Applying Machine Learning Algorithms for Early Prediction of Breast Cancer

B. Srinivas
Bharath Institute of Higher Education and Research, Research Scholar, School of Computing, Department of CSE
Chennai, Tamilnadu, India

M. Sriram
Bharath Institute of Higher Education and Research Associate Professor, School of Computing, Department of IT
Chennai, Tamilnadu, India

V. Ganesan
Bharath Institute of Higher Education and Research Associate Professor, School of Electrical, Department of ECE
Chennai, Tamilnadu, India

ABSTRACT
Breast cancer is a devastating illness impacting millions of women globally. Early prediction is vital for successful treatment and improved survival rates. Machine learning algorithms have come out as one of the most efficient tools for classifying and diagnosing breast cancer, presenting promising solutions for early prediction and enhanced patient outcomes. The study utilised various machine learning classifiers to categorise breast cancer data: MLP (Multi-layer Perceptron classifier), Support Vector Machines (SVMs), Random Forests (RFs), Logistic Regression (LR), K-Nearest Neighbor (KNN), and Decision Trees (DTs) Classifier. Each classifier underwent training and evaluation using the Wisconsin Breast Cancer Dataset, a widely utilised benchmark dataset with 569 instances featuring characteristics extracted from fine needle aspirates of breast mass lesions. The effectiveness of each classifier was assessed employing various metrics, including accuracy, F1-score, specificity, and sensitivity. Experimental results revealed that MLP (Multi-layer Perceptron) displayed superior performance among the tested classifiers, achieving an accuracy of 95.07%, an F1-score of 93.33%, a specificity of 94.44%, and a sensitivity of 99.22%. SVM Classifier closely followed, attaining accuracies of 94.37% and 97.35%, respectively. The findings highlight the potential of machine learning algorithms, especially Multi-Layer Perceptron, which can accurately classify breast cancer datasets and predict breast cancer. The high accuracy and sensitivity achieved by Multi-Layer Perceptron suggest its suitability for early cancer prediction, enabling prompt intervention and improved treatment outcomes.

Keywords
Machine learning, breast cancer, diagnosis, early Prediction, treatment, survival rates.

1. INTRODUCTION
Breast cancer poses a significant global health issue. The worldwide cancer burden is a cause for concern, as one in five individuals is expected to receive a cancer diagnosis in their lifetime. Future projections suggest a troubling trend, with anticipated cancer diagnoses in 2040 set to increase by nearly 50% compared to 2023. According to World Health Organization in early 2020, approximately 2.3 million women were diagnosed with breast cancer with 690,000 deaths reported globally. At the start of 2021, there were approximately 7.8 million women who had been diagnosed with breast cancer in the last 5 years, making breast cancer as one of the world's most widespread cancer. This underscores the crucial need for heightened attention to both cancer prevention and treatment [1]. Machine learning can enable the early prediction of breast cancer thereby improving not only the cancer diagnosis but also the survival rate of cancer patients [2]. Machine learning tools can process and analyse large amount of mammogram datasets thereby extracting even the minute change in image patterns and identify abnormal changes or growths enabling the prediction of breast cancer or likelihood of breast cancer based upon various risk factors [4-6]. Recent research shows a very high capability of machine learning algorithms in terms of breast cancer prediction, achieving high accuracy by surpassing tradition diagnosis methods and thereby revolutionizing the study of oncology that is faster, more accurate and non-invasive [7-8]. Our research paper focuses on relative analysis of five classifiers namely; Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Decision Tree (C4.5), and K-Nearest Neighbors (KNN). Our primary objective is to exploit various machine learning algorithms to forecast and predict breast cancer using mammogram datasets and to identify the efficient classifier through performance evaluations based on metrics namely; accuracy, precision, sensitivity and confusion matrix.

2. WORK METHODOLOGY
The central objective of our study centres on pinpointing the optimal and prognostic algorithm for identifying breast cancer. To accomplish this, we utilized machine learning models viz; Support Vector Machine (SVM), Random Forests (RFs), Logistic Regression (LR), Decision Tree (C4.5), and K-Nearest Neighbors (KNN) on the Wisconsin Breast Cancer dataset. The assessment of outcomes will ascertain the model that produces superior accuracy. For our research we have used "Breast Cancer Wisconsin (Diagnostic) WDBC dataset", available on Kaggle. The data consists of "Digitized images of fine needle aspirates (FNA) of breast mass lesions", having features namely; Perimeter, Texture, Area, Smoothness, Compactness, Symmetry, Concavity, Fractal dimension, and Radius. Each feature has three values namely; mean, standard error, and worst. The total number of instances are 569; wherein, 212 instances have been labelled as malignant (1) and 357 instances have been labelled as benign (0).

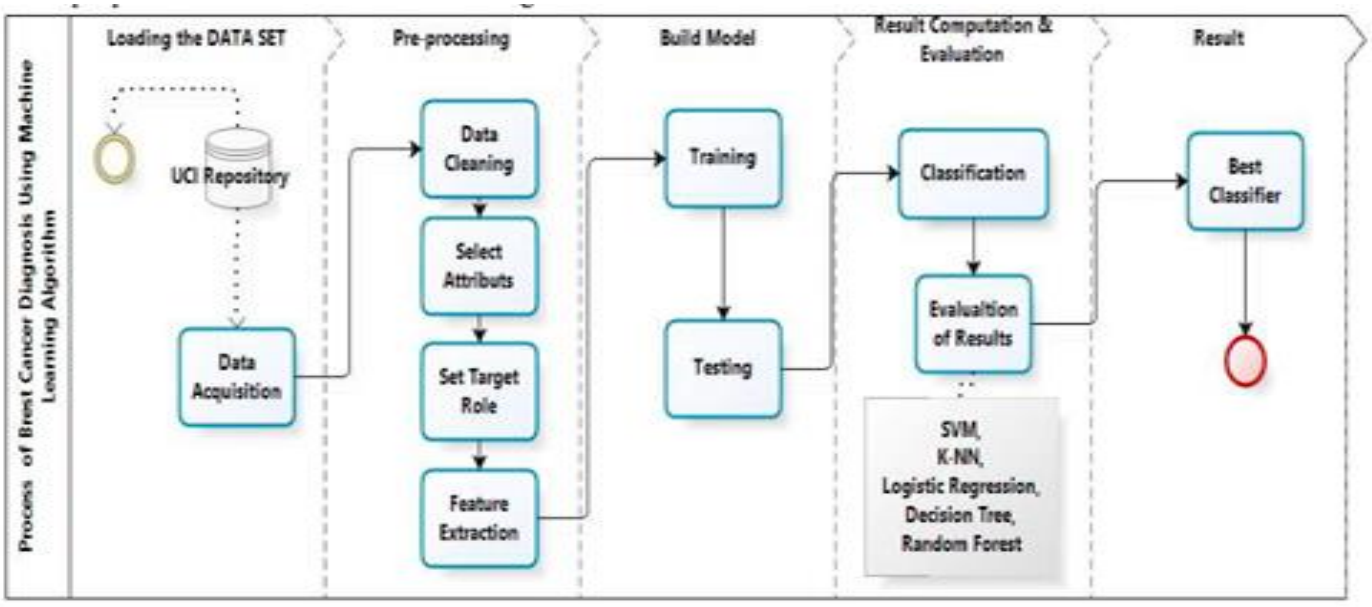

Fig 1: Process Flow Diagram

3. EVALUATING MODEL PERFORMANCE

Following parameters were used to evaluate the performance of our models:

3.1. Precision (P): Precision denotes the measure of exactness or specificity in correctly identifying relevant instances among those predicted as positive. It assesses the precision of positive predictions, determined by dividing the instances of true positives by the total of true positives and false positives.

$$P = \frac{TP}{\text{TP} + \text{FP}} \text{ ----------(1)}$$

3.2. Recall (R): Recall denotes the capacity of a model to recognize all pertinent occurrences within a dataset. Its emphasis lies in capturing authentic positives, achieved by dividing the count of genuine positive instances by the total of true positives and false negatives.

$$R = \frac{TP}{\text{TP} + \text{FN}} \text{ ----------(2)}$$

3.3. F1 Score: The F-1 score serves as a composite metric to furnish a complete evaluation of a model's effectiveness. Calculated through the harmonic mean of Precision and Recall, it presents a succinct portrayal of the model's capability in managing both false positives and false negatives.

$$F1 = \frac{2 \times P \times R}{\text{P+R}} \text{ ----------(3)}$$

3.4. Accuracy (A): Accuracy measures the overall correctness of predictions made by a machine learning model. Its computation involves the division of the cumulative count of true positives and true negatives by the total instances.

$$P = \frac{TP+TN}{\text{TP+TN+ FP+FN}} \text{ -----(4)}$$

4. EXPERIMENTAL SETUP

The implementation of this study was conducted in Jupyter Notebook using the Python language. The outlined procedure encompasses key steps aimed at assisting data analysts or physicians in real-time breast cancer prediction. Initially, relevant Python libraries were imported, followed by the execution of pre-processing steps to eliminate missing values. Subsequently, various performance evaluation metrics were tested to assess the models' effectiveness.

For the creation of training sets, an 80%-20% split was implemented, and output results were defined. Multiple machine learning models, namely MLP (Multi-layer Perceptron classifier), Support Vector Machines (SVMs), Random Forests, Logistic Regression, K-Nearest Neighbors (KNN), and Decision Trees Classifier, were then executed. Cross-validation metrics were utilized to showcase the results. Following the definition of variables and data splitting, two methods were applied to determine the best hyperparameters: Grid Search CV and Recursive Feature Elimination (RFE). Confusion matrices, ROC curves, learning curves, and cross-validation metrics were plotted for both methods. In the third set of prediction models, Grid Search CV was employed to identify optimal hyperparameters for MLP, SVMs, Random Forests, Logistic Regression, KNN, and Decision Trees Classifier. Subsequently, confusion matrices and cross-validation metrics were presented. The final model, denoted as EC, incorporated MLP and SVMs with the voting classifier. Execution steps mirrored the previous methodologies, and results were elucidated through confusion matrices, learning curves, and cross-validation metrics.

5. RESULTS AND DISCUSSION

5.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a vital initial step in data science. It uncovers patterns and anomalies, using summary stats and visualizations to understand variable relationships. EDA addresses data quality, detecting outliers and handling missing values. In classification tasks, it examines target variables, guiding decisions on class distributions. Beyond exploration, EDA sparks hypotheses, informing pre-processing decisions and model selection, serving as a compass for the entire data science project.

5.2 Diagnosis (Target)

Diagnosis classification (M: Malignant, B: Benign)

Higher feature values correlate with malignancy, prompting consideration of data discretization. However,

the decision requires balancing granularity and information loss. Striking the right balance is crucial, considering the observed association and broader dataset context for the classification task

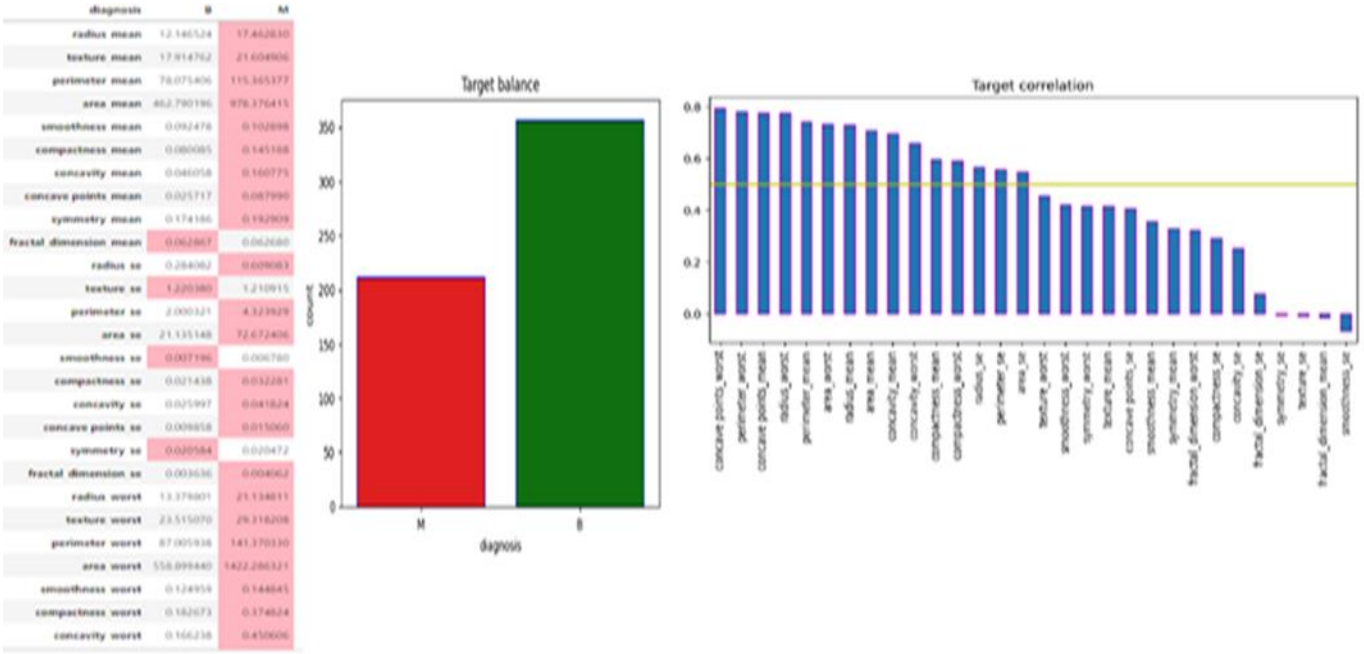

Fig 2: Wisconsin Breast Cancer Diagnostic Datasets

5.3. Correlated Predictors

Overall analysis of all the predictors.

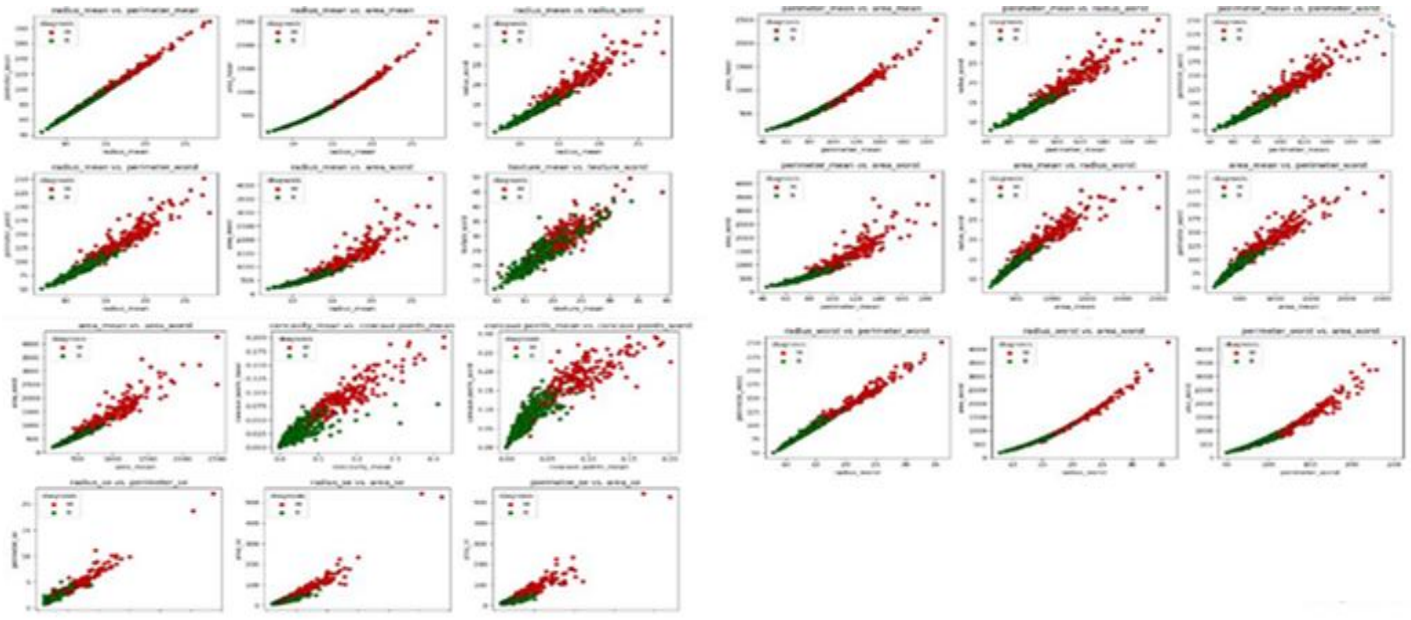

Fig 3: Multifaceted Examination of Predictors in Breast Cancer Assessment

5.4. Target correlations

An exploration of the features that exhibit higher correlations with our target variable entails both univariate and multivariate analysis. This two-fold approach aims to comprehensively understand the individual and collective impacts of these influential features on the target variable. In breast cancer analysis, "area_se," "perimeter_se," and "radius_se" measure variability in cell nucleus areas, perimeters, and radii. Higher values hint at irregularities, possibly indicating

malignancy. These features offer crucial insights into cell characteristics, aiding in the distinction between benign and malignant cases. "Perimeter_se" and "radius_se" show notable correlations with "area_se," reflecting their inherent similarity in measuring diagnostic characteristics. This intuitive correlation stems from the shared attributes of these parameters in capturing variations in perimeter and radius measurements, aligning with the diagnostic context of "area_se.".

In breast cancer analysis, "area_mean," "perimeter_mean," and "radius_mean" reflect cell nucleus characteristics, providing insights into size, shape, and boundary. Larger values in these features may suggest abnormal cell growth patterns, aiding in the identification of potentially malignant cases in diagnosis.

.

	diagnosis
area_se	0.548236
perimeter_se	0.556141
radius_se	0.567134
compactness_worst	0.590998
compactness_mean	0.596534
concavity_worst	0.659610
concavity_mean	0.696360
area_mean	0.708984
radius_mean	0.730029
area_worst	0.733825
perimeter_mean	0.742636
radius_worst	0.776454
concave points_mean	0.776614
perimeter_worst	0.782914
concave points_worst	0.793566

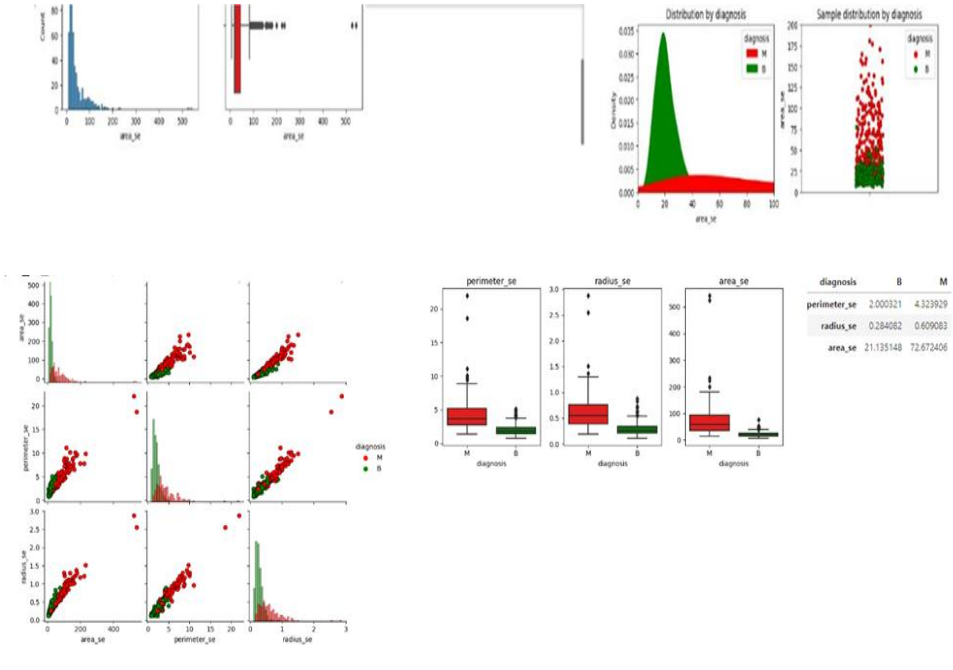

Fig 4: Integrated Analysis of Area, Perimeter, and Radius (SE) in Breast Cancer Assessment

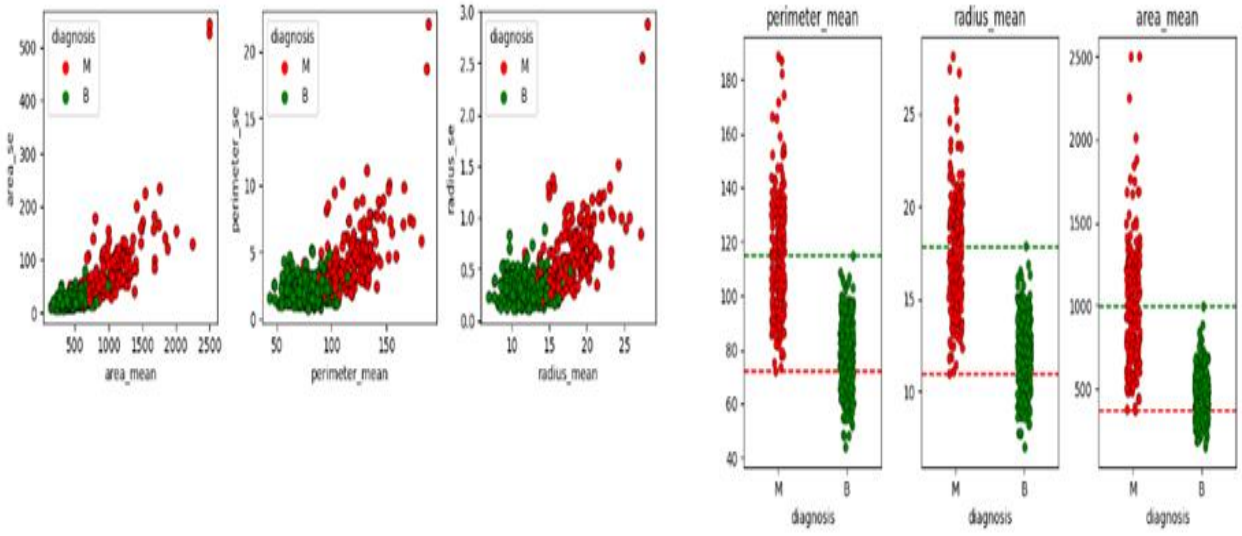

Fig 5: Multivariate Analysis of Area, Perimeter, and Radius (Mean) in Breast Cancer Assessment

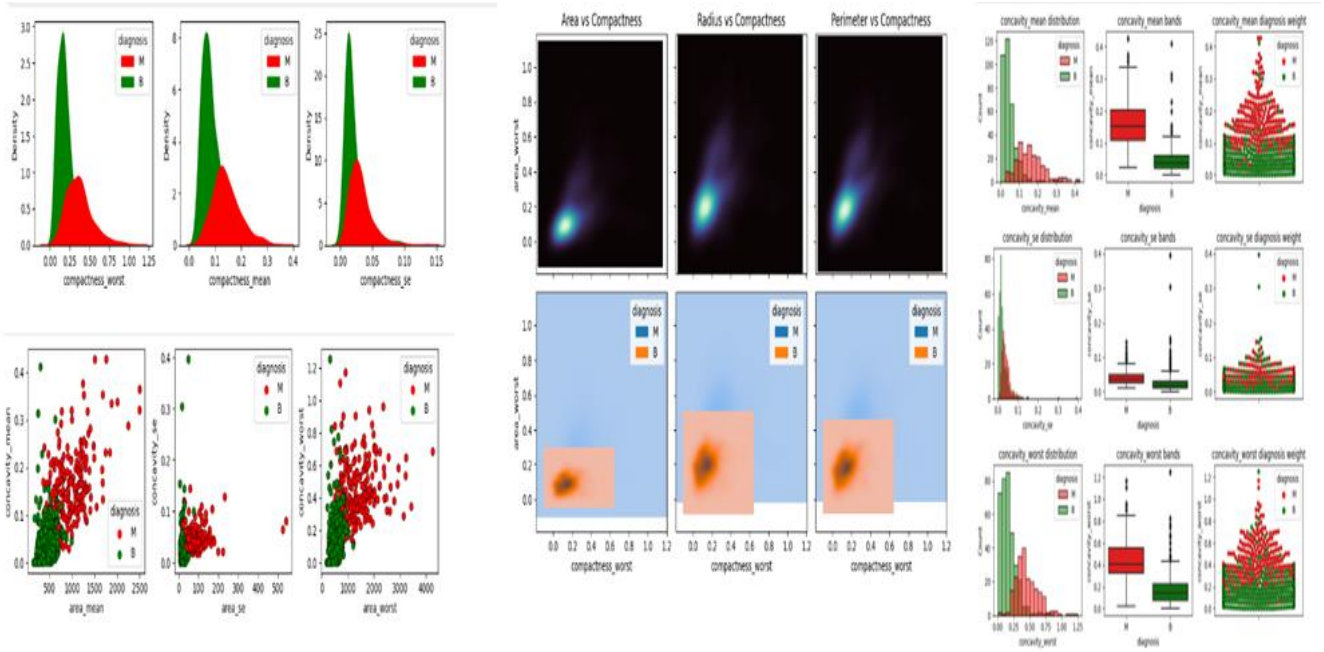

Fig 6: Comprehensive Analysis of Compactness in Breast Cancer Prediction

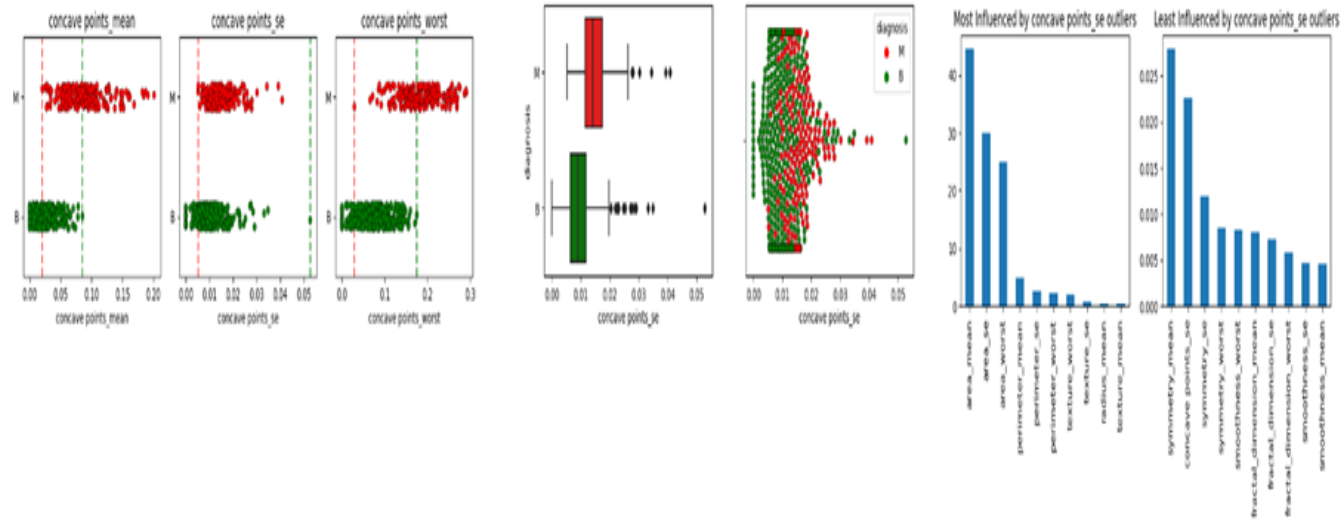

Fig 7: Comprehensive Overview of Concave Points in Breast Cancer Analysis

In breast cancer analysis, "concave points_mean," "concave points_se," and "concave points_worst" offer insights into concave points within cell nuclei. Elevated values indicate intricacy, variability, and complexity, aiding in distinguishing between benign and malignant cases and enhancing diagnostic precision. Combining concavity with other features, notably area, reveals a consistent and potent pattern, emphasizing the predictive power of concavity. The trend in concave points measurements suggests discrimination between benign and malignant diagnoses, with outliers, especially in standard error, requiring careful inspection for data quality and analysis reliability.

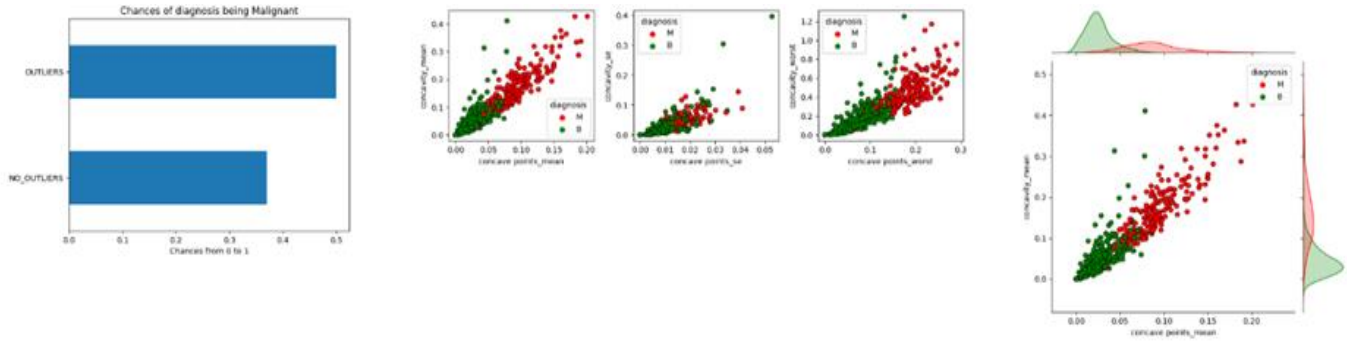

Fig 8: AUC comparison of malignant and benign diagnoses

In our comparison, becomes evident that the presence of outliers is associated with an increased likelihood of a malignant diagnosis. However, it's crucial to contextualize this observation. Rather than being detrimental, these outliers play a valuable role in strengthening our ability to distinguish between malignant and benign diagnoses. High feature values correlate with a higher chance of malignancy. A Principal Component Analysis (PCA) is suggested for deeper insights. Further exploration of correlated variables and additional analyses are also recommended.

5.5. Overall analysis

The plan is to aggregate “mean”, “se”, and “worst” features for a comprehensive view, and use PCA to uncover patterns and reduce dimensionality, enhancing diagnostic capabilities. Feature selection improves model performance by choosing the most relevant features. Metrics like correlation, mutual information, chi-squared test, F-test, recursive feature elimination, L1 regularization, information gain, and permutation importance are used. The choice of metric depends on the data and problem. The data indicates a consistent linear relationship between "mean" and "worst" variables, with higher values correlating to malignancy. Notably, the "SE" features exhibit distinct class separation. Before modeling, conducting PCA without considering labels, focusing on two components, confirms the correlation between higher values and malignancy. This analysis serves as a crucial bridge between data exploration and machine learning, fortifying the foundation for predictive models.

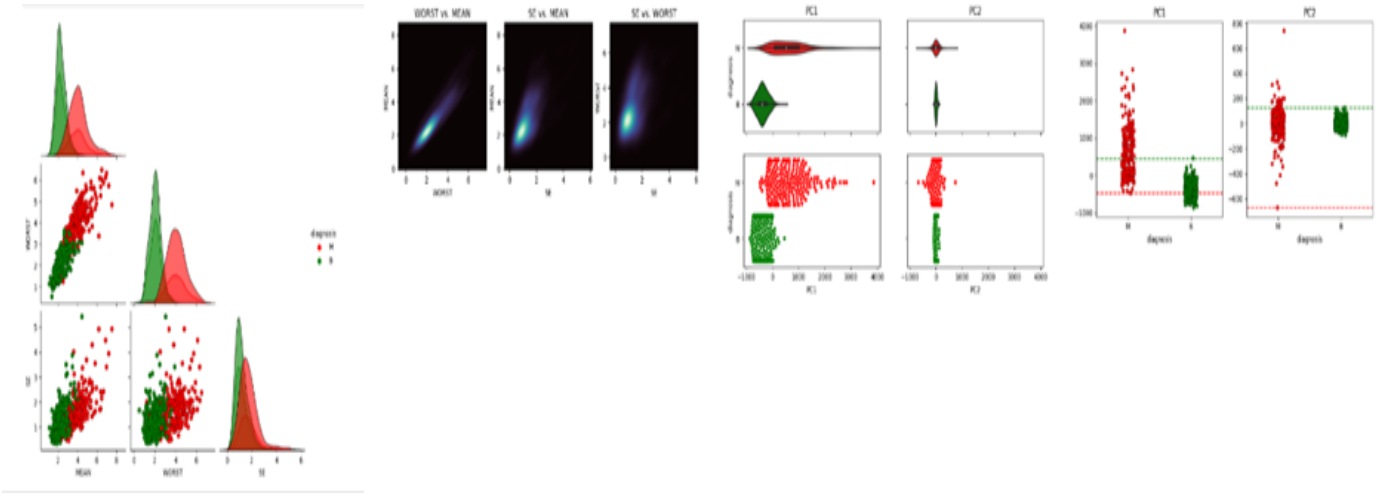

Fig 9: Combined Features and PCA Analysis for Enhanced Breast Cancer Diagnosis

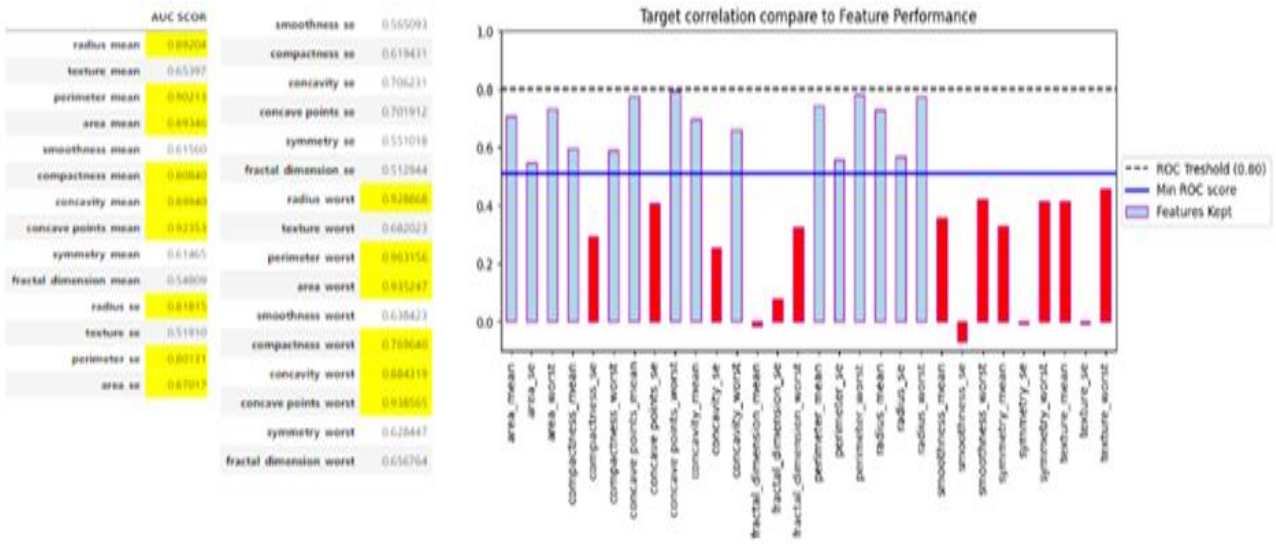

Fig 10: Target correction compare to feature performance

Model					
MLPClassifier	0.992227	0.950783	0.944405	0.924675	0.933349
SVM (RBF)	0.990253	0.943766	0.961144	0.886580	0.921450
Random Forest	0.990105	0.947274	0.939084	0.919913	0.928191
SVM (Linear)	0.987973	0.936717	0.937322	0.891342	0.912613
Logistic Regression	0.986013	0.940257	0.960668	0.877273	0.916100
K-Nearest Neighbors	0.976100	0.936717	0.928307	0.900866	0.913808
Decision Trees	0.918647	0.920927	0.886488	0.910390	0.895919

Fig 11: Performance Comparison of Machine Learning Models for Breast Cancer Prediction

Type	ROC AUC	Accuracy	Precision	Recall	F1-Score
Hard Voting	nan	0.938534	0.984899	0.848701	0.910224
Soft Voting	0.992737	0.950815	0.981169	0.886580	0.929975
MLPC	0.992227	0.950783	0.944405	0.924675	0.933349
SVM	0.991133	0.943797	0.984899	0.862771	0.918456

Fig 12: MLPC and SVM via Hard and Soft Voting for Breast Cancer Prediction

Type	ROC AUC	Accuracy	Precision	Recall	F1-Score
Soft Voting	0.992737	0.950815	0.981169	0.886580	0.929975
MLPC	0.992227	0.950783	0.944405	0.924675	0.933349
SVM	0.991133	0.943797	0.984899	0.862771	0.918456

Fig 13: MLPC and SVM via Soft Voting for Breast Cancer Prediction

Type	ROC AUC	Accuracy	Precision	Recall	F1-Score
Soft Voting	0.992737	0.950815	0.981169	0.886580	0.929975
MLPC	0.992227	0.950783	0.944405	0.924675	0.933349

Fig 14: MLP (Multi-layer Perceptron Classifier) via Soft Voting for Breast Cancer Prediction.

5.6. Final solution

We're deeply analyzing the Soft Voting classifier, examining prediction distribution, probabilities, and procedural steps. Understanding these aspects is crucial. Next, we'll train the classifier on the full dataset, fine-tune parameters, and optimize performance. This thorough approach highlights our commitment to delivering a precise solution for our specific machine learning problem.

5.7. Creating training and testing data

In machine learning, train-test splitting involves dividing data into a training set (70-80%) for model learning and a testing set (20-30%) for evaluation. The model learns patterns from the training set and is then assessed for performance on the testing set, ensuring a comprehensive evaluation of accuracy and robustness. This technique is fundamental in effective machine learning model development.

5.8. Confusion matrix

A vital instrument for assessing classification models is the confusion matrix, which concisely presents predictions in comparison to actual class designations. It encompasses accurate classifications such as true positives (TP) and true negatives (TN), as well as errors like false positives (FP) and false negatives (FN). Correct classifications involve TP and TN, while errors encompass FP and FN. This matrix facilitates the quantification of model performance by evaluating accuracy, precision, recall, and F1-score, offering a thorough insight into class distinctions.

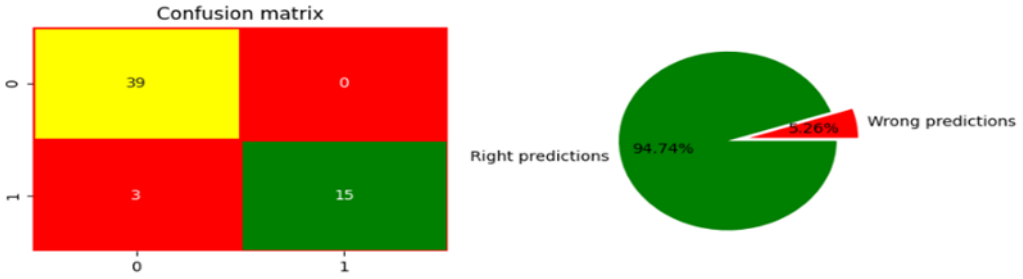

Fig 15: Confusion matrix of breast cancer predictions

5.9. Classification report

A classification report is a crucial tool for assessing a classification model's effectiveness in categorizing data into classes. It provides a detailed breakdown, helping identify issues like overfitting or underfitting. Particularly valuable in domains like medical diagnosis, spam detection, and sentiment analysis.

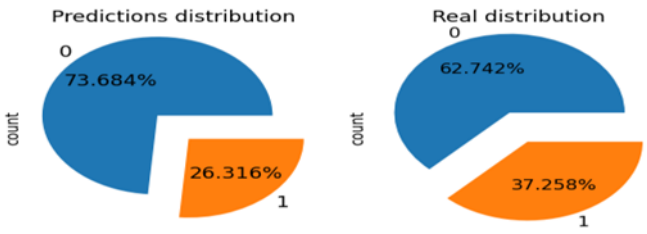

Fig 16: Classification Report for Breast Cancer Prediction

5.10. Wrong predictions in the original data

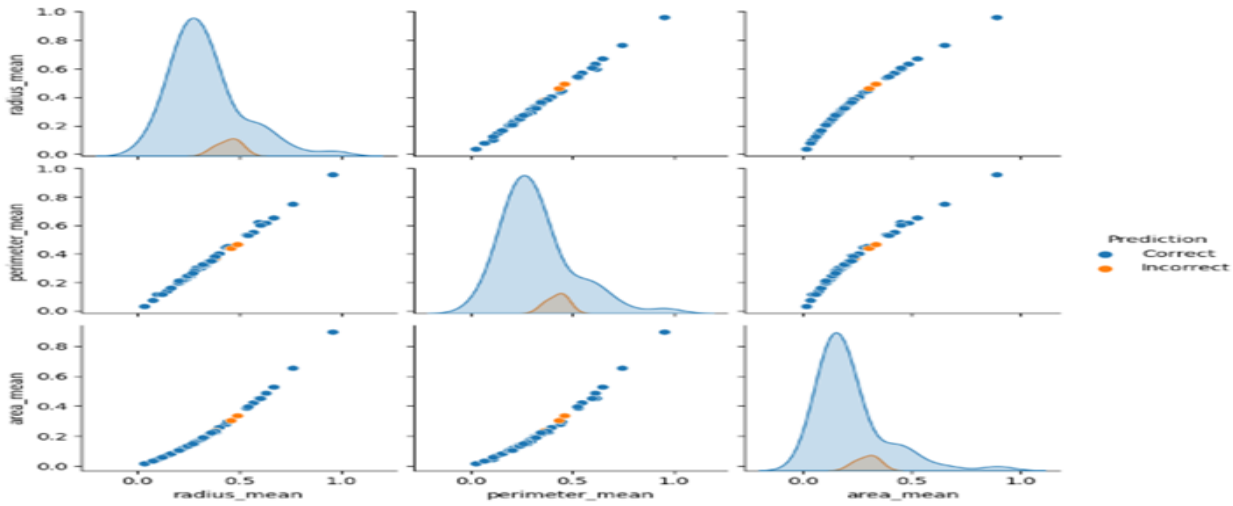

Fig 17: Visualization of Misclassified Data Points in Breast Cancer Prediction

5.11. Probabilities distribution

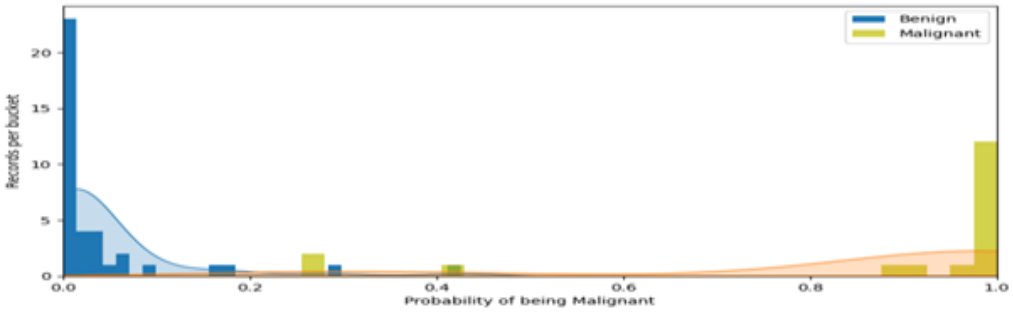

Fig 18: Probability Distribution for Breast Cancer Prediction

The solution involves a unified classifier, consolidating all steps from data pre-processing to model training using the identified winning algorithm. This streamlined approach maximizes the dataset's potential, enhancing predictive capabilities and facilitating easy deployment in real-world applications.

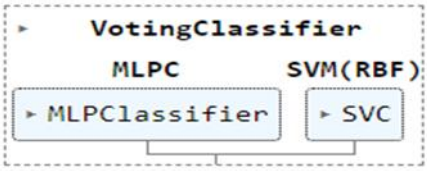

Fig 19: Integrated Workflow for Enhanced Breast Cancer Prediction

6. CONCLUSIONS

Wisconsin Breast Cancer datasets (WBCD) were used and studied by employing five principal algorithms: MLP (Multi-layer Perceptron classifier), Support Vector Machines (SVMs), Random Forests, Logistic Regression, K-Nearest Neighbors (KNN), and Decision Trees Classifier. Subsequently, we computed, compared, and evaluated diverse outcomes based on metrics such as confusion matrix, accuracy, sensitivity, precision, and AUC. This comprehensive analysis aimed to discern the most precise and reliable machine learning algorithm, striving to attain the highest accuracy. Following a meticulous comparison, the Multi-Layer Perceptron classifier emerged as the most efficient, achieving a remarkable accuracy of 95.07%, precision of 94.44%, and an AUC of 99.27%, surpassing the performance of other algorithms. In conclusion, the Multi-Layer Perceptron showcased notable efficiency in Breast Cancer prediction and diagnosis, establishing itself as the top performer in terms of accuracy and precision. It is crucial to acknowledge that the results obtained pertain solely to the WBCD database, posing a limitation to our study. Consequently, future research should apply these algorithms and methodologies to diverse databases to validate the results

7. REFERENCES

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. .

[2] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[3] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems

[4] Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[5] Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[6] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[7] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[8] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[9] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender