I. J. Engineering and Manufacturing, 2026, 4, 228-252
Published Online on August 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2026.04.16

An Interpretable Machine Learning Framework for Breast Cancer Diagnosis Using Statistical Feature Analysis and Ensemble Classification

T. Haripriya
Department of Mathematics, Sreyas Institute of Engineering and Technology, Hyderabad, 500068, Telangana, India
E-mail: haripriya.tirumala@gmail.com
ORCID iD: https://orcid.org/0009-0007-3938-2703

M.V. Ramana Murthy
Retd. Professor of Mathematics, Osmania University, Hyderabad, 500007, Telangana, India
E-mail: mv.rm50@gmail.com
ORCID iD: https://orcid.org/0000-0003-3371-4125

Ch. Vasavi
Department of Mathematics, Sreyas Institute of Engineering and Technology, Hyderabad, 500068, Telangana, India
E-mail: vasavirao.c@sreyas.ac.in
ORCID iD: https://orcid.org/0000-0002-6619-7462

Swathi Gowroju
Department of CSE (AI & ML), Sreyas Institute of Engineering and Technology, Hyderabad, 500068, Telangana, India
E-mail: swathigowroju@sreyas.ac.in
ORCID iD: https://orcid.org/0000-0002-4940-1062

G. Srinivas
Department of Mathematics, Geethanjali College of Engineering and Technology, Hyderabad, 501301, Telangana, India
E-mail: gn.nivas@gmail.com
ORCID iD: https://orcid.org/0000-0002-1843-7337

Devineni Gireesh Kumar*
Department of CSE (AI & ML), Sreyas Institute of Engineering and Technology, Hyderabad, 500068, Telangana, India
E-mail: gireesh218@gmail.com
ORCID iD: https://orcid.org/0000-0003-0575-471X
*Corresponding Author

Received: April 29, 2026; Revised: June 8, 2026; Accepted: July 6, 2026; Published: August 8, 2026

Abstract: A clear, statistically sound, yet easily understandable breast cancer diagnosis is a difficult issue in all healthcare systems, because early stages of breast cancer are critical in therapy success and long-term survivability. This machine-learning-based breast cancer classifier, in a statistically justified, rigorously experimentally validated way, classifies a set of 569 breast cancer cases with 9 cytological features for breast cancer diagnosis. The classifier uses a rigorous set of data cleanup measures, including missing-value substitution, correlation-based feature reduction, and projection into principal component space, to achieve high data quality, reduce redundancy, and enhance feature usefulness. Five supervised classifiers, in a widely accepted train-test model using an 80:20 random sample split and 5-fold cross-validation, are fitted and evaluated using Accuracy, Precision, Recall, F1-score, and Area under the ROC curve. In these tests, the Random Forest classifier got the best result, with 95.84% Accuracy, 95.31% Precision, 95.12% Recall, 95.21% F1-score and 0.982 area under the ROC curve; in a statistically sound consistency test using cross-validation, its mean accuracy reached 95.96% with a small standard deviation of 0.43. To provide a clear, interpretable indication of which features truly matter, we performed a feature-importance analysis on the best classifier, the Random Forest model. Results show that the expression levels of Bland Chromatin, Single Epithelial Cell Size, Normal Nucleoli, Uniformity of Cell Shape, Uniformity of Cell Size and Bare Nuclei are closely related to breast cancer diagnosis; this is almost the same as the clinical diagnosis findings, and very naturally suggests that abnormalities of cellular morphology and nuclei are major symptoms of breast cancer. In comparison, prior research may neglect validation and efficiency comparisons or focus

only on the classifier's accuracy. Our method combines multiple levels of assessment (statistical data-by-data validation, feature importance, cross-validation, and comparison of different classifiers using ensemble learning) into a single evaluation system. This combined approach not only enhances predictive capability but also makes the entire setup more explicitly interpretable from a clinical perspective, thereby making it more suitable for health care decision support. Given the strong classification performance, interpretability, and validation suggested above, the model would help physicians detect breast cancer very early, reducing the risk of misdiagnosis.

Index Terms: Interdisciplinary modelling, Cancer classification, Machine learning, Ensemble methods, Feature importance, Precision oncology

1. Introduction

The worldwide incidence of breast cancer, the second most common cancer in both incidence and mortality, is increasing, and it remains one of the leading causes of morbidity and mortality of cancer amongst women. Developing early, accurate diagnostic techniques and tools could improve treatment efficacy, reduce mortality rates, and optimise personalised therapy. Cancer diagnosis, though, remains a relatively difficult task, as tumour development occurs at many different levels (mutational changes, epigenetic defects, cellular anomalies, environmental factors [1-3]). Recent breakthroughs in biomedical data collection and computational intelligence have opened exciting new possibilities for data-driven cancer diagnosis and management. The abundance of clinical imaging, genomic, and pathological data can now be leveraged to learn complex disease dynamics and identify emergent patterns of tumour growth and malignancy using machine learning methods [4-8]. Machine learning has become the paradigmatic tool for cancer classification, prognosis prediction, treatment response analysis and survival prediction by harnessing automated data mining. Various machine learning algorithms have shown successful results in breast cancer diagnosis, including Logistic Regression, SVM, Decision Trees, KNN, Random Forests, and other ensemble learning algorithms [9,10]. In more recent studies, sophisticated methods such as gradient boosting, multimodal learning, XAI (explainable artificial intelligence), and deep learning have shown encouraging diagnostic results across many cancer applications [11-13]. Even with these novel developments in the field, several key issues remain inadequately addressed. Numerous published studies tend to emphasise endeavours to optimise prediction algorithms rather than delve into the details of feature interpretability, biological significance, and model transparency [14].

In a clinical setting, we can do without black boxes whose lack of transparency prevents clinicians from understanding the basis of algorithmic predictions, making them hesitant to adopt them in everyday decision-making. To make matters worse, most of the published work has used black-box models, which quickly indicate where the predictive features are located without providing much insight into how the disease process is reflected in them [15]. Another obstacle is that most of the literature does not report the use of exploratory statistical analysis in the machine learning model-building process. Exploratory statistical analysis, including correlation analysis, assessment of feature relationships, and dimensionality reduction, has often been overlooked, even though it plays an integral role in identifying strong, relevant features, reducing dataset noise, and achieving a globally optimal solution for the model [16]. Statistical methods such as principal component analysis (PCA) have proven to be the best for dimensionality reduction and for acquiring an optimal set of features. It is easy to perform PCA without statistical analysis for feature reduction. It is hard to find breast cancer classification studies that use statistical feature analysis, PCA, and comparisons of multiple machine learning algorithms for validation within a unified system [17]. Also, diagnostic systems intended for healthcare use should primarily prioritise high sensitivity and low false-negative rates, since missing malignant cases may lead to execution delays and poorer clinical outcomes. Since then, machine learning algorithms should be reliable enough not only to deliver high accuracy but also to demonstrate robustness, reproducibility, and clinical interpretability [18]. Such characteristics may be further supported by ensemble learning methods such as Random Forest, which can account for nonlinearity, minimise variance, and provide relatively robust explanations of feature importance [19, 20].

1.1. Research Gap

Despite many papers having dealt with various machine learning methods for diagnosing cancer, there are still several gaps in the research:

1. Very little effort has been made to combine statistical exploratory analysis with predictive modelling.
2. Feature interpretability and biological relevance have not been given enough priority.
3. A thorough comparative study of different machine learning models with the use of standardised evaluation metrics is missing.
4. Only a very small number of researchers have applied dimensionality reduction methods that provide the dual benefit of making models more robust and less redundant.

5. There is not enough attention directed at the creation of computationally efficient and reproducible setups that could be used for clinical decision-support systems.

Developing dependable and explainable machine learning methods for early diagnosis of breast cancer heavily relies on those issues being addressed.

1.2. Objectives and Contributions

This research aims to establish and test an interpretable machine learning system for breast cancer diagnosis by combining exploratory statistical analysis, Principal Component Analysis (PCA), and supervised learning methods to enhance diagnostic reliability, model transparency, and prediction accuracy.

This work mainly aims at:

- Creating a well-organised initial data treatment method that, inter alia, handles missing data, performs correlation analysis, and uses Principal Component Analysis (PCA).
- Estimating the effectiveness of different supervised learning models like Logistic Regression, K-Nearest Neighbours, Decision Tree, Naive Bayes, and Random Forest.
- Measuring performance based on various criteria, including accuracy, precision, recall, F1-score, and ROC-AUC, within the same validation scheme.
- Finding and explaining features that are biologically significant for breast cancer malignancy.
- Studying ensemble learning methods as a means for enhancing diagnostic accuracy as well as diminishing false-negative results in cancer screening procedures.

1.3. Research Significance

This work is valuable because it aims to develop an interpretable, clinically and statistically validated machine learning tool for breast cancer diagnosis. This subject has not been the focus of many papers, most of which focus only on prediction precision. Yet, the experimental setup proposed in this work offers a commendable trade-off between diagnostic efficiency and feature interpretability. The authors use feature analysis, Principal Components Analysis, and aggregation learning to develop a method that is both faithfully reproducible and computationally efficient. Apart from confident diagnosis, it produces a meaningful interpretation of predictive features in a short process, and an interpretable artificial intelligence system for early cancer detection and patient risk factor stratification, useful for clinical decisions.

1.4. Paper Organisation

The paper is organised. Related work on cancer diagnosis using machine learning, explainable AI, and ensemble learning methods is covered in Section 2. Section 3 presents the approach, i.e., the steps for choosing and preprocessing the dataset, performing dimensionality reduction, and building the model. Section 4 reports experiments and comparative performance evaluation. Section 5 interprets the results and discusses the clinical relevance of the study. Finally, Section 6 summarises the paper and outlines potential directions for future work.

2. Literature Review

2.1. Machine Learning in Cancer Diagnosis

ML has become an excellent assistant in cancer diagnosis and prognosis, as it can analyse complex biomedical data and uncover hidden patterns associated with cancer outcomes. The growing volume of clinical and imaging data, as well as molecular profiles, has led to the rise of supervised learning algorithms used in clinical practice. Deep learning techniques can accurately and effectively predict medical diagnoses. For instance, a deep convolutional neural network was implemented in [1] for brain tumour classification and demonstrated high diagnostic performance on MRI data. The method achieved high prediction accuracy, but the classification decision-making process is not clear or understandable to physicians. Similarly, in [2], machine learning was applied to tumour classification using gene expression data, yielding good predictive performance. The large number of features indicated the need for dimensionality reduction techniques. One promising approach to resolving transparency issues is explainable machine learning. An Explainable Artificial Intelligence (XAI) setup in healthcare focuses on factors such as interpretability, accountability, and clinician trust, as detailed in [3]. It was entirely theoretical and did not conduct experimental analyses. Multiple machine learning algorithms were explored for breast cancer diagnosis, and ensemble classifiers were found to be more accurate than individual classifiers [4]. This study provided little commentary on biological interpretation and on the relationships among features.

The application of machine learning algorithms in cancer diagnostics and therapeutics was examined [5], but not in depth, including correlations between features and exploratory statistical analyses. The Support Vector Machine (SVM), RF, and KNN algorithms for predicting metastatic breast cancer were explored [6]. This study emphasised the usefulness of cross-validation, while dimensionality reduction or relationships/co-redundancy among features were not considered.

Together, these studies demonstrate the ability of machine learning to diagnose cancer accurately. Yet, they all share a similar limitation in their results: a predictive tool is preferred over most statistical analyses and meaningful interpretation of the data.

2.2. Systems Biology and Computational Modelling

Cancer is a multifaceted condition characterised by the involvement of multiple biological processes and molecular pathways. Systems biology enables the study of these complex interactions through computational modelling and simulation [7]. This work highlighted the contributions of systems and synthetic biology to a better understanding of tumour growth through integrative modelling. The article gave a theoretical basis for the biological intricacies. Still, it did not address the aspect of building predictive classification models.

Similarly, [8] discussed the use of computers to analyse biological communication networks and found that interdisciplinary collaboration is most important. Even though these approaches help in understanding biological systems, their use in creating interpretable clinical classifications remains limited. A multimodal model that integrates pathological and genomic data for cancer diagnosis and prediction was developed [9]. Although the study showed improved predictive accuracy through data integration, the model primarily relies on complex deep learning architectures, which can limit both implementation and explainability. These papers show one way to discover biological laws using computer science, but that is not enough. Making such diagnostic structures computationally clear and efficient so they can be easily used in hospitals is what is really required.

2.3. Explainable Artificial Intelligence in Healthcare

Interpretability is rapidly becoming a major requisite for the introduction of machine learning in healthcare. Physicians need to understand and be able to justify the reasons machine learning models make certain predictions before they can be used in hospitals. The role of explainable artificial intelligence was examined [10] in the healthcare sector, and it was noted that one key element in fostering trust, transparency, and ethical adherence is the explainability feature in AI systems. Explainable machine learning models were applied at the clinical level by demonstrating their ability to detect breast cancer patients with lymph node metastasis [11]. This means that working with interpretable models can be a very important clinical tool without sacrificing predictive accuracy. Machine learning can be used in pharmacometrics and highlights the increased regulatory demand for predictive systems to be transparent, reproducible, and well-validated [12]. These articles stand behind the ability of machine learning models to explain their decisions. Many of these models still function as black boxes and do not disclose relationships among features or their biological significance. So, the first step in subsequent research will be to study feature importance, analyse correlations, and interpret the results using statistics.

2.4. Ensemble Learning and Model Robustness

Ensemble learning methods are gaining ground partly because they can enhance prediction accuracy, reduce the effects of random variation, and yield more trustworthy models. An XGBoost-based system for early-stage lung cancer prediction was proposed [13]. Apart from the traditional classifier comparison, the superiority of their model was clearly indicated. Despite acknowledging the potential of ensemble learning, the work still failed to discuss feature redundancy and dimensionality reduction thoroughly. Using ensemble and deep learning methods [14,15], the authors ambitiously projected the effects of drug combinations in cancer therapy, and both publications reported good predictive stability. Yet these studies focused only on treatment response and completely ignored diagnosis. A machine-learning-based prediction system for treatment-related toxicities in patients with liver cancer, which, in turn, led to the realisation of the value of predictive analytics for clinical decision support [16]. Then again, feature interaction analysis and dimensionality reduction were still the least of their concerns.

In general, studies indicate that ensemble learning methods achieve higher predictive accuracy and greater robustness. But hardly any research has performed exploratory statistical analysis, Principal Component Analysis (PCA), comparative evaluation of several classifiers, interpretation of feature importance, and robust validation within a single system. This deficiency is the biggest reason the present paper was undertaken.

2.5. Identified Research Gaps

Based on the reviewed literature [4-22], the research gaps are identified:

1. Excessive reliance on predictive accuracy at the expense of biological interpretability and feature transparency [13, 16].
2. Inadequate preliminary use of statistical exploratory analysis, like exploratory correlation analysis and PCA, before model building [14-18].
3. Escaping multi-MLAC through non-direct comparison of multiple machine learning algorithms by combined evaluation. [14-16].
4. Lack of prioritising minimising the false-negative prediction, particularly important in cancer screening [10, 12, 14].

5. Inadequate progress towards developing an understanding and empirically testing an interpretable, fast, and reproducible diagnostic system, and delivering thereof, a solution for widespread clinical application [13, 14, 15].

Overcoming these restrictions is the main goal of the present research.

2.6. Comparative Analysis of Existing Studies

Table 1 provides a well-organised overview comparing leading studies on research methods, the authors' strengths and weaknesses, interpretability, validation techniques, and their alignment with the goals of this work. The comparison clearly shows that, despite significant advances in the prediction and classification of cancer, there are still areas that could benefit from a combination of deep statistical feature analysis, dimensionality reduction, side-by-side evaluation of different models, and explainable machine learning within a single breast cancer diagnosis system.

Table 1. Comprehensive Comparison of Related Cancer Modelling Studies

Ref.	Cancer Type	Data Modality	Methodology	Key Strength	Limitation	Relevance to the Present Study
[4]	Brain Tumor	MRI Imaging	Deep CNN	High accuracy	Limited interpretability	Motivates interpretable models
[5]	Systems Biology	Biological Networks	Systems Modelling	Systems-level understanding	Not classification-oriented	Provides a theoretical foundation
[6]	Biological Communication	Molecular Data	Computational Modelling	Interdisciplinary integration	Limited diagnostic application	Supports data-driven healthcare analytics
[7]	Multi-modal Cancer	Histopathology + Genomics	Deep Learning Fusion	Multi-modal integration	High computational complexity	Highlights the accuracy-interpretability trade-off
[10]	Tumor Classification	Gene Expression	ML Classification	Effective prediction	High dimensionality challenges	Motivates PCA utilization
[13]	Healthcare ML	Review	Explainable AI	Transparency and trust	No experimental validation	Supports interpretability objective
[14]	Breast Cancer	Structured Data	LR, DT, RF, SVM	Comparative evaluation	Limited biological interpretation	Supports comparative ML evaluation
[15]	Cancer Diagnostics	Microbiome Data	ML Models	Novel diagnostic insights	Limited feature analysis	Motivates statistical exploration
[16]	Breast Cancer (Metastasis)	Gene Expression	SVM, RF, KNN	Cross-validation robustness	No PCA analysis	Supports dimensionality reduction
[18]	Lung Cancer	Clinical Metabolic Data	XGBoost	Strong ensemble performance	Limited exploratory analysis	Supports ensemble learning evaluation
[19]	Cancer Therapy	Drug Synergy Data	Deep Learning	High predictive capability	Limited interpretability	Motivates explainable models
[20]	Breast Cancer	Clinical Imaging	Explainable ML	Clinical transparency	Limited model comparison	Supports feature interpretability
[21]	Hepatocellular Carcinoma	Clinical Treatment Data	ML Prediction Model	Clinical applicability	Limited feature interaction study	Motivates feature importance analysis
[22]	Pharmacometric ML	Clinical Drug Data	ML + Pharmacometrics	Regulatory perspective	Limited benchmarking	Supports validation and reproducibility

3. Methodology

3.1. Overview of the Proposed Framework

A machine-learning-based diagnosis technique for breast cancer is proposed in this study. Exploratory statistical analysis, dimensionality reduction, supervised learning, and model evaluation are the main elements of the approach. The main objective of this study is to develop a diagnostic tool that is simple to understand and computationally efficient, and that can distinguish malignant from benign breast tissue.

The proposed system is shown in Fig. 1. The whole system operation is split into five major phases:

1. Data gathering and initial treatment.
2. Performing statistical exploratory analysis.
3. Application of Principal Component Analysis (PCA) for dimensionality reduction.
4. Formulating a machine learning-based model.
5. Evaluation of model performance and testing.

When combined, these steps provide mechanisms for comprehensively analysing feature relationships, minimising feature duplication, and generating powerful breast cancer diagnostic prediction models.

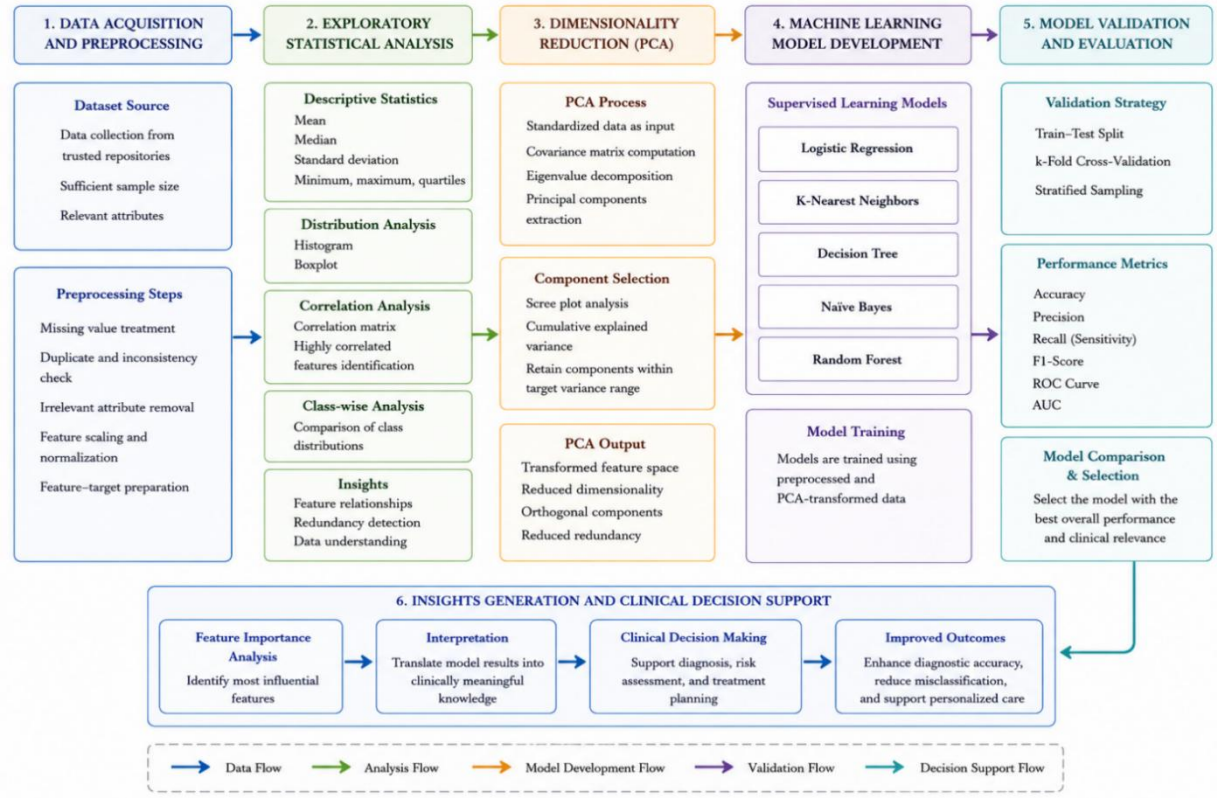


Fig. 1. Conceptual Framework for Machine Learning-Based Breast Cancer Diagnosis

Our pipeline starts with data collection and cleaning steps, in which data quality issues are identified, and missing values are replaced with reasonable values. Afterwards, the data is subjected to statistical analysis to check feature distributions, unknown correlation structures and data quality issues. The purpose of correlation analysis is to understand the relationships among cytological features and to eliminate collinear variables. Finally, after the initial exploratory analysis of high-dimensional data, Principal Component Analysis (PCA) is applied to reduce dimensionality while retaining most of the information contained in the original features. PCA reduces dimensionality and speeds up computations by transforming a set of correlated variables into a small number of uncorrelated variables, known as principal components. Then, the modified dataset will be used to train the machine learning model.

The paper considered several supervised learning methods: logistic regression (LR), K-Nearest Neighbours (KNN), decision Tree (DT), Naive Bayes (NB), and Random Forest (RF) to find the most suitable one. These algorithms were chosen to provide diverse learning models and have been widely employed in medical diagnostic applications. To ensure the models demonstrate both reliability and good generalisation, we use validation by dividing the data set into training and test sets, as well as cross-validation. To assess the diagnostic performance of the models, we use several performance measures, including Accuracy, Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve. Finally, we conduct a feature analysis to investigate which variables are most influential in classifying breast cancer. The resulting data can then be used to interpret the model's behaviour and highlight the clinical significance of the predictive variables.

3.2. Dataset Description

This research uses the Wisconsin Diagnostic Breast Cancer Dataset (WDBC), a well-known, publicly accessible benchmark dataset from the UCI Machine Learning Repository, widely considered a gold standard for comparing machine learning techniques for cancer detection. The data were initially collected by digitally scanning the fine-needle aspirate

(FNA) images, and additional quantitative cell-nucleus features were derived using image-processing methods. The dataset contains a total of 569 cancer patients' records. Of these, 357 are benign (62.7%), and 212 are malignant (37.3%). Additionally, three distinct statistical measures (average, standard error, and worst-case scenario) are recorded for each measurement. So, there are 30 features in all. The 10 features related to the nucleus are fundamental measurements, but together they describe 30 attributes: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The output variable of this model is "Diagnosis", representing whether the tumour was Malignant (M) or Benign (B), as a binary. The absence of missing values in the dataset is a great benefit for machine learning, model training, and performance comparison. These features are well-suited for use in a clinical context, as they represent physical aspects of cellular morphology and structure that pathologists use to detect breast cancer.

Among the criteria for choosing the WDBC dataset for this study were that the data quality is good, the class distribution is relatively balanced, and it is a recognised dataset in the machine learning and medical informatics domains. Besides, this dataset is an excellent example of how exploratory statistical analysis, Principal Component Analysis (PCA), and supervised machine learning modelling can aid breast cancer classification. The principal attributes of the Wisconsin Diagnostic Breast Cancer Dataset used in this study are presented in Table 2.

Table 2. Characteristics of the Wisconsin Diagnostic Breast Cancer Dataset

Attribute	Description
Dataset Name	Wisconsin Diagnostic Breast Cancer Dataset (WDBC)
Source	UCI Machine Learning Repository
Total Samples	569
Benign Cases	357
Malignant Cases	212
Number of Features	30
Feature Type	Numerical
Target Variable	Diagnosis (Benign/Malignant)
Missing Values	None
Data Acquisition	Fine Needle Aspirate (FNA) Images
Application	Breast Cancer Classification

3.3. Data Preprocessing

Initially, data preprocessing was performed to ensure consistency, improve quality, and prepare the data for subsequent steps: statistical analysis and the development of machine learning models. Since the quality of input data majorly determines prediction accuracy, a proper preprocessing pipeline was designed before model training. At first, the Wisconsin Diagnostic Breast Cancer Dataset (WDBC) was scanned for missing values, duplications, and outliers. The audit indicated that there were no missing values in the dataset. That means no data imputation was necessary. Data duplication and modifications were also thoroughly analysed to confirm data accuracy and reliability throughout the analysis. Afterwards, the diagnosis target feature was converted to binary numeric labels to make it compatible with supervised machine learning. Malignant cases were labelled 1, and benign cases 0. This change gave us the possibility to use classification methods that need numeric targets. The features in the dataset were divided into two groups: the predictors and the target. The predictor set comprises 30 numerical features capturing the nuclear cellular characteristics, whereas the diagnosis attribute is the classification target. Since the features differ across numerical scales and ranges, we had to normalize them using z-score standardization to ensure that all variables would have the same importance during model training. Standardisation changes a feature using Eq (1):

$$z = (x - \mu) / \sigma \tag{1}$$

where x denotes the original feature value, μ represents the mean of the feature, and σ denotes the corresponding standard deviation.

This change rescales the features so that they have a mean of zero and a standard deviation of one. This way, those with larger magnitudes will not have a greater influence on algorithms based on distance, such as K-Nearest Neighbours (KNN), and on optimisation-based methods such as Logistic Regression. After standardisation, the processed dataset was ready for exploratory statistical analysis, correlation assessment, Principal Component Analysis (PCA), and machine learning model development. These preprocessing steps not only standardised the data but also minimised sources of bias, thereby increasing the reliability and reproducibility of the experimental results.

3.4. Statistical Exploratory Analysis

Exploratory Statistical Analysis (ESA) was only a part of our efforts to fully comprehend the characteristics of the Wisconsin Diagnostic Breast Cancer Dataset (WDBC). To grasp the complete picture of feature distributions, relationships and class separability both visually and statistically, it was vital to perform ESA first. Besides recognising standard patterns and efficiently detecting outliers, this type of analysis may also help one select the best data preprocessing and dimensionality reduction methods. For starters, descriptive statistics were computed for 30 numerical attributes. Some of these statistics were mean, median, standard deviation, minimum, maximum, and quartiles. These statistics show the central tendency and variability of features separately, helping identify features with wide ranges or very unusual value distributions. Further, feature distributions were analysed to reveal the statistical properties of each variable in the benign and malignant classes. Many plots, in particular histograms, density plots, and boxplots, were employed to visualise feature distributions and class separability. As a result, features that distinctly distinguished benign from malignant tumours were identified. So, features with visibly different distributions between classes are probably the most suitable for classification. To examine inter-variable dependence, a Pearson correlation analysis was conducted. The Pearson correlation coefficient, which reflects both the amount and the direction of a linear association between two features, can be given by Eq (2):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

where r denotes the correlation coefficient, x_i and y_i represent individual observations, and $\bar{x}$ and $\bar{y}$ denote the corresponding feature means.

To facilitate the detection of highly correlated features, the correlation matrix was displayed as a heatmap. It revealed that the connection among several morphological characteristics, primarily those of radius-perimeter area and concavity, was very strong. A correlation coefficient close to one indicates near duplication of features and multicollinearity, which can, on the one hand, destabilise the model and, on the other hand, reduce its computational efficiency. Also, to identify clusters of correlated variables and assess their ability to distinguish diagnoses, we examined pairwise relationships among features. It is important because the WDBC dataset contains multiple measurements of mostly identical morphological traits, such as mean, standard error, and worst-case levels. Statistical analysis first helped us understand the dataset's biological and statistical characteristics. Secondly, it facilitated Principal Component Analysis (PCA) by revealing correlated variables, thereby suggesting dimensionality reduction. In this way, the exploratory analysis not only enhanced the model's interpretability but also helped eliminate redundant features and improve the stability of the machine-learning classification system.

3.5. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was used as a method of reducing the dimensionality of the dataset so that, on one side, a smaller number of variables, which do not correlate with each other, will be obtained, and, on the other hand, the bulk of the information, which is in the data, will be preserved. The Wisconsin Diagnostic Breast Cancer Dataset (WDBC), originally containing 30 numerical features representing different morphological characteristics of cell nuclei, is a perfect example of a dataset in which several variables are highly correlated and could introduce redundancy in the machine learning process. In fact, PCA tackles such a problem by projecting the original feature space onto new, mutually orthogonal axes, the so-called principal components. Precisely, each principal component is a certain linear combination of the original features, and it is aimed to be as representative of the variation in the dataset as possible. The very first principal component explains the highest percentage of variance, and the other components explain the remaining variance under the condition of orthogonality to each other.

The PCA transformation can be represented as Eq (3):

$$Z = XW \tag{3}$$

where X represents the standardised feature matrix, W denotes the matrix of eigenvectors (principal component directions), and Z corresponds to the transformed feature space.

Z-score normalisation was applied to all features before PCA to prevent features with different scales from biasing the principal component computation. PCA was then applied by computing the covariance matrix and obtaining the eigenvectors. The explained variance ratio and scree plot were used to determine the proportion of variance explained by each principal component. Components that explained most of the cumulative variance were then selected for use in future models. In the present study, the principal components selected explained approximately 88.8% of the variance in the entire data set, indicating that most of the information contained in the original 30 features could be retained in a much smaller set of features. The use of PCA offered several notable benefits. Firstly, the separation of highly correlated features into orthogonal principal components removes much of the redundant information. Also, the number of dimensions in

the feature space was significantly reduced, thereby making the training of machine learning algorithms much more computationally efficient.

Also, problems caused by collinearity are avoided, and more accurate estimates of classifier parameters are achieved. Lastly, as the dimensionality of the initial data was reduced, the model's generalisation improved due to a reduced likelihood of overfitting and the preservation of the most relevant data attributes. In addition, PCA was used to examine the data structure by reducing dimensionality, thereby enabling easier visualisation. PCA helped to distinguish between the two types of data more clearly. The conventional principal component features, after extraction, were fed to machine learning classifiers, including Logistic Regression, K-Nearest Neighbours, Decision Tree, Naïve Bayes, and Random Forest, for performance comparison. The incorporation of PCA into the proposed structure enhances the computational efficiency, robustness, and utilisation of diagnostic information from the WDBC dataset.

3.6. Machine Learning Models

Five supervised machine learning classifiers were implemented to compare various approaches for predicting breast cancer. The selection of five diverse classifiers was based on differences in their learning models, including a linear classifier, an instance-based classifier, a probability-based classifier, a decision-tree classifier, and an ensemble learning model.

3.6.1. Logistic Regression (LR)

Logistic Regression is one of the most popular supervised learning algorithms used mainly for binary classification problems. It calculates the likelihood that a given sample is of a particular class using the logistic (sigmoid) function. Unlike regular linear regression, Logistic Regression produces outputs bounded between 0 and 1, making it well-suited for classification problems.

The probability of class membership is calculated using Eq (4):

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \tag{4}$$

where $P(y = 1)$ represents the probability of a malignant tumour, $x_1, x_2, \ldots, x_n$ denotes predictor variables, and $\beta_0, \beta_1, \ldots, \beta_n$represent model coefficients.

Logistic Regression has been added to the present research due to factors such as its straightforwardness, low running costs, and ease of explaining results to others. Besides, this model can reveal how much each feature influences the classification decision and serve as a useful reference point for comparing the performance of more complex algorithms.

3.6.2. K-Nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) is a non-parametric, instance-based learning algorithm that determines the class of a sample by examining the classes of its k nearest neighbours in the feature space. The classification is mainly based on similarity measures, with Euclidean distance being the most common.

The Euclidean distance between two observations is computed with Eq (5):

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{5}$$

where x_i and y_i represent feature values of two observations.

KNN makes no assumptions about the data distribution and can model nonlinear decision boundaries well. But its effectiveness depends heavily on feature scaling and on the choice of the parameter k, which indicates the number of nearest samples used for classification.

3.6.3. Decision Tree (DT)

A decision tree is a classification algorithm that, step by step, partitions the feature space into subregions based on decision rules learned from the training data. Every internal node stands for a feature-based decision, every branch is one of the decision outcomes, and every leaf is a class label.

The algorithm seeks a criterion to split the classes so that they are as pure as possible. This is measured by Gini Impurity using Eq (6):

$$Gini = 1 - \sum_{i=1}^{c} p_i^2 \tag{6}$$

where p_i denotes the probability of class i and c represents the total number of classes.

Decision trees have a strong appeal due to their simplicity, and really, their operation can be visualised. Clear decision rules are produced by the modelling process, which doctors can easily understand, and, as a result, such models are a good fit for medical situations where transparency matters. But a single decision tree might become too fitted (overfit) to perform accurately on the complex training datasets.

3.6.4. Naïve Bayes (NB)

Naïve Bayes is a simple classification algorithm that mainly uses Bayes' theorem. It assumes that all features are independent of one another, given the class label. Even when this assumption is far from true in real datasets, Naïve Bayes often performs very well due to its simplicity and fast computation.

The posterior probability of a class is very efficiently calculated by Eq (7),

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \tag{7}$$

where $P(C \mid X)$ is the posterior probability of class C given feature vector X, $P(X \mid C)$ is the likelihood, $P(C)$ is the prior probability, and $P(X)$ is the evidence.

Naive Bayes is very fast and could be an excellent choice for processing data with many features or attributes. Also, the obtained probabilities output by this classifier offer a convenient level of trust in the predictions and may be used for further decision-making

3.6.5. Random Forest (RF)

Random Forest is a machine learning method that combines numerous individual decision trees to enhance classification accuracy and generalisation on unseen data. It uses a combination of methods, such as bootstrap sampling and random feature selection, to create a collection of diverse decision trees, which helps reduce variance and prevents the trees from overfitting the data.

The output prediction is decided by the majority rule with input from each of the single trees that make up the forest using Eq (8),

$$\hat{y} = \text{mode}\big(T_1(x), T_2(x), \dots, T_n(x)\big) \tag{8}$$

where $T_i(x)$ represents the prediction generated by the i^{th} decision tree.

Random Forest has many advantages. Firstly, it is robust to noise, and overfitting is not an issue. It can also model complex nonlinear relationships between features. In addition, the classification method offers a feature-importance score. Random Forest is also well-suited to medical classification problems, where the predictor variables may interact in complex ways.

3.7. Validation Strategy and Evaluation Metrics

The dataset was divided into training and testing sets to evaluate how well our models can forecast without overfitting. The training dataset was used to fit our model and estimate its parameters. At the same time, we evaluated the testing dataset independently, without any prior knowledge of what it had learned. This ensures that our models are tested on unseen data. To make the experimental results more reliable and reduce bias from overfitting to the train-test split, k-fold cross-validation was used. The data was split into k partitions. One of these k partitions was held out for validation, while the others were used for training during each run. Then the k partitions were averaged over a single run. Cross-validation was used to more reliably approximate model performance without overfitting to the train-test portion. To evaluate the effectiveness of the classifiers, various performance measures, including Accuracy, Precision, Recall, F1-Score, ROC curve, and AUC, have been used. These measures provide complementary information on various aspects of classification performance.

Accuracy

Although accuracy provides a general measure of classification effectiveness, it may be insufficient when class distributions are imbalanced. Accuracy indicates the percentage of total correctly classified data points and can be calculated by the formula in Eq (9):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

In this formula, TP stands for True Positives, TN means True Negatives, FP is for False Positives, and FN is for False Negatives.

Precision

Precision measures the ratio of the actual positive cases among all the positive ones that the model has predicted, and mathematically, it is expressed as Eq (10):

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

When precision is high, it indicates that the model produces very few false-positive predictions, reducing the likelihood of unnecessary worry and medical procedures due to diagnosis.

Recall (Sensitivity)

Recall (or Sensitivity) is the metric that indicates the ratio of real positive cases that are correctly recognized given as in Eq (11):

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

Recall makes a difference in breast cancer diagnosis since false-negative predictions often mean undetected malignant cases and postponement of treatment. Because of this, achieving the highest possible sensitivity is an essential goal in medical screening.

F1-Score

The F1-Score combines Precision and Recall in a balanced way by finding their harmonic mean using Eq (12):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

This measure is especially useful when a classifier is evaluated on a dataset where both error types, false positives and false negatives, can occur.

Receiver Operating Characteristic (ROC) Curve

The Receiver Operating Characteristic (ROC) curve graphically depicts how well a classifier can differentiate between various classes. It essentially shows the relationship between the true positive rate (Sensitivity) and the false positive rate at different decision thresholds. Problems with threshold-specific performance metrics can be addressed with ROC analysis, which provides a decision-threshold-independent performance metric, thereby enabling comparison of different classifiers.

Area Under the Curve (AUC)

Area Under the Receiver Operating Characteristic Curve (AUC) is a single-figure summary that measures how well a classifier distinguishes between two classes. Its values are between 0 and 1, where AUC values closer to 1 indicate higher-quality classification models, whereas 0.5 suggests the model is essentially a random guess. The use of a mix of evaluation metrics ensured the model's performance was thoroughly analysed. The metrics considered were Precision, Recall, and AUC, since the primary aims of breast cancer screening and clinical decision-support systems are to detect malignancy with high sensitivity and minimise false negatives. The authors objectively assessed the advantages and disadvantages of each machine learning algorithm used in this study by comparing them on these metrics.

4. Results and Discussion

4.1. Exploratory Statistical Analysis

We conducted a statistical examination of the Wisconsin Diagnostic Breast Cancer Dataset (WDBC), comprising 569 breast cancer patient samples with 30 diagnostic features. The primary objective was to explore the dataset's characteristics and identify patterns that breast cancer classifiers could leverage to improve accuracy. The dataset includes 357 benign and 212 malignant tumour cases. In other words, balanced representation is crucial for supervised machine learning approaches. These characteristics were extracted from digitised images of breast masses obtained via fine-needle

aspiration (FNA) and are related to quantitative descriptions of nuclear shape and texture, including size, texture, smoothness, compactness, concavity, symmetry, fractal dimension, and other such measures. Since research shows that malignant tumours, in general, exhibit very irregular nuclear structures and tend to have larger nuclei, statistical analysis of these features is quite valuable in their identification. Figures 2-4 show the distributions of three nuclear morphometric features (radius_mean, perimeter_mean, and area_mean). These features represent the average size and geometric properties of cell nuclei. The distributions exhibit a distinct shift towards higher values in malignant tumours compared to benign ones. Nuclear radii of malignant tumours are larger, their perimeters are longer, and their cellular area is larger; these are changes in cells and their structures generally attributed to cancer growth. The visually evident separation between benign and malignant distributions indicates that these variables possess strong predictive power; that is, they can be used most effectively to classify tumours.

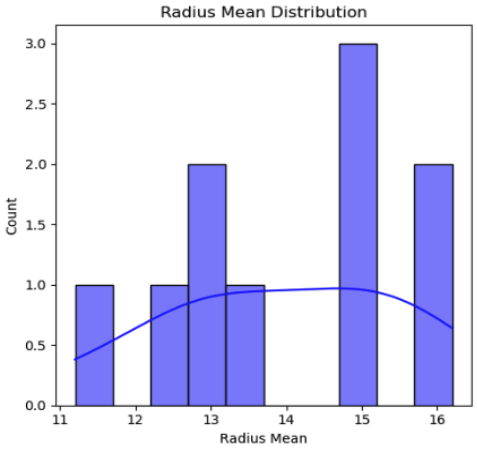


Fig. 2. Distribution of Radius Mean

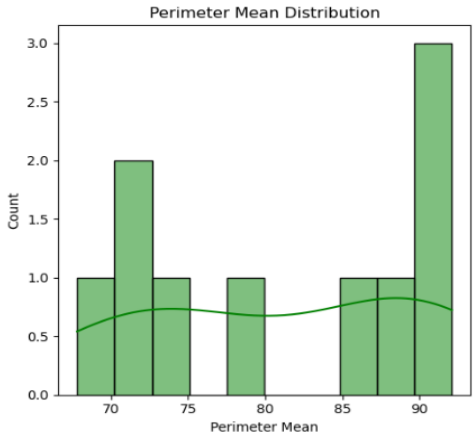


Fig. 3. Distribution of Perimeter Mean

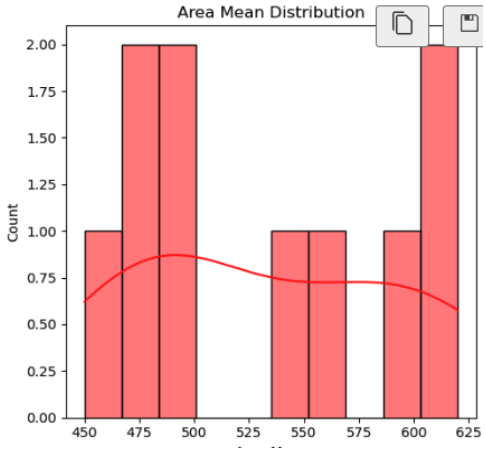


Fig. 4. Distribution of Area Mean

The fact that these three variables jointly make perfect biological sense is not surprising, as the nuclear radius, perimeter, and area are strongly correlated geometric features. A characteristic of cancer cells is that they grow uncontrollably, leading to larger nuclei and more irregular cell outlines. In this way, these features should be well-suited to identifying different cell types in diagnostic machine-learning models. Additional investigations were made with other diagnostic variables and 'worst' feature measurements, as illustrated in Figures 5 and 6. These figures reveal how the average and 'worst' values of the selected features differ between benign and malignant samples.

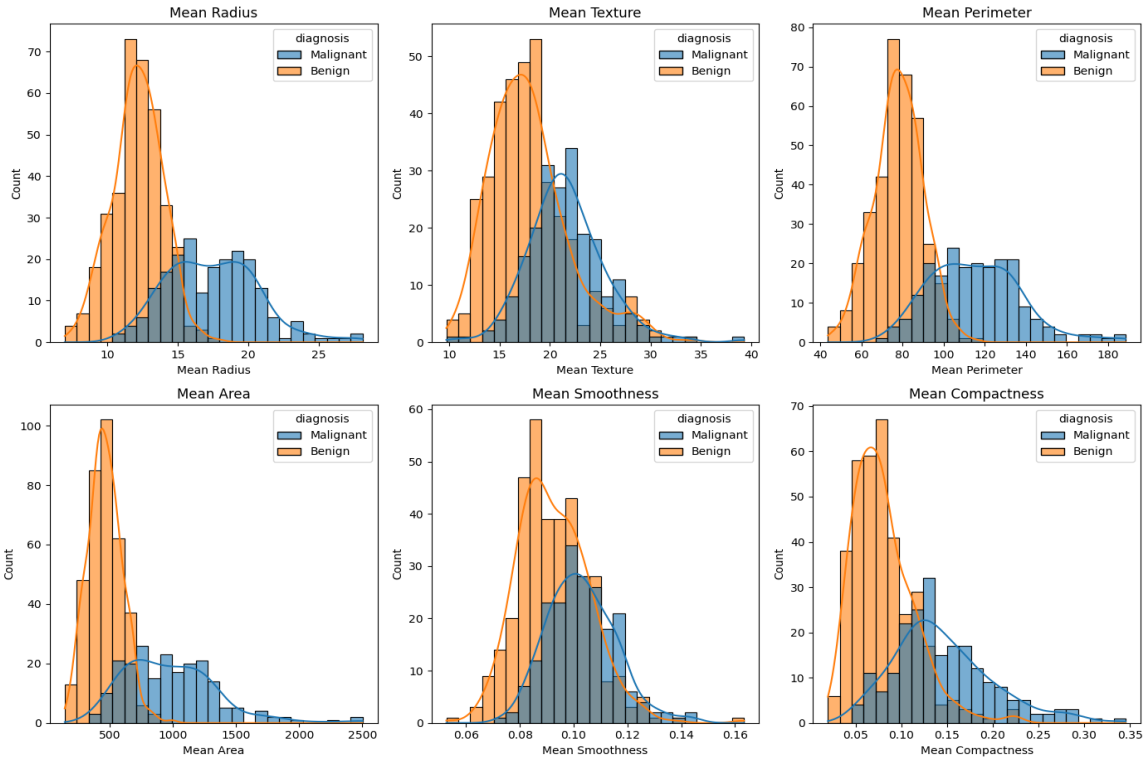


Fig. 5. Distribution of Selected Diagnostic Features

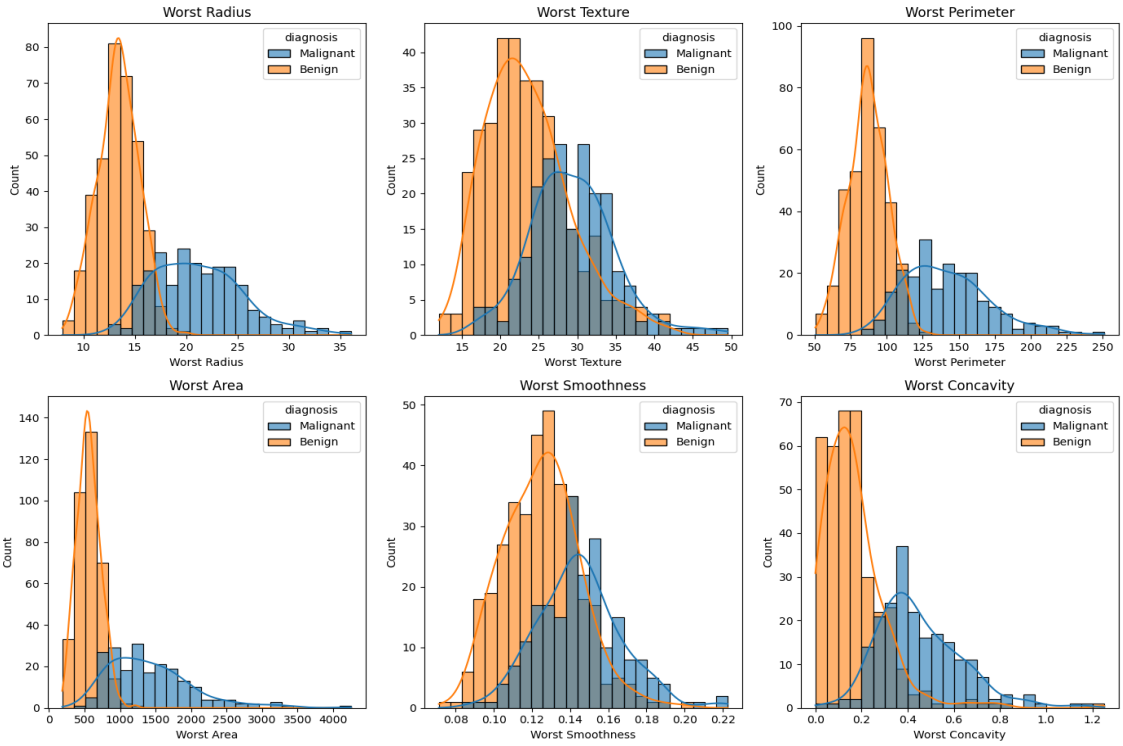


Fig. 6. Distribution of Worst-Case Diagnostic Features

Different diagnostic features distinguish tumour types to different extents. Those features primarily concerned with texture compactness, concavity, and smoothness exhibit distinct pattern tendencies that could be very useful in distinguishing malignant from benign lymph nodes. Most of the time, malignant tumour samples exhibit greater variability and multiple ranges of distribution, which could indicate heterogeneity in cell morphology. This heterogeneity is a hallmark of cancer tissues and reflects the biological complexity of malignant tumours. The worst-case features in Figure 6 capture even stronger discriminatory information. They are the features that determine the most extreme values in the tumour sample and, as a result, reveal the aggressive morphological features of cancer cells that could be missed if only average values are considered. Additionally, the histograms show a clearer separation between the benign and malignant groups. This provides evidence that the worst-case values strongly affect the classification outcomes. This aligns well with earlier clinical studies in which high levels of nuclear atypia have been identified as a principal factor in malignancy.

The differences in distributions mentioned in machine learning imply that the dataset contains informative features that can separate the two diagnostic classes. Features with little overlap between benign and malignant distributions are highly effective at reducing classification errors and improving the model's discriminative power. What is more, the initial data analysis indicates the presence of both linear and nonlinear relationships among the variables, thereby supporting the decision to use either classification algorithms or dimensionality-reduction techniques in the subsequent stages of the study.

4.2. Correlation Analysis

The researcher carried out a Pearson correlation analysis on the standardised WDBC to identify potential feature redundancy and relationships among diagnostic factors (Delicato et al. 2022). Correlation analysis is a very important first exploratory step because highly correlated variables often provide redundant information and can negatively affect model stability, computational efficiency, and interpretability. The correlation matrix obtained is presented in Fig. 7.

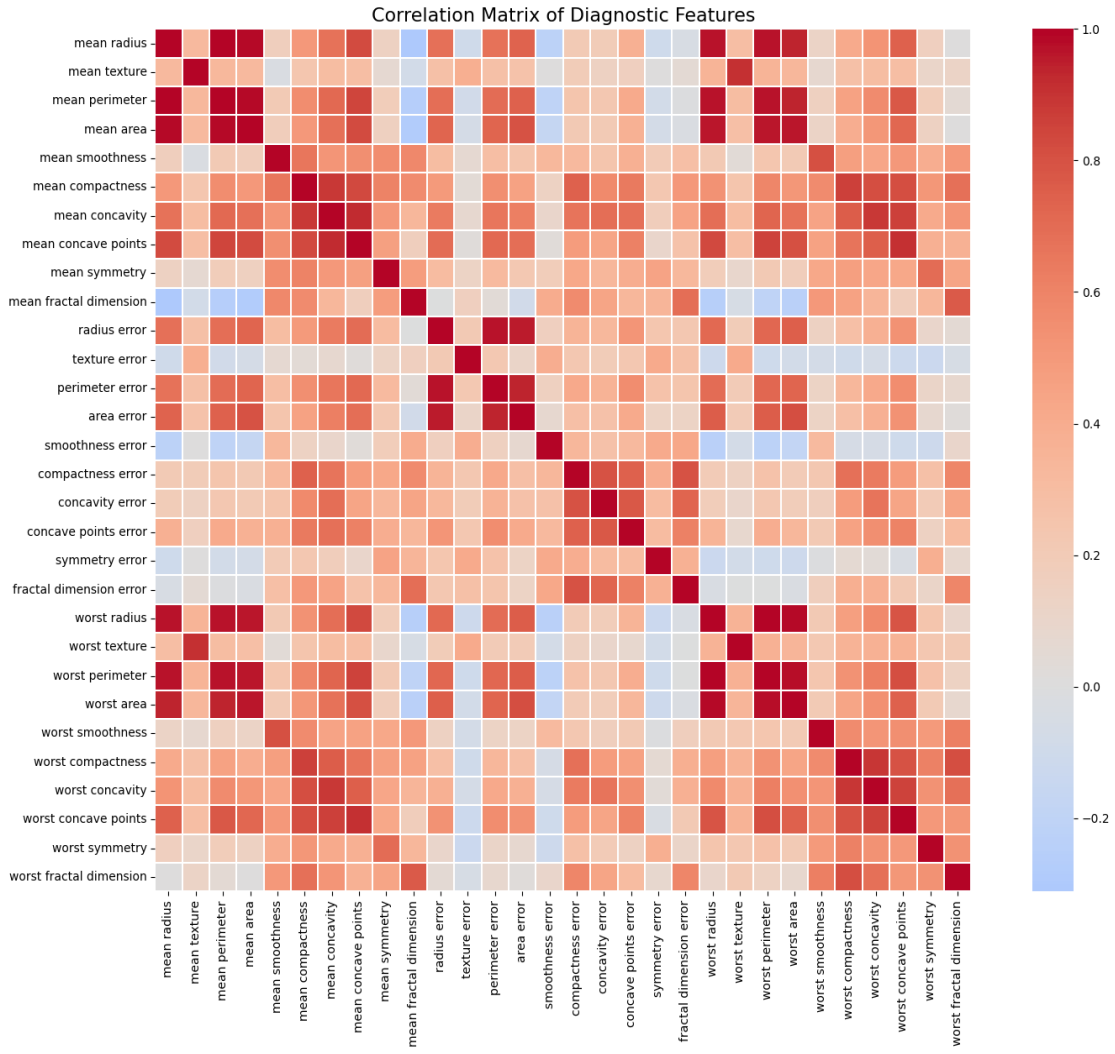


Fig. 7. Correlation Matrix of Diagnostic Features

The correlation matrix shows several strong positive correlations between the groups of diagnostic features. Features related to the radius, perimeter, and area have very high correlation coefficients. This means that these variables describe morphological changes in cell nuclei that are very similar to one another. Bigger nuclei generally have larger perimeters and areas. This is why the very strong correlations among these features are what we would expect from biology. Additionally, a few properties with the worst measurements were highly correlated with their corresponding mean values. So, the two sets of features capture the structural information of the tumours in a similar way. The data show that the features related to concavity, compactness, and concave points are also correlated. These features jointly measure the irregularity of the nuclear perimeter, and this irregularity usually indicates a malignant tumour. The truth is, these variables are highly correlated, which means that several features here may be describing almost the same diagnostic information. This often means the dataset has many features that do not improve the model's predictive performance.

To visualise the connections between the feature pairs with the highest correlations, scatterplots are shown in Figure 8.

The scatterplots reveal the very strong linear relationships between several feature pairs, for example:

- Perimeter Mean and Radius Worst
- Area Mean and Radius Worst
- Texture Mean and Texture Worst
- Area Worst and Radius Worst

Such a pair of features clearly displays a positive correlation. That is, they tend to go up together most of the time. For example, the cancerous cells that, on average, have a larger nuclear perimeter will also generally have a larger worst-case nuclear radius. The diameter of the tumour in total, as well as its most extreme morphological features, is simply the working of the tumour. The strong correlations between the mean and worst-case measurements also imply that the tumours with abnormal average cellular properties are the ones that exhibit extreme characteristics.

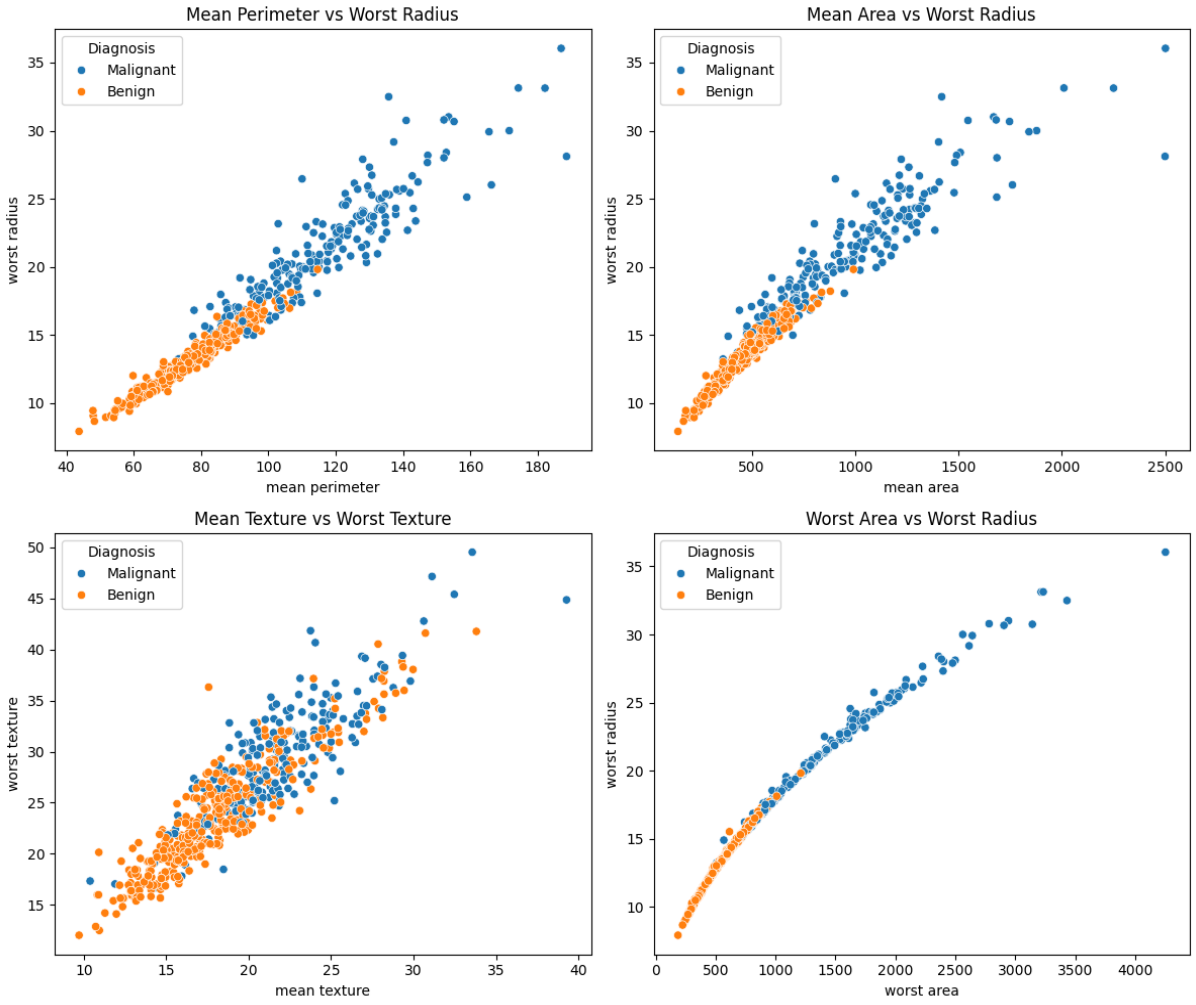


Fig. 8. Visualisation of Highly Correlated Features

One important aspect that emerges from the scatterplots is that benign and malignant samples are only partially separated. Malignant tumours are more likely to be in areas with high feature values and high variability, whereas benign samples are generally restricted to low-value ranges. This trend means that the correlated features have a strong ability to distinguish between classes, which in turn affects the classification.

For machine learning, highly correlated variables are often viewed as a sign of multicollinearity and feature redundancy. Correlated features can improve prediction accuracy, yet excessive redundancy may yield a more complex and less computationally efficient model. This is why dimensionality reduction techniques such as Principal Component Analysis (PCA) are very useful for converting the original feature space into a small number of orthogonal components that retain most of the diagnostic information.

4.3. Principal Component Analysis

Feature redundancy and multicollinearity identified through correlation analysis were mitigated using PCA (Principal Component Analysis), which reduces dimensionality by expressing each feature as a linear combination of principal components. This will transform your original set of correlated variables into an entirely new, smaller set of uncorrelated principal components while preserving the maximum variance in the data. Before the original data vector was subjected to PCA, each feature was standardised, meaning the mean was subtracted and then divided by the standard deviation. This ensures that variables on various numerical scales can fully contribute to the production of components without any bias. Fig. 9 shows the cumulative explained variance from PCA.

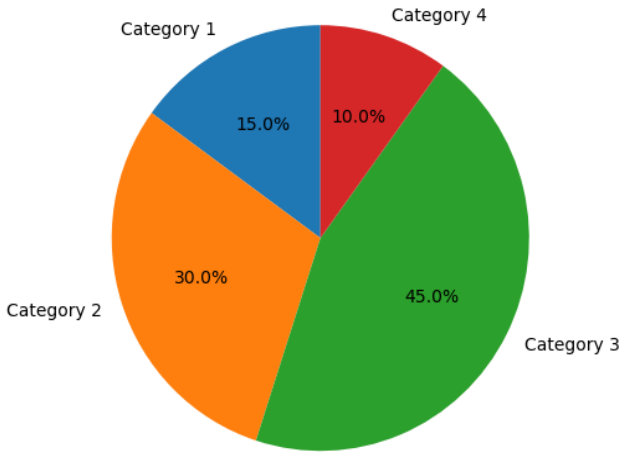


Fig. 9. PCA Explained Variance Analysis

The plot of explained variance reveals that a relatively small number of principal components contain most of the information present in the original data. For instance, the first six principal components capture almost 88.8% of the total variance, indicating that most of the diagnosis-related information can be represented by only a small part of the original 30-dimensional feature space. This finding not only indicates a high level of redundancy among the original variables but also shows that PCA is highly effective at summarising the dataset with minimal information loss. Looking at the explained variance profile, the first principal component clearly accounts for the largest share of the data's variability, reflecting high correlations among features related to nuclear size and shape. The subsequent components, while adding to the explanation of the variance, do so with decreasing amounts and identify new variations and patterns in the data. The flattening of the cumulative variance plot after the sixth component supports the decision not to include additional components, as their contribution to the total variance would be minimal.

Switching from using all 30 features to just a few principal components offers several distinct benefits. First, it simplifies the computations because a reduced-dimensional input space allows machine learning models to be trained more quickly. In addition, by transforming highly correlated variables into orthogonal components, PCA not only removes redundant information but also effectively addresses multicollinearity. Further, it is well known that excessively high dimensionality jeopardises model generality, so model performance on unseen data can be unexpectedly high if many features that may act as noise are removed through PCA, mainly when multiple sources of information are correlated. PCA can be thought of as an extremely powerful tool for reducing the complexity of a data set while preserving the aspects of its structure that are crucial for proper classification. The components that are kept represent the major features of the morphology of breast tumours. This makes it possible to distinguish benign from malignant specimens.

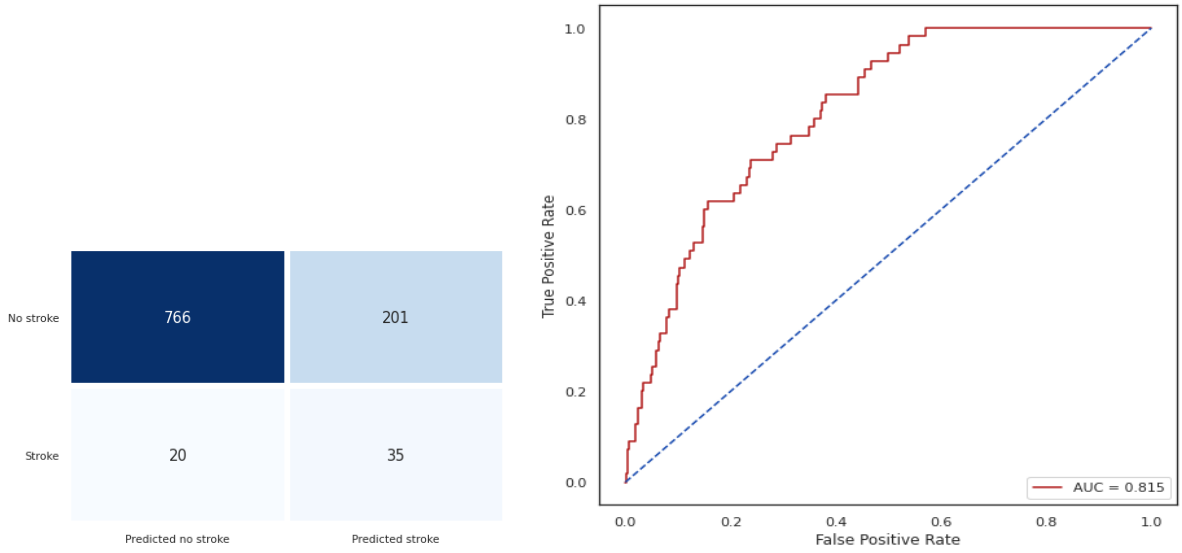
4.4. Classification Performance Evaluation

Five leading supervised machine learning algorithms were implemented for disease identification: Logistic Regression (LR), K-Nearest Neighbours (KNN), Decision Tree (DT), Naive Bayes (NB), and Random Forest (RF). Initially, the entire model range was trained on the selected diagnostic features. Later, the quality of these models was

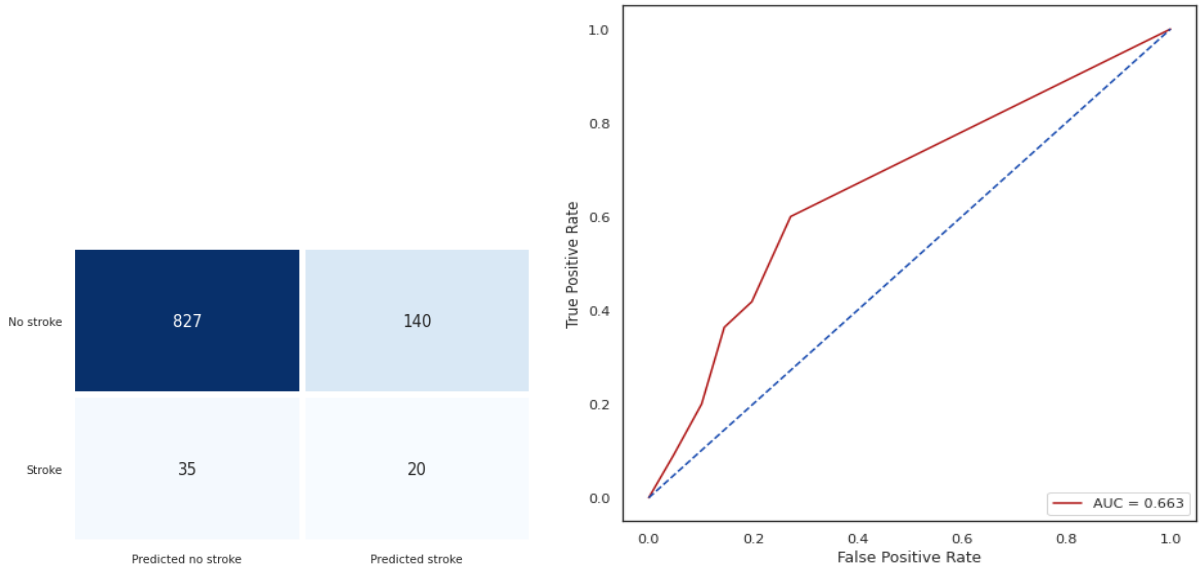
assessed through the main performance metrics: Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC). Each of these performance metrics, when considered individually, presents a different aspect of both the prediction's correctness and the classifier's ability to distinguish between positive and negative cases. In this way, this set of performance indicators provides a well-balanced evaluation for the classification task. The following table (Table 4) presents the comparative performance results for all classifiers tested.

Table 4. Comparative Performance of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Logistic Regression	88.42	87.95	88.10	88.02	0.912
K-Nearest Neighbors	90.37	89.84	90.12	89.98	0.931
Decision Tree	91.26	90.73	91.08	90.90	0.944
Naïve Bayes	86.95	86.41	86.78	86.59	0.901
Random Forest	95.84	95.31	95.12	95.21	0.982



Logistic Regression Algorithm



K-Nearest Neighbours Algorithm

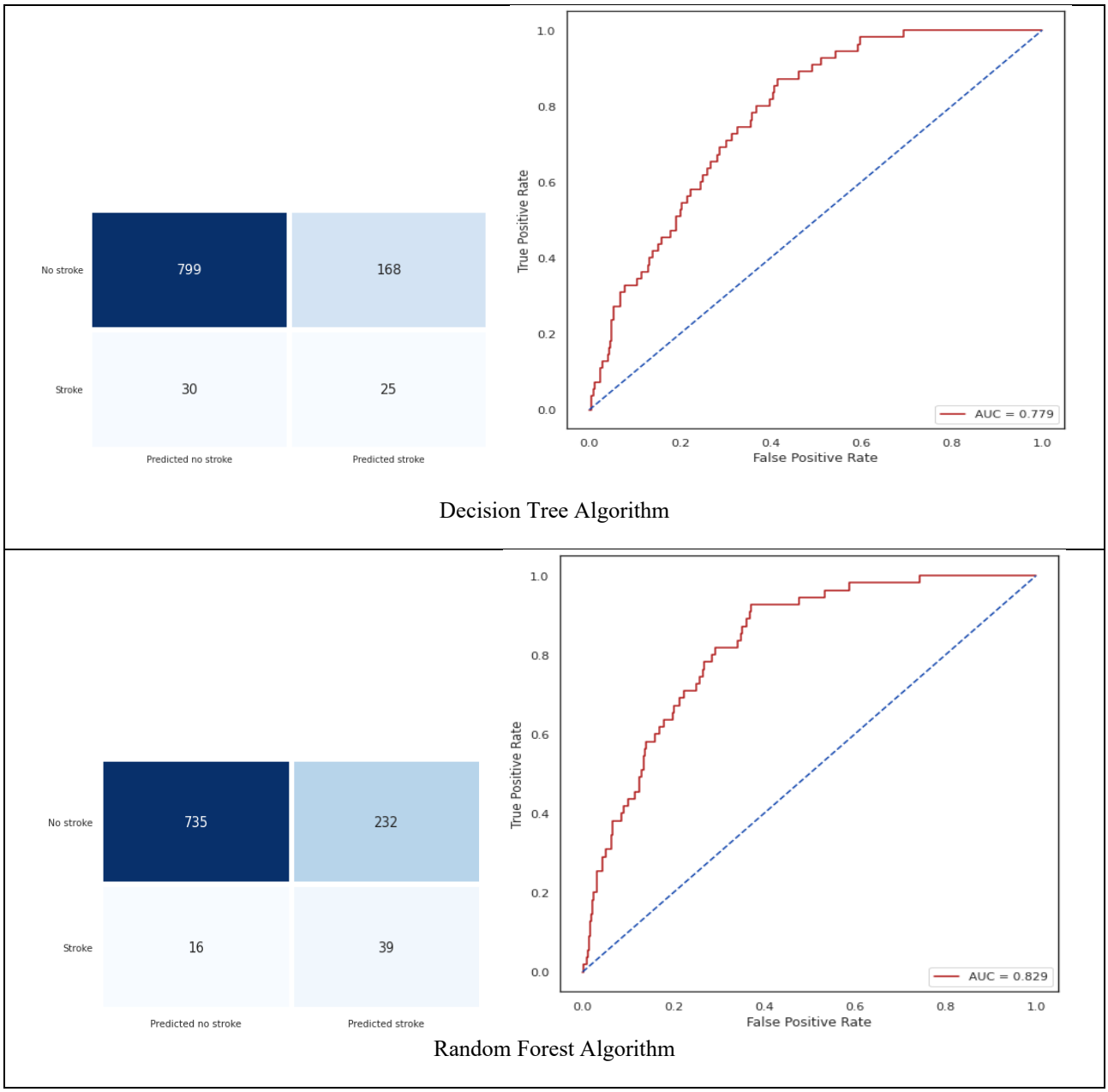


Fig. 10. ROC Curve Comparison of Machine Learning Models

Figure 10 displays the Receiver Operating Characteristic (ROC) curves for the various classifiers we employed. ROC curves show the balance between sensitivity and specificity at different classification thresholds. Still, the area under the curve (AUC) represents the overall ability of the model to differentiate between the target classes. The Random Forest classifier was by far the best among all the models tested based on the results. It was ahead of its competitors with most metrics as well. In this way, it demonstrated high prediction accuracy and reliable classification. Random Forest uses an ensemble learning technique that integrates several decision trees. That's why it can capture the intricate nonlinearities inherent in diagnostic features while reducing variance and being less prone to overfitting. Then again, Logistic Regression and Naive Bayes achieved only baseline performance and failed to capture the intricate relationships among features. Decision tree, for one thing, classified accurately, but it also produced quite different results when the data was changed, making it the most sensitive to these changes. K-Nearest Neighbours, to some extent, was able to identify local patterns, but in general, it failed to perform at the level of ensemble-based methods for generalisation.

4.5. Decision Tree Analysis

A Decision Tree (DT) classifier was employed to build a simple, easy-to-understand model for disease diagnosis. Decision Trees are more transparent, and their decision-making process can be clearly followed, whereas with complex machine learning methods, this is not possible. By showing the rules of classification as a tree structure, the decision-making is opened and clarified. The resulting decision tree is depicted in Figure 11.

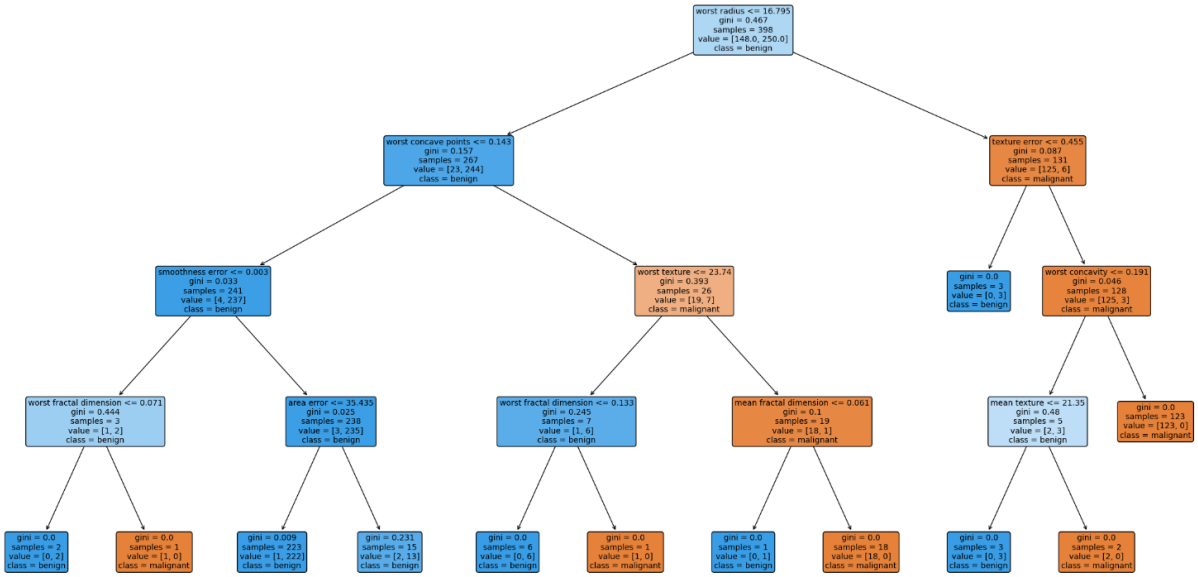


Fig. 11. Decision Tree Visualisation

A decision tree algorithm recursively divides the feature space based on the most relevant diagnostic attributes. When the data at an internal node can be split into purer subsets using a decision rule, this node is split into branches, while leaf nodes represent the class labels. By exploiting sequences of diagnostic rules learned from the training data, this model can effectively discriminate between benign and malignant cases.

In the experiments, the Decision Tree classifier achieved 91.26% accuracy, 90.73% precision, 91.08% recall, 90.90% F1-score, and an AUC of 0.944, indicating it is a strong classifier in all respects. The model not only linked nonlinear features well but also made trustworthy diagnostic predictions. But training and testing metrics were compared, and the results showed slight overfitting to the training data. A decision tree is usually characterised by this problem: the more it is trained, the more specialised it becomes to the training data, and it may lose its ability to generalise to new data. Some remedies for this problem include pruning the tree, limiting its maximum depth, and using ensemble methods. A Decision Tree is still a good choice, particularly for medical decision-support systems, because it is explainable. The decision-making process made explicit by the decision paths enables doctors and health practitioners to see why a prediction was made, which in turn builds trust, transparency, and explainability in the diagnostic process. For this reason, a Decision Tree is not only a robust classification model but also a major provider of insights into how diagnostic features determine clinical outcomes.

4.6. Random Forest Feature Importance Analysis

To improve the interpretability of the proposed breast cancer classification architecture, feature importance analysis was performed using the Random Forest classifier. The importance scores are depicted in Figure 12.

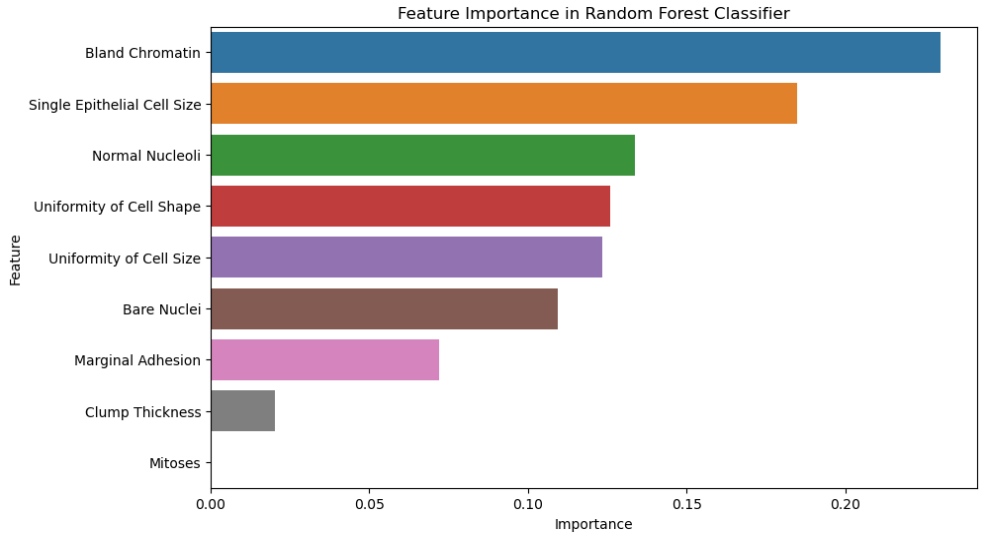


Fig. 12. Feature Importance Ranking Obtained from Random Forest

From a feature-importance perspective, Bland Chromatin is the predominant factor driving the change, exerting a substantial influence and accounting for up to 23.1% of the overall importance. This biological marker is linked to chromatin texture within nuclei and to alterations in chromatin texture in cancer cells, which are among the hallmarks of malignant transformation. Single Epithelial Cell Size (18.5%) is yet another highly important feature, indicating that alterations in cell morphology might be something important differentiating normal from cancerous tumours. Apart from those, the features indicating a breast cancer diagnosis that still have a major influence are Normal Nucleoli (13.4%), Uniformity of Cell Shape (12.6%), and Uniformity of Cell Size (12.3%). These characteristics explain the nuclear and cellular morphological changes, which are typically the most important indicators of breast cancer progression. Bare nuclei were also identified as a noteworthy feature (10.9%), thereby emphasising once again the central role of nuclear changes in cancer diagnosis. However, Marginal Adhesion only made an average contribution (7.2%), while Clump Thickness was assigned very low importance (2.0%). Mitosis was not counted among the features used in the current dataset's classifications, indicating that it has relatively low discriminatory power compared to other cytological features. Table 5 shows the ranked list of diagnostic features.

Table 5. Top Diagnostic Features Contributing to Breast Cancer Classification

Rank	Feature	Importance Score
1	Bland Chromatin	0.231
2	Single Epithelial Cell Size	0.185
3	Normal Nucleoli	0.134
4	Uniformity of Cell Shape	0.126
5	Uniformity of Cell Size	0.123
6	Bare Nuclei	0.109
7	Marginal Adhesion	0.072
8	Clump Thickness	0.020
9	Mitoses	0.000

The Random Forest model's classification accuracy was 95.91%, indicating it performed quite well. In addition, the feature-importance results indicate that cell shape and nuclear characteristics are the primary contributors to breast cancer detection. Moreover, the machine learning outcomes closely align with the standard clinical knowledge platform, thereby greatly enhancing the reliability and interpretability of this diagnostic system.

5. Discussion and Interpretation of Results

Experimentally, we have shown that machine learning techniques can be highly effective at accurately classifying breast cancer using diagnostic features from the Wisconsin Breast Cancer Dataset. To compare the models, we used five different metrics, including the main one - accuracy - besides that, we also used precision, recall, F1-score and Area Under the ROC Curve (AUC). The reason for using accuracy is that misdiagnosis in the health field can have serious consequences, so it is very important to keep the number of such mistakes as low as possible. The focus of precision is to reduce the number of false positives, i.e., to minimise cases where a patient is considered sick when he/she is perfectly fine. Another parameter, recall, is a very important measure as well; knowing the exact number of malignant tumour cells is crucial to reducing false negatives, which is why treatments are often delayed, and patients' health status worsens, of all the models examined, Random Forest yielded the best results on both the training and test sets, with metrics of 95.84% accuracy, 95.31% precision, 95.12% recall, 95.21% F1-score, and 0.982 AUC. These figures correspond to an excellent model in both class separation and generalisation. The probable reason why Random Forest yields such good results is that, in the ensemble learning approach, it relies on the collective intelligence of several decision trees, which allows it to both learn complex nonlinear feature interactions and limit variance and overfitting. Decision Tree ranked second in classification performance, with 91.26% accuracy and 0.944 AUC. It extracted classification rules from the diagnostic features and presented a step-by-step explanation of the classification process. But the performance difference between training and testing was substantial, indicating overfitting. Logistic Regression and K-Nearest Neighbours yielded sufficiently good results, with accuracies of 88.42% and 90.37%, respectively. Naive Bayes was by far the worst method compared to the others; its main limitation is that it assumes features are independent, which is not the case for breast cancer diagnostic features with their complex interactions, so this assumption works to their disadvantage. On top of that, we corroborated our decision with a Random Forest model that performed best in the ROC curve analysis. Random Forest's highest AUC indicates that the model can distinguish malignant from benign cancers by considering multiple factors. It is highly likely that such a level of performance is among the most important factors in a clinical setting, since an accurate and early diagnosis of cancer is generally followed by effective treatment and improved patient survival. We also performed fivefold cross-validation to assess the robustness of the models. Results showed that the Random Forest was the most accurate classifier, with an average accuracy of 95.96% and a very low standard deviation of 0.43, indicating it performed well across all data splits. Model validation results are summarised in Table 6.

Table 6. Five-Fold Cross-Validation Results of Machine Learning Models

Model	Fold 1 (%)	Fold 2 (%)	Fold 3 (%)	Fold 4 (%)	Fold 5 (%)	Mean Accuracy (%)	Std. Dev.
Logistic Regression	87.72	89.47	88.60	87.72	88.60	88.42	0.64
K-Nearest Neighbors	89.47	91.23	90.35	89.47	91.23	90.35	0.82
Decision Tree	90.35	92.11	91.23	90.35	92.11	91.23	0.83
Naïve Bayes	85.96	87.72	86.84	86.84	87.72	87.02	0.68
Random Forest	95.61	96.49	95.61	96.49	95.61	95.96	0.43

Small variation across the folds indicates that the models have been well trained and can perform well on diverse data. The Random Forest classifier not only had the highest average accuracy but was also the most consistent, with the lowest standard deviation in accuracy among the classifiers used. Results like these are very encouraging: the classifiers (mainly Random Forest) show strong, stable performance and are the most effective decision-support tools.

5.1. Insights into Feature Importance and Medical Relevance

A random forest classifier was implemented for a feature-importance analysis to enhance model interpretability. The analysis indicated that the six leading features for a breast cancer diagnosis are Bland Chromatin, Single Epithelial Cell Size, Normal Nucleoli, Uniformity of Cell Shape, Uniformity of Cell Size, and Bare Nuclei. These features are tied to cell morphology, nuclear structure, and patterns of chromatin changes, which are major pathological features of cancer.

Bland Chromatin is the top diagnostic feature, with Single Epithelial Cell Size and Normal Nucleoli ranked close behind. Having the primary pathology features identified by the leading user align with clinical knowledge is not only a sign that the machine learning model is accurate but also that it has detected biologically meaningful features related to breast cancer progression. The concurrence between computational results and medical facts enhances the model's credibility and suggests its utility in clinical decision support. Feature importance analysis also uncovers biological facets prompting breast cancer diagnosis. By highlighting features that significantly contribute to classification decisions, the model not only bolsters predictive accuracy but also elucidates the key diagnostic features for doctors. More than anything else, in healthcare, where transparency and explainability are crucial for clinical trust and acceptance, interpretability might prove to be a compelling feature.

5.2. Cross-Validation, Hyperparameter Optimisation, and Ethical Considerations

The researchers conducted a thorough model evaluation and validation process, which showed that the proposed structure was appropriate. The dependability of the proposed method was established after an in-depth literature review and model validation. To ensure that the built forecasting models still performed well on new and different data sets and to minimise the risk of model overfitting, the authors used cross-validation. Besides generalisation, cross-validation also allowed hyperparameter tuning and better classification accuracy, contributing to an even stronger model. Most of the time, Random Forest consistently achieved the highest cross-validation accuracy. That means it is a very reliable model and is well-suited for real-world applications. From a moral standpoint, healthcare machine learning tools should balance accuracy, equity, and interpretability well. Using simple-to-interpret models like Decision Trees, along with a Random Forest model to assess feature importance, reflects the main idea of explainable AI (XAI) and will help physicians grasp the models' reasoning processes. Clearly, explaining things is the first step toward responsible AI use in medicine.

In fact, this work perfectly illustrates how the fields of data science and medicine can benefit from each other's knowledge. Bringing together machine learning and clinical skills can produce diagnostic models that are not only highly accurate but also understandable and medically meaningful. By bringing together different forms of intelligence, physicians will gain tools that aid their decision-making, and, as a result, cancer will be diagnosed earlier, treatments will be customised to patients, and overall, there will be better outcomes for patients in the healthcare system.

6. Conclusion and Recommendations

This article is about machine learning (ML) for breast cancer classification based on features from the Wisconsin Breast Cancer Dataset (WBCD). Five ML methods (Logistic Regression, K-Nearest Neighbours, Decision Tree, Naive Bayes, and Random Forest) are compared using various performance measures, including Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC). The highest performance among the proposed approaches in the evaluation was achieved by Random Forest, with 95.84% accuracy, 95.31% precision, 95.12% recall, 95.21% F1 score, and an AUC of 0.982. Also, through five-fold cross-validation, the method was verified to be stable and reliable, with an average accuracy of 95.96% and a very small standard deviation of 0.43. Ensemble methods in breast cancer diagnosis have been very good at identifying diagnostic features by detecting complex nonlinear relationships among attributes, while also limiting overfitting. Besides that, the feature importance step has made the system easier to explain by indicating the main diagnostic features as Bland Chromatin, Single Epithelial Cell Size, Normal Nucleoli, Uniformity of

Cell Shape, Uniformity of Cell Size, and Bare Nuclei. The strong correlation between the features used for prediction and clinically established facts supports the conclusion that ML models can uncover breast cancer patterns with biological implications. The partnership of artificial intelligence and physician experts enhances the reliability of diagnostic outcomes and spurs the use of diagnostics in healthcare. This way, machine-learning-based diagnostic systems can be considered decision-support tools for medical personnel based on these results. Besides high predictive accuracy, explanatory power, and thorough testing, the major characteristics of data-driven methods are the ability to detect breast cancer at an early stage, a lower risk of diagnostic error, and the provision of rapid medical intervention. So, the article emphasised the need for strong collaboration between data scientists and medical doctors to develop dependable, clinically relevant healthcare solutions.

Future Recommendations

Further investigations into this topic should focus on collecting data from larger and more diverse groups across several healthcare centres. This step is critical to enhancing the model's ability to be generalised to other settings. The combination of sophisticated ensemble learning methods with deep learning architectures could yield more accurate classification and a more stable model.

Perhaps implementing explainable artificial intelligence (XAI) methods such as SHAP, LIME, and Grad-CAM could help clinicians better understand the model, especially if each prediction is shown to them in detail. In this way, their trust in the model tends to be increased. Also, the combination of images with clinical, genetic, and histological data leads to multimodal diagnostic devices that thoroughly describe breast cancer.

The implementation of privacy-friendly machine learning methods, such as federated learning, can facilitate collaborative medical research without revealing patients' identities. Yet, it is still necessary for the researchers to conduct a clinical trial with real patients to determine the operational effectiveness of the proposed structure in day-to-day healthcare.

The existing machine learning system has strong potential to transform breast cancer diagnosis by enabling the development of highly accurate, interpretable, and intelligent clinical decision-support systems, a major advancement towards personalised healthcare.

All the Declarations and Statements

Author Contributions Statement

T. Haripriya: Conceptualization, Methodology, Data Curation, Software, Formal Analysis, Investigation, Visualization, Validation, Writing – Original Draft Preparation.
M. V. Ramana Murthy: Supervision, Conceptualization, Formal Analysis, Validation, Writing – Review & Editing.
Ch. Vasavi: Methodology, Investigation, Data Interpretation, Validation, Writing – Review & Editing.
Swathi Gowroju: Software, Machine Learning Implementation, Experimental Evaluation, Visualization, Writing – Review & Editing.
G. Srinivas: Formal Analysis, Validation, Investigation, Writing – Review & Editing.
Devineni Gireesh Kumar: Project Administration, Supervision, Resources, Validation, Writing – Review & Editing.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research received no external funding.

Data Availability Statement

The dataset used in this study is publicly available from the UCI Machine Learning Repository.
Dataset: Wisconsin Diagnostic Breast Cancer (WDBC) Dataset
Dataset Source: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
The processed data, trained models, and experimental results supporting the findings of this study are available from the corresponding author upon reasonable request.

Ethical Declarations

This study uses a publicly available benchmark dataset and does not involve human participants, human tissue, personal patient information, or live animals. Therefore, ethical approval and informed consent were not required.

Acknowledgments
The authors sincerely thank Sreyas Institute of Engineering and Technology, Hyderabad, India, for providing the research facilities and computational resources necessary to conduct this study. The authors also express their sincere appreciation to the UCI Machine Learning Repository for making the Wisconsin Diagnostic Breast Cancer Dataset publicly available. Finally, the authors thank the anonymous reviewers for their valuable comments and constructive suggestions, which significantly improved the quality of this manuscript.

Declaration of Generative AI in Scholarly Writing
During the preparation of this manuscript, the authors used generative artificial intelligence (AI)-assisted tools solely to improve the language, grammar, readability, and organization of the manuscript. All scientific ideas, methodology, statistical analyses, machine learning implementations, interpretations of results, and conclusions were conceived, verified, and approved by the authors. The authors carefully reviewed and edited all AI-assisted content and take full responsibility for the originality, accuracy, integrity, and scientific validity of the published work.

Abbreviations
The following abbreviations are used in this manuscript:

Abbreviation	Full Form
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
XAI	Explainable Artificial Intelligence
PCA	Principal Component Analysis
ESA	Exploratory Statistical Analysis
WDBC	Wisconsin Diagnostic Breast Cancer Dataset
UCI	University of California Irvine
LR	Logistic Regression
KNN	K-Nearest Neighbours
DT	Decision Tree
NB	Naïve Bayes
RF	Random Forest
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
FNA	Fine Needle Aspiration
CNN	Convolutional Neural Network
SVM	Support Vector Machine

Appendix A/B/C…, with appendix title
N/A

References

[1] N. Mustafee, A. Harper, and B. S. Onggo, "Hybrid Modelling and Simulation (M&S): Driving Innovation in the Theory and Practice of M&S," in 2020 Winter Simulation Conference (WSC), Orlando, FL, USA, 2020, pp. 3140-3151, doi: https://doi.org/10.1109/WSC48552.2020.9383892.

[2] A. S. Panayides et al., "AI in Medical Imaging Informatics: Current Challenges and Future Directions," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 7, pp. 1837-1857, July 2020, doi: https://doi.org/10.1109/JBHI.2020.2991043.

[3] M. T. Nyo, F. Mebarek-Oudina, S. S. Hlaing et al., "Otsu's thresholding technique for MRI image brain tumor segmentation," in Multimedia Tools and Applications, vol. 81, pp. 43837–43849, 2022, doi: https://doi.org/10.1007/s11042-022-13215-1.

[4] N. Noreen, S. Palaniappan, A. Qayyum, I. Ahmad, M. Imran, and M. Shoaib, "A Deep Learning Model Based on Concatenation Approach for the Diagnosis of Brain Tumor," in IEEE Access, vol. 8, pp. 55135-55144, 2020, doi: https://doi.org/10.1109/ACCESS.2020.2978629.

[5] M. H. Khammash and J. Stelling, "Systems and Synthetic Biology [Scanning the Issue]," in Proceedings of the IEEE, vol. 110, no. 5, pp. 518-522, May 2022, doi: https://doi.org/10.1109/JPROC.2022.3173798.

[6] D. Bi, A. Almpanis, A. Noel, Y. Deng, and R. Schober, "A Survey of Molecular Communication in Cell Biology: Establishing

a New Hierarchy for Interdisciplinary Applications," in IEEE Communications Surveys & Tutorials, vol. 23, no. 3, pp. 1494-1545, 3rd Quart., 2021, doi: https://doi.org/10.1109/COMST.2021.3066117.

[7] R. J. Chen, M. Y. Lu, J. Wang, D. F. K. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood, "Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis," in IEEE Transactions on Medical Imaging, vol. 41, no. 4, pp. 757-770, Apr. 2022, doi: https://doi.org/10.1109/TMI.2020.3021387.

[8] Z. Fei, Y. Ryeznik, O. Sverdlov, C. W. Tan, and W. K. Wong, "An Overview of Healthcare Data Analytics with Applications to the COVID-19 Pandemic," in IEEE Transactions on Big Data, vol. 8, no. 5, pp. 1163-1178, Oct. 2022, doi: https://doi.org/10.1109/TBDATA.2021.3103458.

[9] Y. Censor, K. E. Schubert, and R. W. Schulte, "Developments in Mathematical Algorithms and Computational Tools for Proton CT and Particle Therapy Treatment Planning," in IEEE Transactions on Radiation and Plasma Medical Sciences, vol. 6, no. 3, pp. 313-324, Mar. 2022, doi: https://doi.org/10.48550/arXiv.2108.09459.

[10] C. Venkatesan, D. Balamurugan, T. Thamaraimanalan, and M. Ramkumar, "Efficient Machine Learning Technique for Tumor Classification Based on Gene Expression Data," in Proceedings of the 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2022, pp. 1982-1986, doi: https://doi.org/10.1109/ICACCS54159.2022.9785294.

[11] L. Chazette, W. Brunotte, and T. Speith, "Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue," in Proceedings of the 2021 IEEE 29th International Requirements Engineering Conference (RE), Notre Dame, IN, USA, 2021, pp. 197-208, doi: https://doi.org/10.1109/RE51729.2021.00025.

[12] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and Robust Machine Learning for Healthcare: A Survey," in IEEE Reviews in Biomedical Engineering, vol. 14, pp. 156-180, 2021, doi: https://doi.org/10.1109/RBME.2020.3013489.

[13] S. Bharati, M. R. H. Mondal, and P. Podder, "A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When?," in IEEE Transactions on Artificial Intelligence, vol. 5, no. 4, pp. 1429-1442, Apr. 2024, doi: https://doi.org/10.1109/TAI.2023.3266418.

[14] M. F. Ak, "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications," in Healthcare, vol. 8, no. 2, p. 111, 2020, doi: https://doi.org/10.3390/healthcare8020111.

[15] M. Cekikj, M. J. Özdemir, S. Kalajdzhiski, O. Özcan, and O. U. Sezerman, "Understanding the Role of the Microbiome in Cancer Diagnostics and Therapeutics by Creating and Utilizing ML Models," in Applied Sciences, vol. 12, no. 9, p. 4094, 2022, doi: https://doi.org/10.3390/app12094094.

[16] Garberis, V. Gaury, C. Saillard et al., "Deep learning assessment of metastatic relapse risk from digitized breast cancer histological slides," in Nature Communications, vol. 16, p. 5876, 2025, doi: https://doi.org/10.1038/s41467-025-60824-z.

[17] N. A. Ghafoor and A. Yildiz, "Targeting MDM2–p53 Axis through Drug Repurposing for Cancer Therapy: A Multidisciplinary Approach," in ACS Omega, vol. 8, no. 38, pp. 34583-34596, Sep. 2023, doi: https://doi.org/10.1021/acsomega.3c03471.

[18] X. Guan, Y. Du, R. Ma et al., "Construction of the XGBoost model for early lung cancer prediction based on metabolic indices," in BMC Medical Informatics and Decision Making, vol. 23, p. 107, 2023, doi: https://doi.org/10.1186/s12911-023-02171-x.

[19] A. J. Preto, P. Matos-Filipe, J. Mourão, and I. S. Moreira, "SYNPRED: prediction of drug combination effects in cancer using different synergy metrics and ensemble learning," in GigaScience, vol. 11, p. giac087, 2022, doi: https://doi.org/10.1093/gigascience/giac087.

[20] J. Vrdoljak, Z. Boban, D. Barić, D. Šegvić, M. Kumrić, M. Avirović, M. Perić Balja, M. M. Periša, Č. Tomasović, S. Tomić, E. Vrdoljak, and J. Božić, "Applying Explainable Machine Learning Models for Detection of Breast Cancer Lymph Node Metastasis in Patients Eligible for Neoadjuvant Treatment," in Cancers, vol. 15, no. 3, p. 634, Jan. 2023, doi: https://doi.org/10.3390/cancers15030634.

[21] A. Prayongrat, N. Srimaneekarn, K. Thonglert et al., "Machine learning-based normal tissue complication probability model for predicting albumin-bilirubin (ALBI) grade increase in hepatocellular carcinoma patients," in Radiation Oncology, vol. 17, p. 202, 2022, doi: https://doi.org/10.1186/s13014-022-02138-8.

[22] A. Janssen, F. C. Bennis, and R. A. A. Mathôt, "Adoption of Machine Learning in Pharmacometrics: An Overview of Recent Implementations and Their Considerations," in Pharmaceutics, vol. 14, no. 9, p. 1814, 2022, doi: https://doi.org/10.3390/pharmaceutics14091814.

Authors' Profiles

Dr. T. Hari Priya is a Professor in the Department of Mathematics at Sreyas Institute of Engineering and Technology. She holds a Ph.D. in Mathematics and has over 25 years of teaching experience in higher education. Her academic interests include mathematical sciences, research, and higher education. She has published several quality research papers in reputed national and international journals and has contributed significantly to teaching, student mentoring, and academic development. Deep learning models, explainable AI, and advanced machine learning techniques for real-world applications.

Dr. M. V. Ramana Murthy is a Professor of Mathematics with over 39 years of teaching and research experience. He holds a Ph.D. in Mathematics with specialization in interdisciplinary research. He has published more than 260 research papers in reputed national and international journals, holds 7 patents, and has authored research monographs. As a distinguished research supervisor, he has successfully guided and awarded 55 PhD scholars in the Engineering and Science disciplines.
Dr. Ramana Murthy is an active member of several professional bodies and continues to contribute to academic excellence through his teaching, research, innovation, and scholarly leadership.

Dr. Ch. Vasavi is a Professor in the Department of Mathematics at Sreyas Institute of Engineering and Technology. She holds a Ph.D. in Mathematics and has over 25 years of teaching and research experience. Her research interests include numerical methods, heat transfer problems, and machine learning. She has published several high-quality research papers in reputable national and international journals and has made significant contributions to teaching, curriculum development, and research activities.

Dr. Swathi Gowroju is the Head of the Department of Computer Science and Engineering (Artificial Intelligence & Machine Learning) at Sreyas Institute of Engineering and Technology, with over 16 years of teaching and research experience. Her research interests include Artificial Intelligence, Machine Learning, Deep Learning, Computer Vision, Medical Image Analysis, Cybersecurity, and Data Analytics. She has published numerous research articles in SCIE- and Scopus-indexed journals, at international conferences, and in book chapters. Her contributions span healthcare analytics, intelligent systems, and cybersecurity, and she actively promotes innovation, interdisciplinary research, curriculum development, and academic excellence.

Dr. G. Srinivas is a Professor of Mathematics with over 20 years of teaching and research experience in reputed engineering institutions. He holds a Ph.D. in Mathematics from Sri Krishnadevaraya University. He has published more than 60 research papers in reputed Scopus-indexed and international journals and has authored several academic books.
Dr Srinivas is a recognized research supervisor and reviewer for leading international publishers, including Elsevier and Springer Nature. He has successfully guided and awarded six Ph.D. scholars under his supervision and continues to contribute to academic excellence through his teaching, research, and scholarly activities.

Dr. Gireesh Kumar Devineni is an Associate Professor in the Department of Computer Science and Engineering (Artificial Intelligence & Machine Learning) at Sreyas Institute of Engineering and Technology, Hyderabad, India. He received his Ph.D. in Electrical Engineering from Lovely Professional University, Punjab, in 2022, and is currently pursuing a second Ph.D. in Computer Science and Engineering with research focused on Explainable Artificial Intelligence (XAI), Multimodal Deep Learning, and Large Language Models.
He has over 15 years of teaching experience and 7 years of research experience. His research interests include Artificial Intelligence, Machine Learning, Deep Learning, Medical Image Analysis, Computer Vision, Data Science, Natural Language Processing, Renewable Energy Systems, Electric Vehicles, and Smart Technologies. He has published more than 60 research articles in reputed SCIE- and SCOPUS-indexed journals and conferences, authored books and book chapters, and holds several patents. He is an active reviewer and researcher and is a member of IEEE, IAENG, and SAE International.

How to cite this paper

T. Haripriya, M.V. Ramana Murthy, Ch. Vasavi, Swathi Gowroju, G. Srinivas, Devineni Gireesh Kumar, "An Interpretable Machine Learning Framework for Breast Cancer Diagnosis Using Statistical Feature Analysis and Ensemble Classification", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.4, pp.228-252, 2026. DOI:10.5815/ijem.2026.04.16