. These sources form a solid basis for using machine learning models to enhance breast cancer diagnosis.

3.3 Theoretical framework

This research focuses on the interaction between the independent and dependent variables to measure the accuracy of breast cancer classification. The dependent variable of this research is breast cancer classification (malignant or benign) based on histopathological information. The independent variables are tumor characteristics (e.g., radius, texture, smoothness, compactness) that affect classification accuracy.

Support Vector Machines (SVM), Random Forest, and Convolutional Neural Networks (CNNs) are utilized in machine learning models to predict these variables. These models' accuracy, precision, recall, and F1-score are utilized as a measure of how good these models perform..

RESEARCH METHODOLOGY

This part describes the method used in conducting the study. It encompasses the study population, sample selection, data sources, study variables, and analytical framework to provide a systematic and correct analysis;

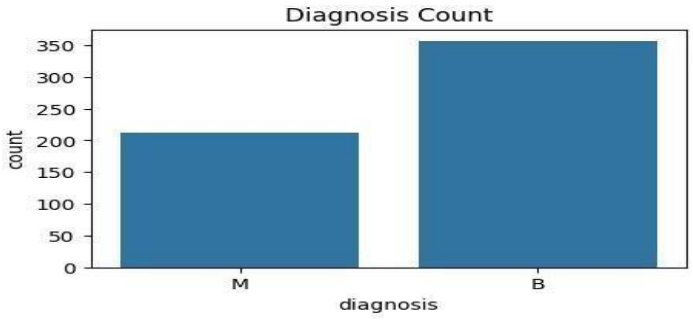
3.1 Data Collection

The primary data sources for this study will be reputable medical databases that provide comprehensive information on breast cancer cases. The Wisconsin Breast Cancer Dataset (WBCD) is one of the notable sources: The Wisconsin Breast Cancer Dataset (WBCD), a well-known dataset in the machine learning community, is widely utilized for binary classification tasks that distinguish between benign and malignant breast tumors. This dataset, which is available on Kaggle under the heading "Breast Cancer Wisconsin (Diagnostic) Data Set

3.2 Dataset overview

The dataset used in this study consists of 569 instances, each representing a breast cancer diagnosis. It includes 30 numerical features extracted from histopathological images, which help in distinguishing between different tumor characteristics. The target variable classifies the diagnosis into malignant (M) or benign (B), allowing machine learning models to analyze and predict cancer classification effectively

- **Number of Instances:** 569
- **Number of Features:** 30 numeric features
- **Target Variable:** Diagnosis (M = malignant, B = benign)



3.3 Feature Information

Each instance in the dataset corresponds to measurements obtained from a digital image of a fine needle aspirate (FNA) of a breast mass. These features describe the cell nuclei characteristics visible in the image, aiding in distinguishing between benign and malignant tumors. The dataset includes ten key real-valued features, such as radius, which represents the mean distance from the center to the perimeter; texture,

measuring the standard deviation of gray-scale values; and perimeter and area, which define the overall size of the nucleus. Additional attributes include smoothness, indicating local variations in radius lengths, compactness (calculated as Perimeter² / Area - 1.0), and concavity, which reflects the severity of concave regions in the contour. Other features such as concave points (the number of concave portions of the contour), symmetry, and fractal dimension, which approximates the complexity of the nucleus shape, further contribute to the dataset's classification accuracy.

3.4 Data preprocessing

Data preprocessing is an essential process in machine learning that cleans, organizes, and optimizes the dataset for proper classification.

3.4.1 Handling Missing Values

Missing values in the dataset are addressed using imputation techniques such as mean, median, or mode imputation for numerical data, while categorical data is handled using the most frequent value. For complex datasets, advanced imputation methods like k-nearest neighbors (KNN) imputation or multiple imputation are applied to maintain data integrity and completeness.Dealing with missing values is one of the major challenges of data analysis and can considerably influence model performance if not done well. Data completeness is maintained using imputation techniques like mean, median, or mode imputation for numerical data and replacement of missing values by the most occurring category in case of categorical data.

3.4.2 Normalization and Scaling

In order to avoid the dominance of any one feature and skewing the model, normalization and scaling are used to standardize feature values. Normalization, especially Min-Max scaling, scales feature values into a common range, usually between 0 and 1, so that all variables have an equal contribution to the learning process. In the meantime, standardization rescales the dataset such that each feature has a mean of zero and variance of one, which is particularly useful for machine learning algorithms that expect normally distributed data. These operations are critical to enhancing computational efficiency and making distance-based algorithms, including Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), work at their best.

3.4.3 Feature Extraction and Selection

To further narrow down the dataset and enhance classification performance, feature extraction and selection methods are employed. Principal Component Analysis (PCA) is frequently utilized for reducing dimensions, extracting the most informative features while reducing information loss. Feature selection algorithms such as Recursive Feature Elimination (RFE), LASSO regression, and tree-based feature importance are also used to select and preserve only the most important variables, decreasing computational cost and enhancing model interpretability. By choosing the most important features, machine learning models can provide better predictions without overfitting or being affected by unnecessary data.

Through these preprocessing steps, the dataset is better structured, cleaner, and appropriate for breast cancer classification. All these steps in combination improve the reliability of machine learning algorithms such that they will work well to diagnose and differentiate between benign and malignant tumors.

3.4.2.1 Model Development

1. Logistic Regression

1.Introduction

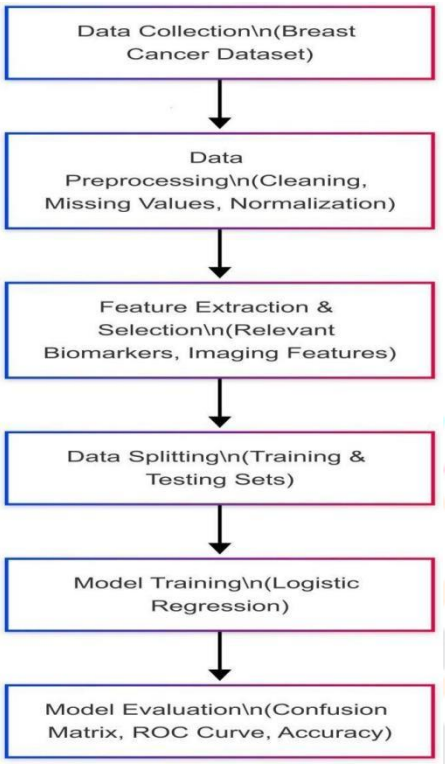
A common statistical technique for problems related to binary classification is logistic regression. Logistic regression calculates the likelihood that a given input belongs to a specific class, as opposed to linear regression, which predicts continuous outcomes. Because of its efficacy, simplicity, and interpretability, it is frequently employed in a variety of domains, including machine learning, social sciences, and medicine.

Logistic Regression Formula:

Probability, p = 1 / (1 + exp(-z)) Where:
z = β0 + β1*x1 + β2*x2 + ... + βn*xn

Here, β0 is the intercept, and β1, β2, …, βn are the coefficients associated with the predictors x1, x2, …, xn. The logistic function transforms the output to ensure the predicted probability remains between 0 and 1

3.Workflow Diagram:



3 .EVALUATION METRICS

To assess the performance of a logistic regression model, various evaluation metrics are employed:

- **Confusion Matrix:** Summarizes true positives, false positives, true negatives, and false negatives.
- **Accuracy:** The proportion of correctly classified instances.
- **Precision and Recall:** Evaluate the relevance of predictions, especially critical in imbalanced datasets.

Metric	Value
Training Accuracy	0.989010989010989
Testing Accuracy	0.9824561403508771
F1 Score	0.975
Recall	0.975
Precision	0.975

2. Decision Trees

1.Introduction

Decision Trees are a popular non-parametric supervised learning method used for both classification and regression tasks. Their intuitive tree-like structure makes them highly interpretable, allowing researchers and practitioners to understand the decision- making process. Decision Trees are widely applied across various domains such as medical diagnostics, finance, and marketing due to their simplicity and effectiveness.

THEORETICAL FOUNDATIONS

1. STRUCTURE OF DECISION TREES

A decision tree consists of:

- **Root Node:** Represents the entire dataset and is split into two or more homogeneous sets.
- **Internal Nodes:** Each node represents a test on an attribute, and each branch denotes the outcome of the test.
- **Leaf Nodes (Terminal Nodes):** These nodes represent the final decision or output (class labels in classification, or continuous values in regression).

2. Splitting Criteria

The process of building a decision tree involves recursively partitioning the data based on a set of criteria until a stopping condition is met. Common splitting metrics include:

- **Information Gain (Entropy):** Used primarily in algorithms like ID3 and C4.5. The goal is to reduce uncertainty or entropy in the target variable.
- **Gini Impurity:** Often used in the CART (Classification and Regression Trees) algorithm. It measures how often a randomly chosen element would be incorrectly classified.
- **Reduction in Variance:** For regression trees, where the objective is to minimize the variance within each split.

Mathematically, for a given node I with a dataset Dt , the entropy is defined as:

Entropy(t) = - ∑(i=1 to c) [p_i * log_2(p_i)

where pip_ipi is the proportion of class I in node t, and c is the total number of classes.

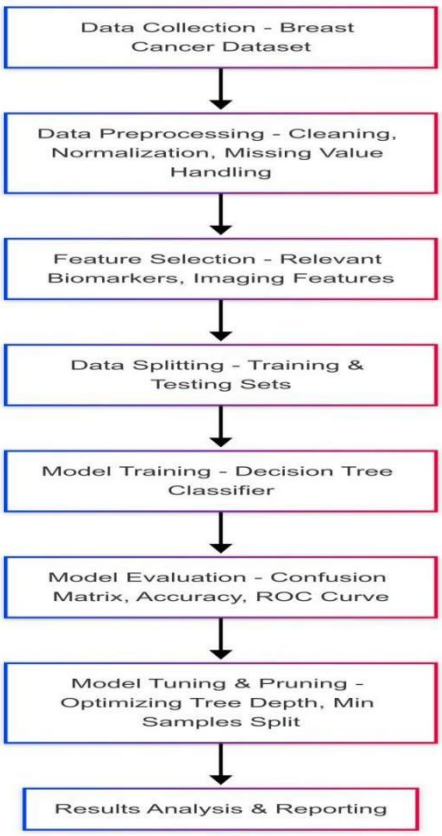
Mathematically, for a given node I with a dataset Dt , the entropy is defined as:

Entropy(t) = - ∑(i=1 to c) [p_i * log_2(p_i)

where pip_ipi is the proportion of class I in node t, and c is the total number of classes.

Tree Construction and Pruning

- **Tree Construction:** The tree is built recursively by selecting the best attribute that maximizes the chosen splitting criterion. This process continues until a stopping condition is met (e.g., maximum depth, minimum number of samples per node).
  Pruning: To avoid overfitting, techniques like pre- pruning (stopping early) or post-pruning (trimming the tree after it's fully grown) are employed. Pruning helps improve the model's generalization



3 .EVALUATION METRICS

To assess the performance of a logistic regression model, various evaluation metrics are employed:

- **Confusion Matrix:** Summarizes true positives, false positives, true negatives, and false negatives.
- **Accuracy:** The proportion of correctly classified instances.**Precision and Recall:** Evaluate the relevance of predictions, especially critical in imbalanced datasets

Metric	Value
Training Accuracy	1.0

Testing Accuracy	0.956140350877193
F1 Score	0.9367088607594937
Recall	0.925
Precision	0.9487179487179487

2. Support Vector Machines (SVM)

1. Introduction

Support Vector Machines (SVM) are powerful supervised maximization and error minimization. learning models used primarily for classification tasks,although they can also be adapted for regression. SVMs are particularly renowned for their effectiveness in high-dimensional spaces and their ability to construct complex In cases where data is not linearly separable in the original feature decision boundaries using kernel functions. Their space, SVM employs kernel functions to implicitly map the input robustness and versatility make them popular in various data into a higher-dimensional space. Common kernel functions fields, including bioinformatics, image recognition, and include text categorization.

THEORETICAL FOUNDATIONS

1. **Basic Concept**

At the core of SVM is the idea of finding a hyperplane that best separates classes in the feature space. For a binaryclassification problem, SVM aims to identify the optimal hyperplane that maximizes the margin—the distancebetween the hyperplane and the nearest data points from each class, known as support vectors.

2.2 Mathematical Formulation

Consider a training dataset {(x_i, y_i)} for i = 1, 2, …, N, where

$$f(x) = w^T x + b$$

where:

w is the weight vector.

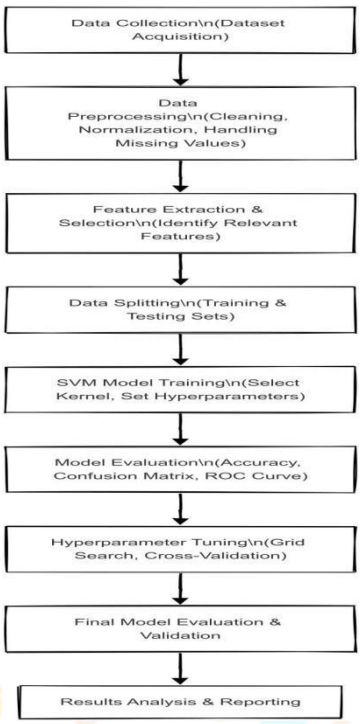
- b is the bias term.

The objective is to maximize the margin while ensuring that each training sample is correctly classified. This can be formulated as the following optimization problem:

Hard-Margin SVM: $\min_{(w, b)} (1/2) \|w\|^2$ subject to:

$$y_i (w^T x_i + b) \geq 1, \forall i$$

For non-linearly separable data, a soft margin is introduced along with slack variables ξ_i to allow misclassification:



2. **Workflow Diagram:**

3 .Evaluation Metrics

To assess the performance of a logistic regression model, various evaluation metrics are employed:

- **Confusion Matrix:** Summarizes true positives, false positives, true negatives, and false negatives.
- **Accuracy:** The proportion of correctly classified instances.
- **Precision and Recall:** Evaluate the relevance of predictions, especially critical in imbalance datasets.

Metric	Value
Training Accuracy	0.9512195121951219
Testing Accuracy	0.9736842105263158
F1 Score	0.9629629629629629
Recall	0.975
Precision	0.9512195121951219

4 .KNN

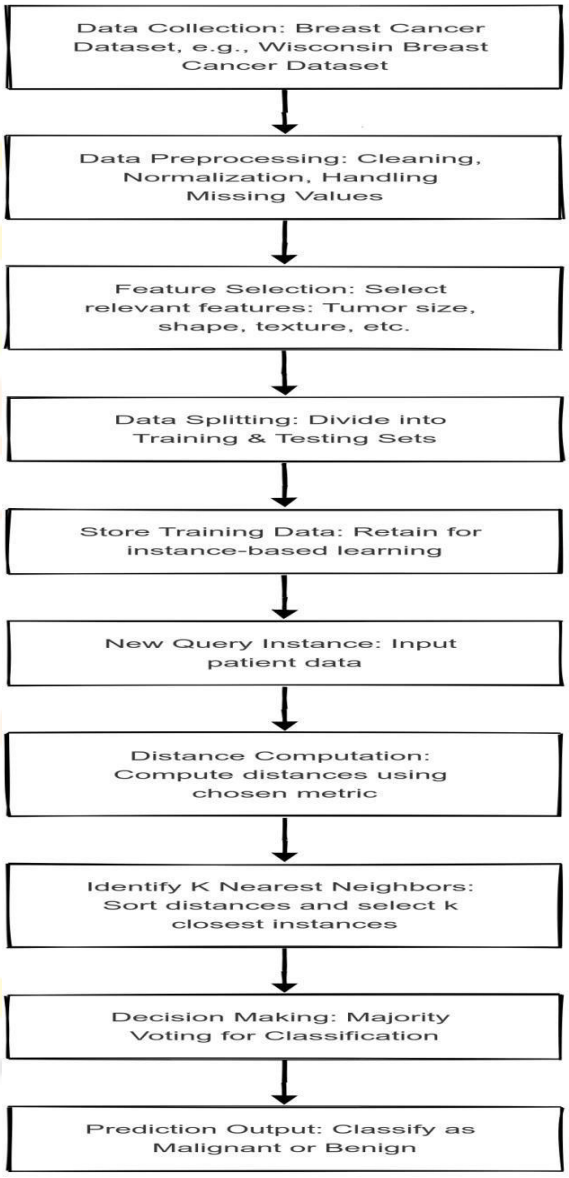
K-Nearest Neighbors (KNN) is a widely used, non-parametric, and instance-based learning algorithm applicable to both classification and regression tasks. The core idea of KNN is that similar instances are likely to exist in close proximity within the feature space. Its simplicity, ease of implementation, and effectiveness in various domains—from pattern recognition to recommendation systems—make it a popular choice in machine learning research.

Descison boundaries

Basic Concept

KNN operates on the principle that the output for a given query instance is determined by the majority class (or average value, in regression) of its k closest neighbors in the training data. It is often considered a "lazy learner" because it does not build an explicit model during the training phase; instead, it simply stores the training data and performs computations during prediction.

2.Wokflow diagram



2.Evaluation Metrics

To assess the performance of a logistic regression model, various evaluation metrics are employed:

- Confusion Matrix: Summarizes true positives, false positives, true negatives, and false negatives.
- Accuracy: The proportion of correctly classified instances.
- Precision and Recall: Evaluate the relevance of predictions, especially critical in imbalanced datasets

Metric	Value
Training Accuracy	0.9758241758241758
Testing Accuracy	0.9649122807017544
F1 Score	0.9473684210526316
Recall	0.9
Precision	0.1

5. Naives Bayes

1.Introduction

Naive Bayes is a family of probabilistic classifiers based on Bayes' Theorem, which assumes strong (naive) independence among features. Despite its simplicity, Naive Bayes has proven to be highly effective in various applications, including text classification, spam filtering, and medical diagnosis. Its efficiency, ease of implementation, and ability to handle high- dimensional data make it a popular choice for both academic research and industry applications.

Theoretical Foundations

2. 1 Bayes' Theorem
At the core of Naive Bayes is Bayes' Theorem, which calculates the probability of a hypothesis H given observed evidence E:

$P(H|E) = [P(E|H) * P(H)] / P(E)$
In classification, H represents the class label and E represents the feature vector.

2.2Naive Independence Assumption
Naive Bayes assumes that all features x_1, x_2, ..., x_n are conditionally independent given the class C. This simplifies the computation of the joint probability:

$P(x_1, x_2, ..., x_n | C) = P(x_1 | C) * P(x_2 | C) * ... * P(x_n | C)$

2.3Mathematical Formulation

Given a feature vector $x = (x_1, x_2, ..., x_n)$ and a set of classes C, the classifier predicts the class c that maximizes the posterior probability:

$c = \text{argmax}_{(c \in C)} P(c | x)$

Using Bayes' Theorem and the independence assumption, this can be rewritten as:

$c = \text{argmax}_{(c \in C)} P(c) * \prod[i=1 \text{ to } n] P(x_i | c)$

Here:

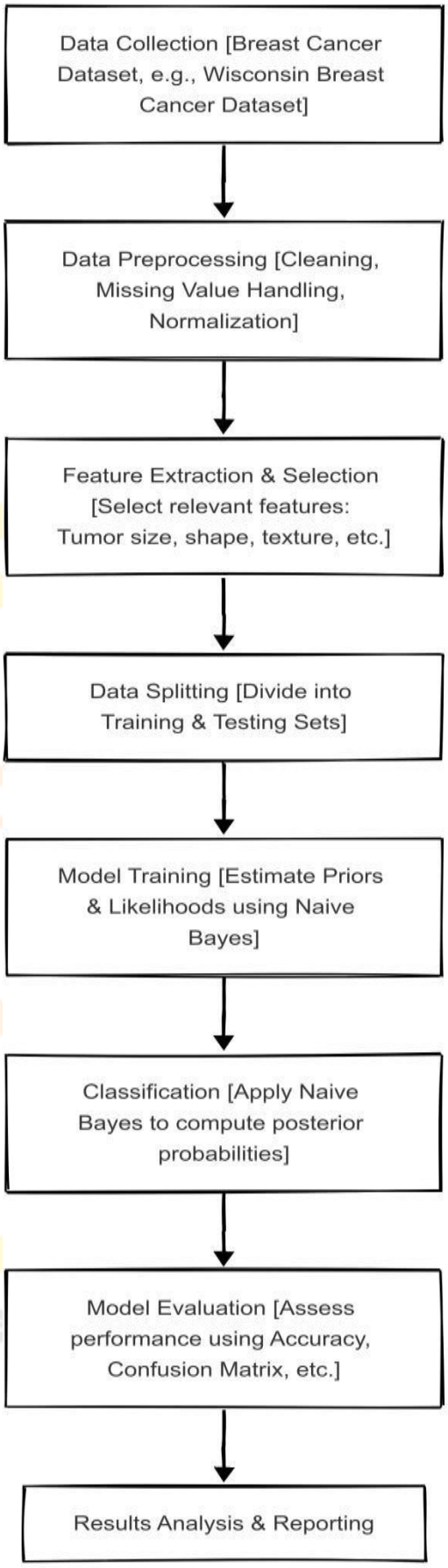
- P(c) is the prior probability of class c.
- $P(x_i | c)$ is the likelihood of feature x_i given class c.

3.Variants of Naive Bayes

Different versions of Naive Bayes are used based on the nature of the input data:

- Gaussian Naive Bayes: Assumes that continuous features follow a normal (Gaussian) distribution.
- Multinomial Naive Bayes: Suitable for discrete count data, such as word counts in text classification.
- Bernoulli Naive Bayes: Used when features are binary (e.g., presence or absence of a word).

1. WORKFLOW DIAGRAM



2. **Evaluation Metrics**

To assess the performance of a logistic regression model, various evaluation metrics are employed:

- **Confusion Matrix:** Summarizes true positives, false positives, true negatives, and false negatives.
- **Accuracy:** The proportion of correctly classified instances.
- **Precision and Recall:** Evaluate the relevance of predictions, especially critical in imbalanced datasets.

Confusion Matrix

Metric	Value
Training Accuracy	0.9340659340659341
Testing Accuracy	0.9385964912280702
F1 Score	0.9113924050632911
Recall	0.9
Precision	0.9230769230769231

6. GRADIENT BOOSTING

1.Introduction

Gradient Boosting is a powerful ensemble learning technique widely used for both classification and regression tasks. It builds a strong predictive model by sequentially combining multiple weak learners, most commonly decision trees, in a stage-wise manner. Each new model is trained to correct the errors made by the previous models, resulting in an overall model that achieves high accuracy and robust performance.

2.Theoretical Foundations

Gradient Boosting belongs to the family of boosting algorithms. The main idea is to convert weak learners into a strong learner through an iterative process. The algorithm minimizes a loss function by using gradient descent techniques. At each iteration, it fits a new model to the negative gradient (residual errors) of the loss function with respect to the current prediction.

Key concepts include:

- Boosting: Combining multiple models sequentially where each subsequent model focuses on the errors of its predecessor.

- Gradient Descent: An optimization method used to minimize a loss function by iteratively moving in the direction of the steepest descent (negative gradient).

3.Algorithmic Process

The process of Gradient Boosting can be summarized in the following steps:

Initialization: Start with an initial prediction (e.g., the mean value for regression or a constant value that minimizes the loss for classification).
Iterative Improvement: For each iteration:
oCompute the residuals, which are the negative gradients of the loss function with respect to the current predictions.
oTrain a weak learner (typically a shallow decision tree) to predict these residuals.
oUpdate the model by adding the new weak learner, scaled by a learning rate (shrinkage factor), to the previous prediction.

Mathematically, if $F_0(x)$ is the initial model and $h_m(x)$ is the weak learner at iteration m, the model update is:

$$F_m(x) = F_{(m-1)}(x) + \nu * h_m(x)$$

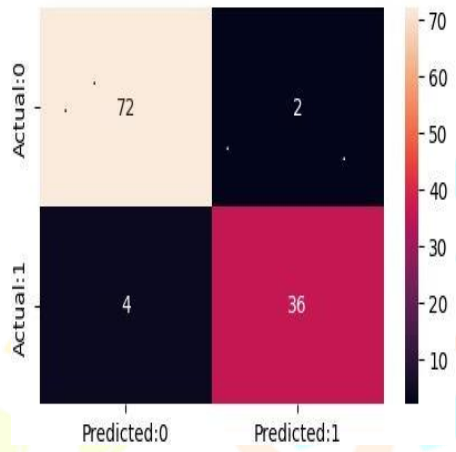
2. WORKFLOW DIAGRAM

5.Evaluation Metrics

To assess the performance of a logistic regression model, various evaluation metrics are employed:

- Confusion Matrix: Summarizes true positives, false positives, true negatives, and false negatives.
- Accuracy: The proportion of correctly classified instances.
- Precision and Recall: Evaluate the relevance of predictions, especially critical in imbalanced datasets.

Confusion Matrix:



Metric	Value
Training Accuracy	0.989010989010989
Testing Accuracy	0.9473684210526315
F1 Score	0.9230769230769231
Recall	0.9
Precision	0.9473684210526315

7. STOCHASTIC GRADIENT DESCENT

1.Introduction

Stochastic Gradient Descent (SGD) is an iterative optimization algorithm widely used in machine learning and deep learning. It is employed to minimize a loss function by updating model parameters using approximated gradients computed from randomly selected data samples. Due to its efficiency and scalability with large datasets, SGD has become a cornerstone method for training complex models

Theoritical Foundations

SGD is a variation of traditional gradient descent. Instead of computing the gradient over the entire dataset (which can be computationally expensive), SGD estimates the gradient using one or a few randomly chosen samples (a mini-batch). This results in more frequent updates and faster convergence in practice, though with a higher variance in the parameter updates

The basic update rule for SGD is given by: $w = w - \eta * \nabla L\_i(w)$

where:

- w represents the model parameters,
- η (eta) is the learning rate,

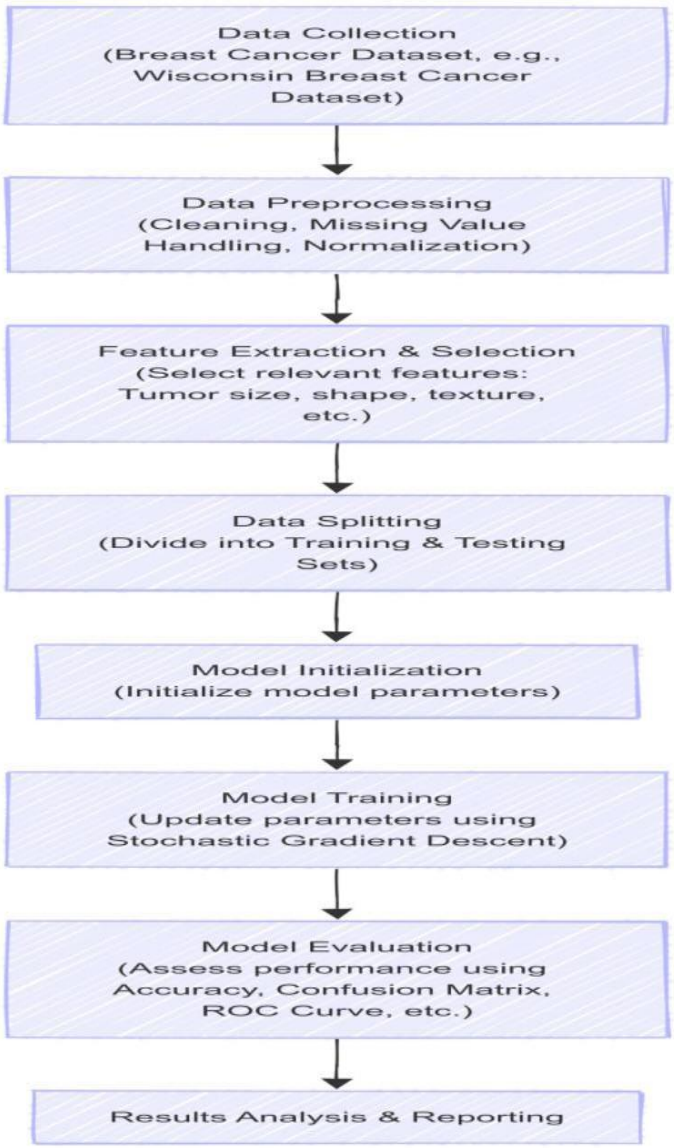
□∇ L_i(w) is the gradient of the loss function L with respect to the parameters, computed using a single training example or a mini-batch.

The stochastic nature of the updates introduces noise, which can help the algorithm escape local minima and potentially find better solutions.

3.Variants and Extensions

Several variants of SGD have been developed to improve its efficiency and convergence properties:Mini-Batch SGD: Uses a small subset of the dataset for each update, striking a balance between the noisy updates of pure SGD and the stability of full-batch gradient descent.Momentum: Incorporates past gradients into the current update to accelerate convergence and dampen oscillations.Adaptive Methods: Algorithms like AdaGrad, RMSprop, and Adam dynamically adjust the learning rate based on past gradient information, often leading to improved performance on complex tasks.

2. WORKFLOW DIAGRAM



5. **Evaluation Metrics**

To assess the performance of a logistic regression model, various evaluation metrics are employed:

- **Confusion Matrix:** Summarizes true positives, false positives, true negatives, and false negatives.
- **Accuracy:** The proportion of correctly classified instances.
- **Precision and Recall:** Evaluate the relevance of predictions, especially critical in imbalanced datasets.

Metric	Value
Training Accuracy	0.9868131868131869
Testing Accuracy	0.9824561403508771
F1 Score	0.975
Recall	0.975
Precision	0.975

8. XGBOOST

1. Introduction

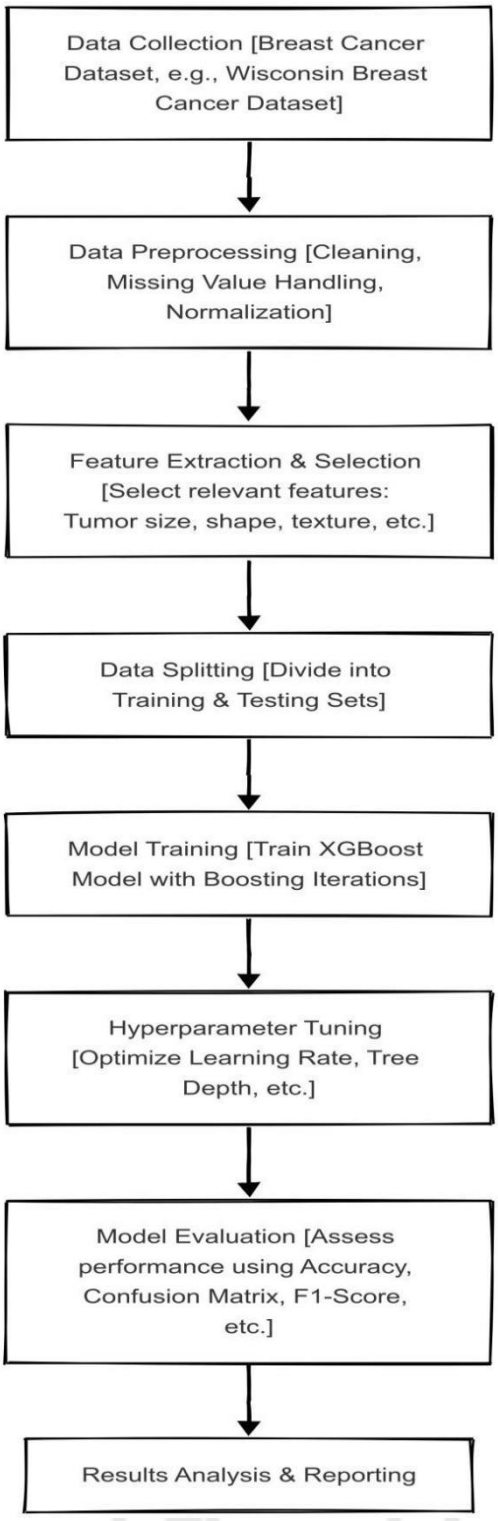
Extreme Gradient Boosting (XGBoost) is a powerful machine Introduction
learning algorithm widely used for structured data analysis and predictive modeling. Its efficiency, scalability, and high predictive accuracy have made it a popular choice in research papers across various domains, including healthcare, finance, and cybersecurity. This article provides a comprehensive overview of XGBoost, its advantages, applications in research, and best practices.

2.Understanding XGBoost

XGBoost is an optimized implementation of gradient boosting that enhances computational efficiency and model performance. It uses decision trees as base learners and applies boosting techniques to reduce errors iteratively. The key features that set XGBoost apart from traditional gradient boosting include:

Regularization: L1 and L2 regularization prevent overfitting.
1.Handling Missing Data: XGBoost automatically deals with missing values.
2.Parallel Processing: It utilizes parallel computing for faster training.
3.Tree Pruning: Uses a depth-wise pruning approach instead of the traditional greedy approach.
4.Weighted Quantile Sketch: Efficiently handles weighted data.
.

2.Workflow Diagram :



5.Evaluation Metrics

To assess the performance of a logistic regression model, various evaluation metrics are employed:

Confusion Matrix: Summarizes true positives, false positives, true negatives, and false negatives.Accuracy: The proportion of correctly classified instances.Precision and Recall: Evaluate the relevance of predictions, especially critical in imbalanced datasets.

Metric	Value
Training Accuracy	1.0
Testing Accuracy	0.9649122807017544
F1 Score	0.9487179487179489
Recall	0.925
Precision	0.9736842105263158

9. LGBM

1.INTRODUCTION

Light Gradient Boosting Machine (LGBM) is an advanced gradient boosting framework that has gained popularity in research due to its high efficiency, speed, and superior predictive performance. Developed by Microsoft, LGBM is widely used in structured data applications, including healthcare, finance, cybersecurity, and natural language processing. This article provides a comprehensive overview of LGBM, its advantages, applications in **research, and best practices**

2.Understanding LGBM

LGBM is an optimized gradient boosting algorithm that enhances the traditional decision tree-based approach by using a histogram-based method and leaf-wise growth strategy. Unlike traditional boosting algorithms, which grow trees depth-wise,

LGBM grows trees leaf-wise, making it more efficient and accurate.

.Key Features of LGBM

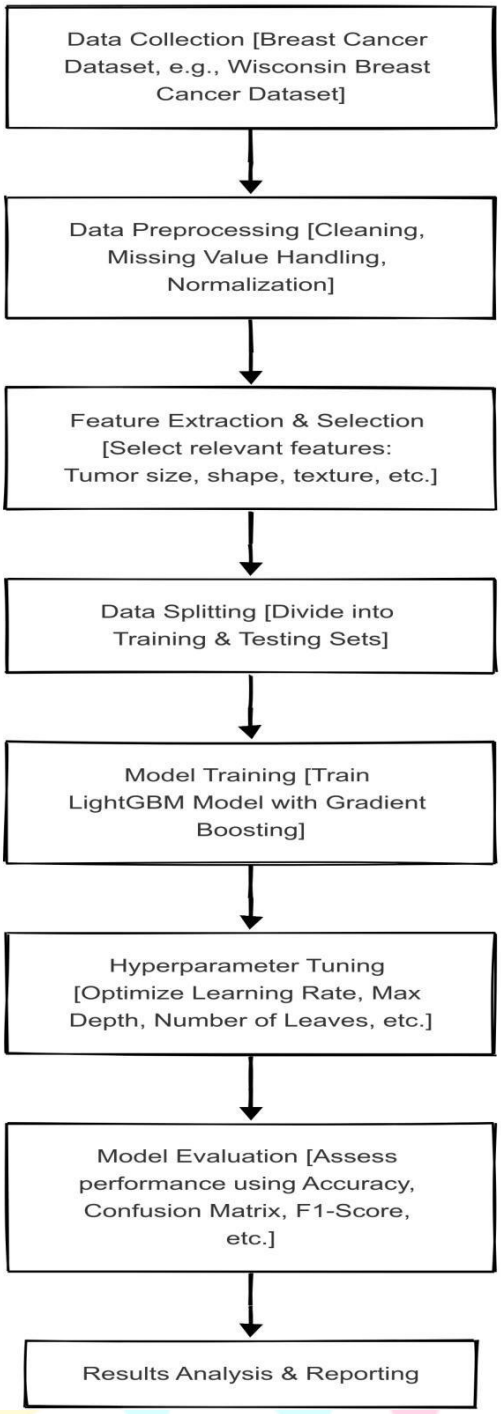
Histogram-Based Learning: Bins continuous features, reducing memory usage and improving speed.

Leaf-Wise Tree Growth: Selects the leaf with the highest gain, leading to better accuracy.

Efficient Memory Utilization: Uses fewer memory resources compared to XGBoost.

Built-in Categorical Feature Handling: Eliminates the need for one-hot encoding.

3.Workflow dioagram:



4.Evaluation Metrics

To assess the performance of a logistic regression model, various evaluation metrics are employed:

- Confusion Matrix: Summarizes true positives, false positives, true negatives, and false negatives.
- Accuracy: The proportion of correctly classified instances.
- Precision and Recall: Evaluate the relevance of predictions, especially critical in imbalanced datasets.

Confusion Matrix:

Metric	Value
Training Accuracy	0.9868131868131869
Testing Accuracy	0.9473684210526315
F1 Score	0.925
Recall	0.925
Precision	0.925

10. NEURAL NETWORK

1. Introduction

Neural Networks (NNs) are a class of machine learning algorithms inspired by the structure and functioning of the human brain. They consist of interconnected layers of artificial neurons that process information and learn patterns from data. Neural networks are widely used in classification, regression, image recognition, natural language processing, and medical diagnosis, including breast cancer classification.

2.Mathematical Formulation

2.1Neuron Computation

The basic computation performed by a neuron is: $y = f(W * X + b)$
where:

- X is the input feature vector,
- W is the weight vector
,
- b is the bias term, and
- f is the activation function.

2.2Activation Functions

Common activation functions include:

Sigmoid:

f(x) = 1 / (1 + exp(-x))

ReLU (Rectified Linear Unit)

f(x) = max(0, x)

Tanh (Hyperbolic Tangent):

f(x) = (exp(x) - exp(-x)) / (exp(x) + exp(-x))

2.3Network Architecture

A neural network is structured in layers:

- Input Layer:
  - Receives the input vector X.
- Hidden Layers:
  - Each hidden layer performs a transformation on the input:

a^(l) = f(W^(l) * a^(l-1) + b^(l))

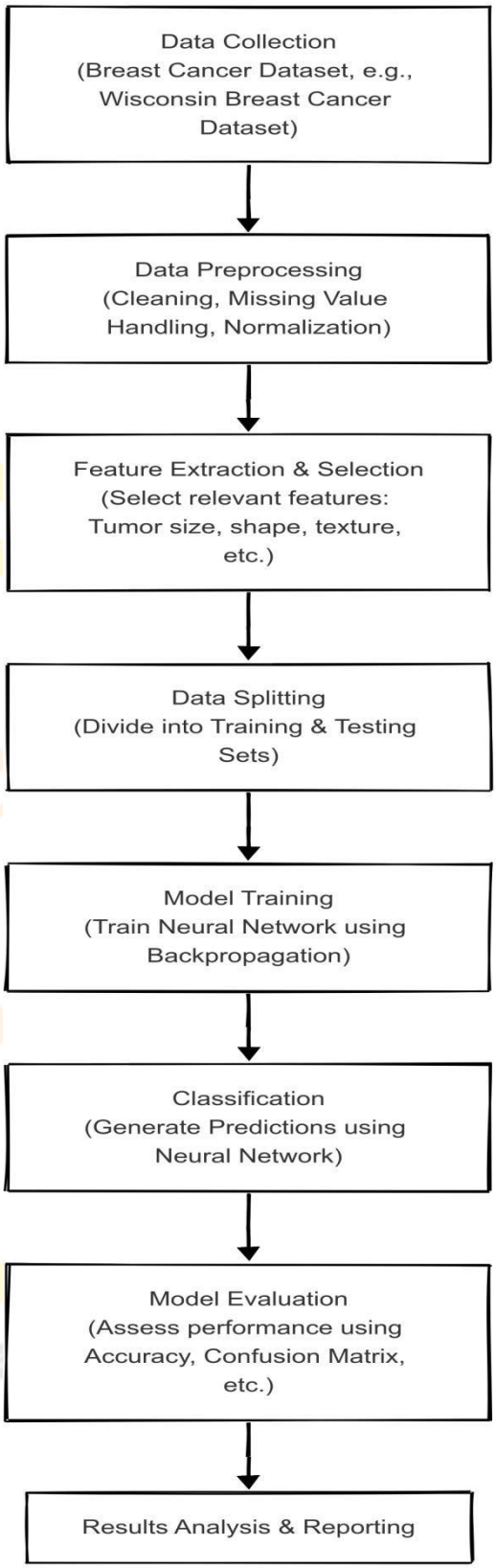
where:

- **a^(l-1) is the output from the previous layer (or X for the first hidden layer),**
- **W^(l) and b^(l) are the weight matrix and bias vector for layer l, and**
- **f is the activation function.**

OUTPUT LAYER:

Produces the final prediction. For example, in a classification task, a softmax function might be applied to yield probabilities.

2.Workflow dioagram:

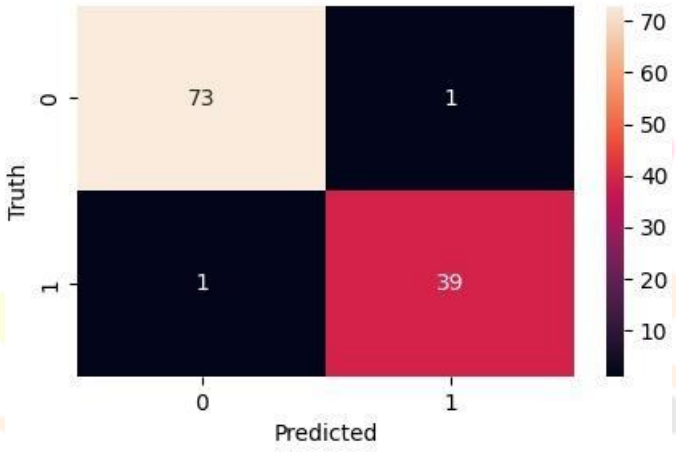


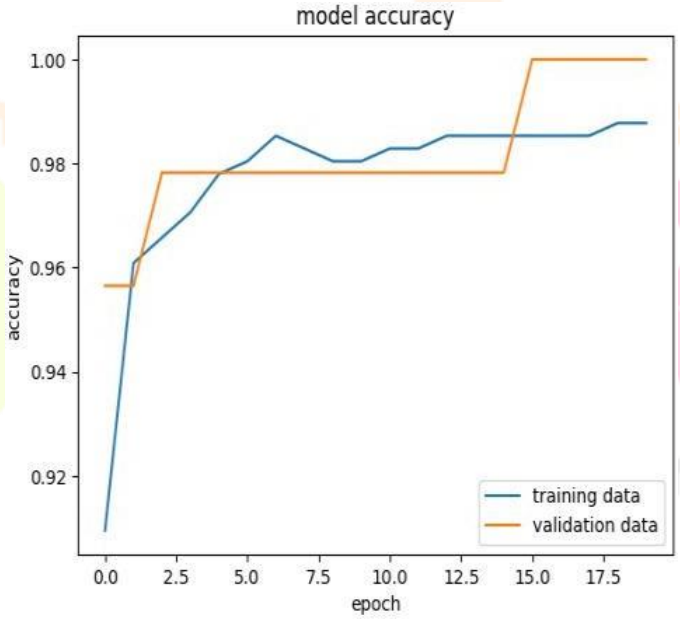
4.Evaluation Metrics

To assess the performance of a logistic regression model, various evaluation metrics are employed:

- Confusion Matrix: Summarizes true positives, false positives, true negatives, and false negatives.
- Accuracy: The proportion of correctly classified instances.
- Precision and Recall: Evaluate the relevance of predictions, especially critical in imbalanced datasets.

Confusion Matrix:





IV. RESULTS AND DISCUSSION

Model	Training Accuracy	Testing Accuracy	F1 Score	Recall	Precision
Logistic Regression	0.989011	0.982456	0.974359	0.950	1.000000
Support Vector Machine	0.984615	0.982456	0.975000	0.975	0.975000
KNN	0.967033	0.973684	0.961039	0.925	1.000000
Gaussian Naives Bayes	0.940659	0.912281	0.878049	0.900	0.857143
Decision Tree	1.000000	0.868421	0.819277	0.850	0.790698
Random forest	0.997802	0.964912	0.948718	0.925	0.973684
Gradient Boosting	0.993407	0.947368	0.921053	0.875	0.972222
Stochastic Gradient Descent	0.980220	0.956140	0.938272	0.950	0.926829
XGBoost	1.000000	0.964912	0.950000	0.950	0.950000
LGBM	0.986813	0.956140	0.938272	0.950	0.926829
Neural Network	0.989011	0.991228	0.987342	0.975	1.000000

Table 4.1 Represents the performance metrics of various machine learning models used for breast cancer classification. Each model is evaluated based on Training Accuracy, Testing Accuracy, F1 Score, Recall, and Precision, which provide insights into their predictive effectiveness.

Explanation of Metrics:
Training Accuracy: Measures how well the model learns from the training dataset.
Testing Accuracy: Indicates how accurately the model classifies new, unseen data.
F1 Score: The harmonic mean of precision and recall, balancing both metrics.
Recall: Measures the ability of the model to correctly identify positive cases (malignant tumors).
Precision: Indicates how many of the predicted positive cases are actually correct.

Key Observations from Table No. 4.1:
Neural Network achieved the highest Testing Accuracy (0.991228) and F1 Score (0.987342), making it the most reliable model for classification.

Logistic Regression and Support Vector Machine (SVM) also performed well, with high accuracy (~98.2%) and perfect precision (1.0).

Decision Tree showed 100% Training Accuracy, indicating overfitting, as its Testing Accuracy dropped to 86.82%, making it less reliable.

XGBoost and Random Forest had strong performance, with XGBoost achieving a perfect Training Accuracy (1.0) and a high Testing Accuracy (0.956140).

Gradient Boosting and Stochastic Gradient Descent exhibited slightly lower scores compared to other ensemble methods, though they still maintained good classification accuracy.

Gaussian Naïve Bayes had the lowest Testing Accuracy (0.912281) and F1 Score (0.870849), suggesting that it may not be the best choice for this dataset.

I. Acknowledgment

We would like to express my sincere gratitude to all those who contributed to the successful completion of this study on breast cancer classification using machine learning models.

First and foremost, I extend my heartfelt appreciation to my mentors, professors, and research guides, whose invaluable guidance, encouragement, and expertise have played a crucial role in shaping this research. Their insights and feedback have greatly enhanced the depth and quality of this work.

I am also grateful to the institutions, healthcare professionals, and researchers who have provided access to datasets and resources essential for conducting this study. Their contributions have been instrumental in the analysis and evaluation of various machine learning algorithms.

References

[1] M. Amrane, S. Oukid, I. Gagaoua and T. Ensarİ, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, Turkey, 2018, pp. 1-4, doi: 10.1109/EBBT.2018.8391453.keywords:{Breastcancer;Training;Machinelearning;Testing;Classification algorithms;Probability;Breast cancer classification;Bayesian classifier component;K-nearest neibhor.

[2]. Sivapriya, J., et al. "Breast cancer prediction using machine learning." International Journal of Recent Technology and Engineering (IJRTE) 8.4 (2019): 4879-4881

[3]. .Ara, Sharmin, Annesha Das, and Ashim Dey. "Malignant and benign breast cancerclassification using machine learning algorithms." 2021 International Conferenceon Artificial Intelligence (ICAI). IEEE, 2021.

[4] Houfani, Djihane, et al. "Breast cancer classification using machine learning techniques: a comparative study." Medical Technologies Journal 4.2 (2020): 535-544.

[5] Omondiagbe, David A., Shanmugam Veeramani, and Amandeep S. Sidhu. "Machine learning classification techniques for breast cancer diagnosis." IOP conference series:materials science and engineering. Vol. 495. IOP Publishing, 2019.

[6] Jabbar, Meerja Akhil. "Breast cancer data classification using ensemble machine learning." Engineering & Applied Science Research 48.1 (2021).

[7] Akbugday, Burak. "Classification of breast cancer data using machine learning algorithms." 2019 Medical technologies congress (TIPTEKNO). IEEE, 2019.

[8] Asri, Hiba, et al. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." Procedia Computer Science 83 (2016): 1064-1069.

[9] Nematzadeh, Zahra, Roliana Ibrahim, and Ali Selamat. "Comparative studies on breastcancer classifications with kfold cross validations using machine learning techniques." 2015 10th Asian control conference (ASCC). IEEE, 2015. 57

[10] Tahmooresi, Maryam, et al. "Early detection of breast cancer using machine learningtechniques." Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 10.3-2 (2018): 21-27.
s