

Machine Learning-Based Prediction of Breast Cancer

Maab A. Srour^{1,*}, Mohammed M. Dweib², and Ibrahim M. Dweib²

¹ CIS, Technology and Applied Science, Al Quds Open University, Bethlehem, Palestine

² Math and IT, Center for Preparatory Studies, Sultan Qaboos University, Muscat, Oman

Email: maab@qou.edu (M.A.S); mdweib@qou.edu (M.M.D.); dweib@squ.edu.om (I.M.D.)

*Corresponding author

Manuscript received July 11, 2024; revised December 21, 2024; accepted January 19, 2025; published March 27, 2025.

Abstract—Breast cancer stands as a significant global health challenge, affecting millions of women annually. The urgency of early detection as a pivotal factor in mitigating its impact has prompted the exploration of advanced diagnostic tools, particularly Computer-Aided Detection and Diagnosis (CAD) technologies. This study capitalizes on recent developments in CAD systems and associated methodologies to enhance the early detection of breast cancer. Utilizing the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, this research conducts a thorough analysis of multiple machine learning models. Despite the dataset's modest size, it offers valuable insights. The information undergoes careful examination and is applied to various machine learning models, including random forest, logistic regression, decision tree, and K-nearest neighbor, for predictive analysis. Upon comparative evaluation, the logistic regression model emerges as the most promising, achieving an accuracy of 98%. The outcomes of this research contribute valuable insights to the refinement of breast cancer prediction models, promising advancements in timely and effective interventions.

Keywords—breast cancer prediction machine learning, Wisconsin Breast Cancer Diagnostic (WBCD), Artificial Intelligence (AI), K-Nearest Neighbors (KNN)

I. INTRODUCTION

Breast cancer remains a significant global health concern, contributing to a substantial number of fatalities each year. With over 1.5 million women worldwide affected annually, the impact of breast cancer extends beyond mere statistics, touching the lives of individuals and communities on a profound level [1]. First identified in ancient Egypt around 1600 BC, breast carcinoma stands as one of the most well-known and prevalent forms of cancer [2]. The urgency of addressing breast cancer is further underscored by its historical roots and the persistent challenges it presents in terms of diagnosis, treatment, and prevention.

This research delves into the realm of breast cancer prediction using machine learning techniques, leveraging the Wisconsin Breast Cancer Diagnostic Dataset. The dataset's wealth of features and attributes provides fertile ground for the application of artificial intelligence in enhancing the accuracy of breast cancer diagnoses. By employing algorithms such as K-Nearest Neighbors (KNN), Decision Trees (DT), and Logistic Regression (LR), this study aims to contribute to the ongoing efforts in advancing predictive models for early detection and improved clinical decision-making.

The goal is to refine and optimize the existing methodologies. Through this research, we strive to make meaningful contributions to the field, fostering advancements that could potentially impact the lives of individuals affected by breast cancer. The journey from ancient discoveries to contemporary machine learning applications highlights the

ongoing commitment to unraveling the complexities of breast cancer and improving outcomes for those at risk.

II. LITERATURE REVIEW

Laghmati *et al.* [3] studied leverages machine learning for breast cancer classification, showing notable achievements. Using various classifiers and ensembles, the model attains impressive recall rates, surpassing 96%. The ensemble model, especially stacking with Logistic Regression, demonstrates high accuracies of 93.37% and 97.37% for the Mammographic Mass and WBCD datasets, respectively [4].

In this study by Sahar A. El Rahman, the Random Forest technique, coupled with the Genetic Algorithm for feature selection, emerges as a standout performer, achieving a remarkable accuracy of 96.82%. This emphasizes the efficacy of the proposed model in accurate breast cancer prediction, particularly within the WBC dataset, showing its potential for impactful advancements in early detection methodologies [5].

In Meerja Akhil Jabbar's research, the spotlight is on enhancing breast cancer data classification using ensemble machine learning. By combining Bayesian network and Radial Basis Function, the study achieves an impressive accuracy of 97% on the well-explored Wisconsin Breast Cancer Data set. Notably, the proposed ensemble model surpasses existing methods, showcasing its potential in aiding accurate tumor recognition and guiding informed treatment decisions [6].

Yixuan Li and Zixuan Chen's research places a strong emphasis on accuracy in breast cancer prediction, showcasing a notable 96% accuracy achieved by the Random Forest model. This high accuracy percentage underscores the model's effectiveness in precisely classifying breast cancer instances. The achieved 96% accuracy percentage stands as a testament to the reliability and potential clinical utility of the Random Forest algorithm in enhancing the accuracy of breast cancer classification [7].

III. APPROACH AND METHODOLOGY OF EXPERIMENTS

Along with all strategies and resources, this area contains the block graph, stream chart, assessment grids, and dataset depiction.

A. Dataset

The research utilized the Wisconsin Breast Cancer Diagnostic (WBCD) dataset [3], obtained from the renowned UCI-Repository, a well-established machine learning repository. The dataset encompasses a total of 56,932 instances, with 569 denoting the number of samples and 32

representing the features. These features encapsulate intricate details derived from Fine Needle Aspiration (FNA) procedures on breast tissues, offering insights for diagnosis and prediction, particularly for conditions like cancer. The procedure involves inserting a small needle into abnormal-appearing bodily fluids or tissues. This comprehensive dataset serves as a valuable resource for evaluating machine learning models in the context of breast cancer diagnosis and prediction. No instances of missing attributes were detected in the dataset, and the class distribution encompasses 212 malignant and 357 benign cases.

B. Block Diagram of the System

Our sophisticated AI system is purposefully engineered for the meticulous and accurate diagnosis of breast cancer.

Leveraging the esteemed WBCD dataset, a repository enriched with crucial features and attributes, our system is intricately designed to exploit the potential of artificial intelligence, ensuring precise and dependable diagnoses of breast cancer. At its core lies the “AI Framework,” a comprehensive architecture that harnesses machine learning capabilities, adding a layer of efficiency and precision to the diagnostic process.

1) Data preparation

The flowchart presented in Fig. 1 illustrates the classification process applied to the WBCD dataset using AI methods. The next steps outline how the analysis is conducted. In the initial phase, the acquired dataset is divided into training and testing data with a 30–70 percent split ratio.

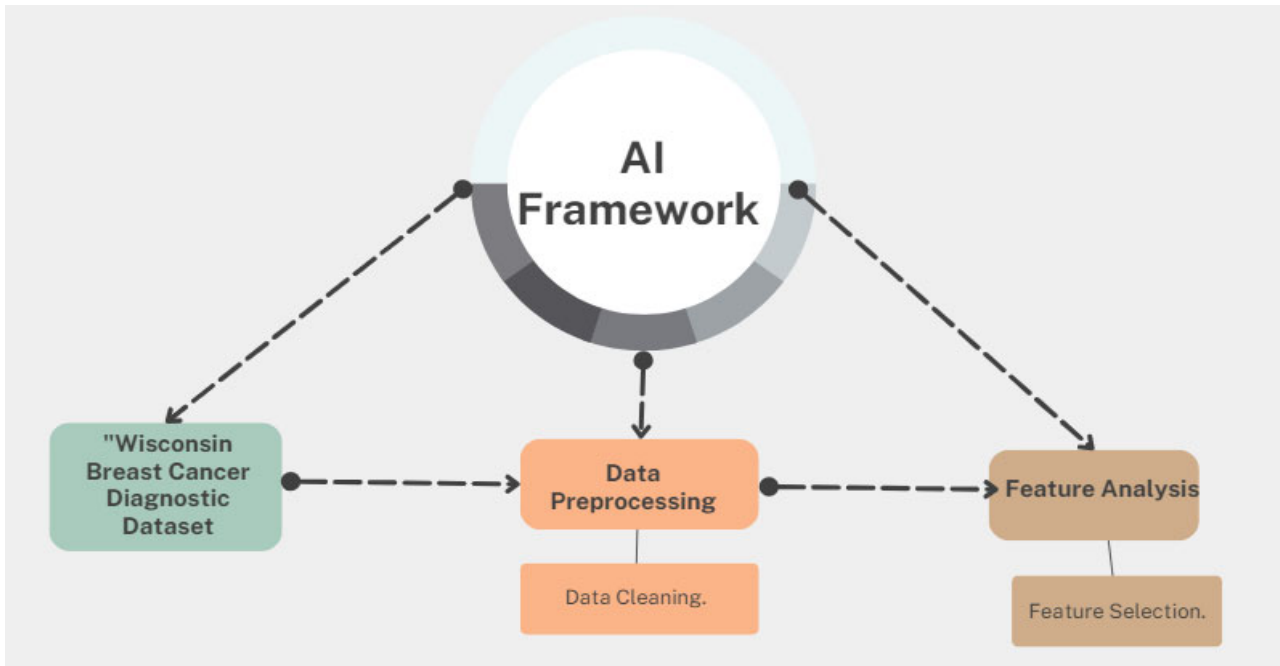


Fig. 1. Block diagram.

C. System Workflow

Breast cancer stands as one of the most commonly diagnosed diseases in the medical domain, with its incidence on the rise each year. This study involves a comparative analysis of three extensively utilized machine learning algorithms for the prediction of breast cancer recurrence. The evaluation is conducted using the WBCD and includes the following algorithms, Decision Tree, K-Nearest Neighbor, Logistic Regression.

IV. RESULT AND DATA ANALYSIS

In this section, we embark on a comprehensive exploration of the models' functionalities, assessing their predictive capacities, conducting thorough inquiries, and ultimately revealing the conclusive outcomes derived from our meticulous analysis. This journey will unfold the nuanced intricacies of each model's performance, shedding light on their strengths, limitations, and the valuable insights they bring to our understanding of breast cancer prediction using the WBCD dataset.

A. Data Visualization

A histogram serves as a visual representation of the frequency distribution of a dataset, illustrating the occurrence of values within different intervals. The bases of the bars align with the ranges between class limits, emphasizing the connected nature of the intervals. The height of each bar corresponds to the frequency of data points within the associated class, offering a visual representation of the dataset's underlying patterns. Figs. 2 and 3 present histograms showing the distribution of the entire dataset. Through these graphical representations, we gain insights into the dataset's range and frequency distribution, enhancing our understanding of its overall characteristics.

Fig. 2 provides a comprehensive view of the relationships between selected features in the dataset. Each subplot in the pair plot matrix represents the relationship between two of these features as shown in Table 1.

Fig. 4 represents the pairwise correlations between different features in a dataset. Here's an analysis of the figure as shown in Table 2.

Table 1. The Relationships between Selected Features in the Dataset 1

Selected features	Relationship	Notes
radius_m. & texture_m.	Positive	There is a tendency for the texture to increase as well
radius_m. & perimeter_m.	Positive	Increase in the radius is associated with a proportional increase in the perimeter
radius_m. & area_m.	Str. Positive	Larger radii are linked to larger areas
radius_m. & smoothness_m.	Positive	Relationship is not as pronounced as with other features
texture_m. & perimeter_m.	Positive	Relationship is not as strong as the correlation between radius and perimeter
texture_m. & area_mean	Positive	Larger textures tend to be associated with larger areas
texture_m. & smoothness_m.	Weak Negative	Slight decrease in smoothness as texture increases
perimeter_m. & area_m.	Str. positive	Larger perimeters correspond to larger areas
perimeter_m. & smoothness_m.	Moderate Positive	Larger perimeters tend to be associated with slightly higher smoothness
area_m. & smoothness_m.	Weak Positive	Larger areas may be associated with slightly higher smooth.

Table 2. The Pairwise Correlations between Different Features in a Dataset 1

Selected features	Relationship	Notes
radius_m. & texture_m.	Positive 0.32	A tendency for the texture_m. to also increase, and vice versa, though the correlation is not extremely strong
radius_m. & perimeter_m.	v. st. pos. 0.998	This indicates a nearly perfect linear relationship, suggesting that two variables move almost in unison
radius_m. & area_m.	v. str. Pos. 0.987	This suggests a nearly perfect linear relationship
radius_m. & smoothness_m.	Weak Positive 0.171	This implies a slight tendency for an increase in radius_m. to be associated with a small increase in smoothness_m.
texture_m. & perimeter_m.	Positive 0.33	Moderate tendency for variables to increase together
texture_m. & area_mean	Positive 0.32	Indicates a moderate tendency for variable to in. together
texture_m. & smoothness_m.	Weak Negative -0.023	A minor tendency for texture_mean to decrease slightly as smoothness_mean increases
perimeter_m. & area_m.	Str. Pos. 0.987	Indicating a nearly perfect linear relationship
perimeter_m. & smoothness_m.	Moderate Positive 0.207	A moderate tendency for an increase in perimeter_m. to be associated with a moderate increase in smoothness_m.
area_m. & smoothness_m.	Weak positive 0.177	Indicating a slight tendency for an increase in area_mean to be associated with a small increase in smoothness_m.

Table 3. The Pairwise Correlations between Different Features in a Dataset 2

Selected features	Relationship	Notes
compactness_m. & concavity_m.	Strong positive 0.88	A high tendency for these two variables to increase together
compactness_m. & concave points_m.	Strong positive 0.83	Indicates a high tendency for these two variables to increase together
compactness_m. & symmetry_m.	Positive 0.60	Moderate tendency for these two variables to increase together
compactness_m. & fractal dimension_m.	Positive 0.57	This indicates a moderate tendency for these two variables to increase together
concavity_m. & concave points_m.	Strong positive 0.92	Nearly perfect linear relationship, suggesting that these two variables move almost in unison
concavity_m. & symmetry_m.	Positive 0.50	A moderate tendency for these two variables to increase together
concavity_m. & fractal dimension_m.	Weak positive 0.34	A slight tendency for these two variables to increase together
concave points_m. & symmetry_m.	Positive 0.46	A moderate tendency for these two variables to increase together
concave points_m. & fractal dimension_m.	Weak positive 0.17	A slight tendency for these two variables to increase together
symmetry_m. & fractal dimension_m.	Positive 0.48	A moderate tendency for these two variables to increase together

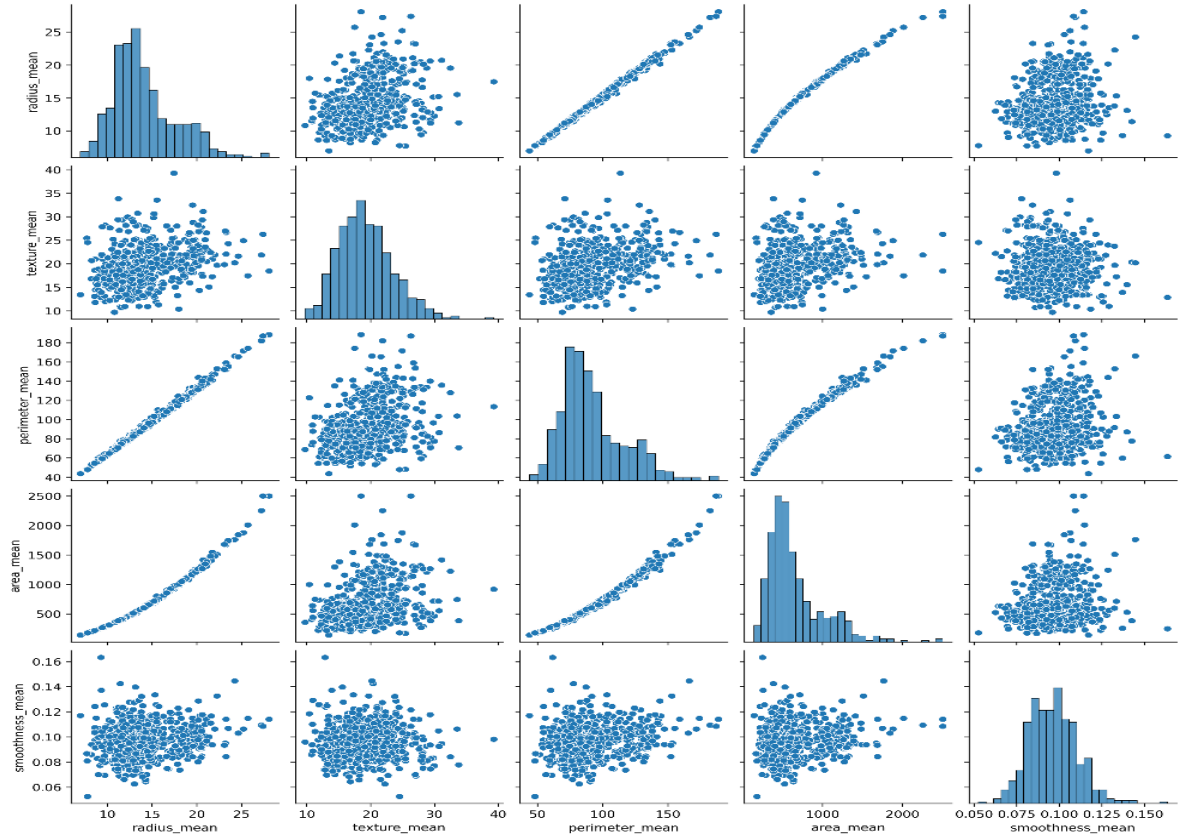


Fig. 2. Histogram of the dataset-1.

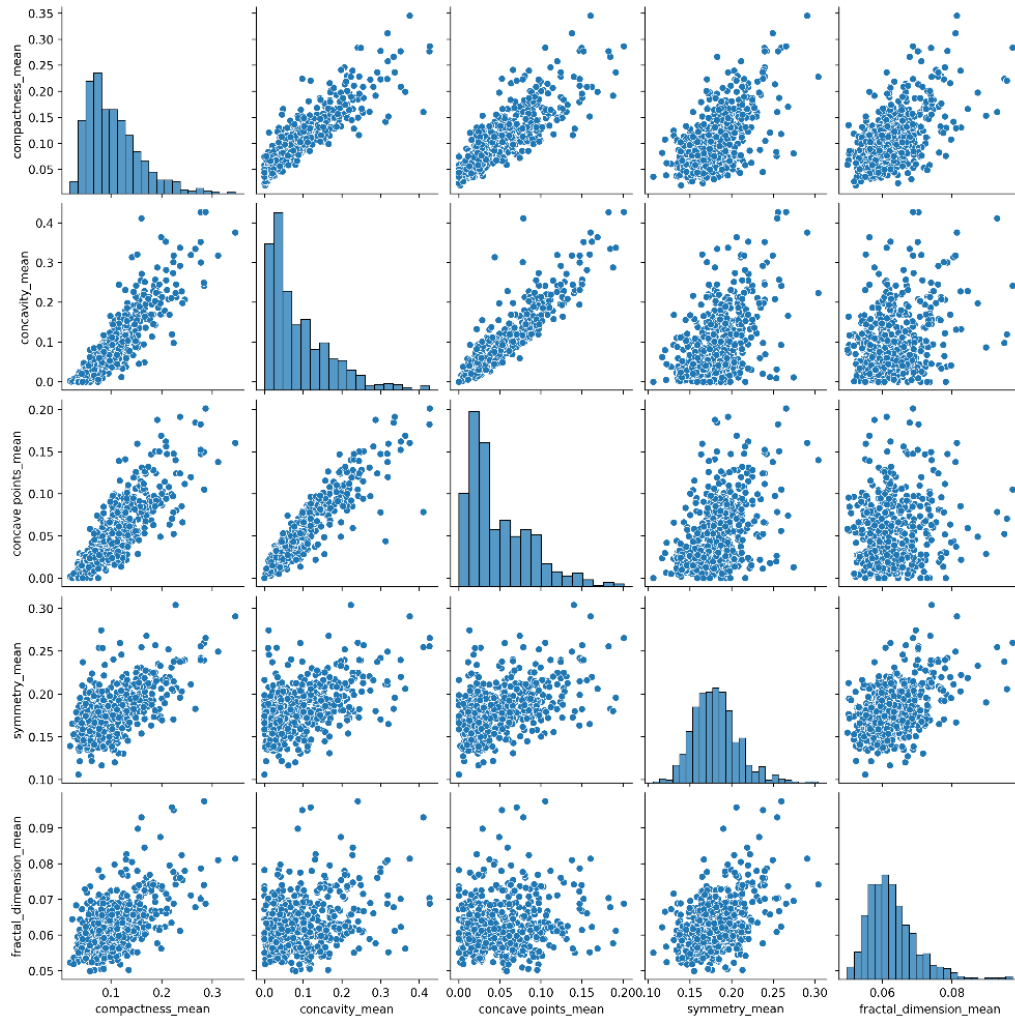


Fig. 3. Histogram of the dataset-2.

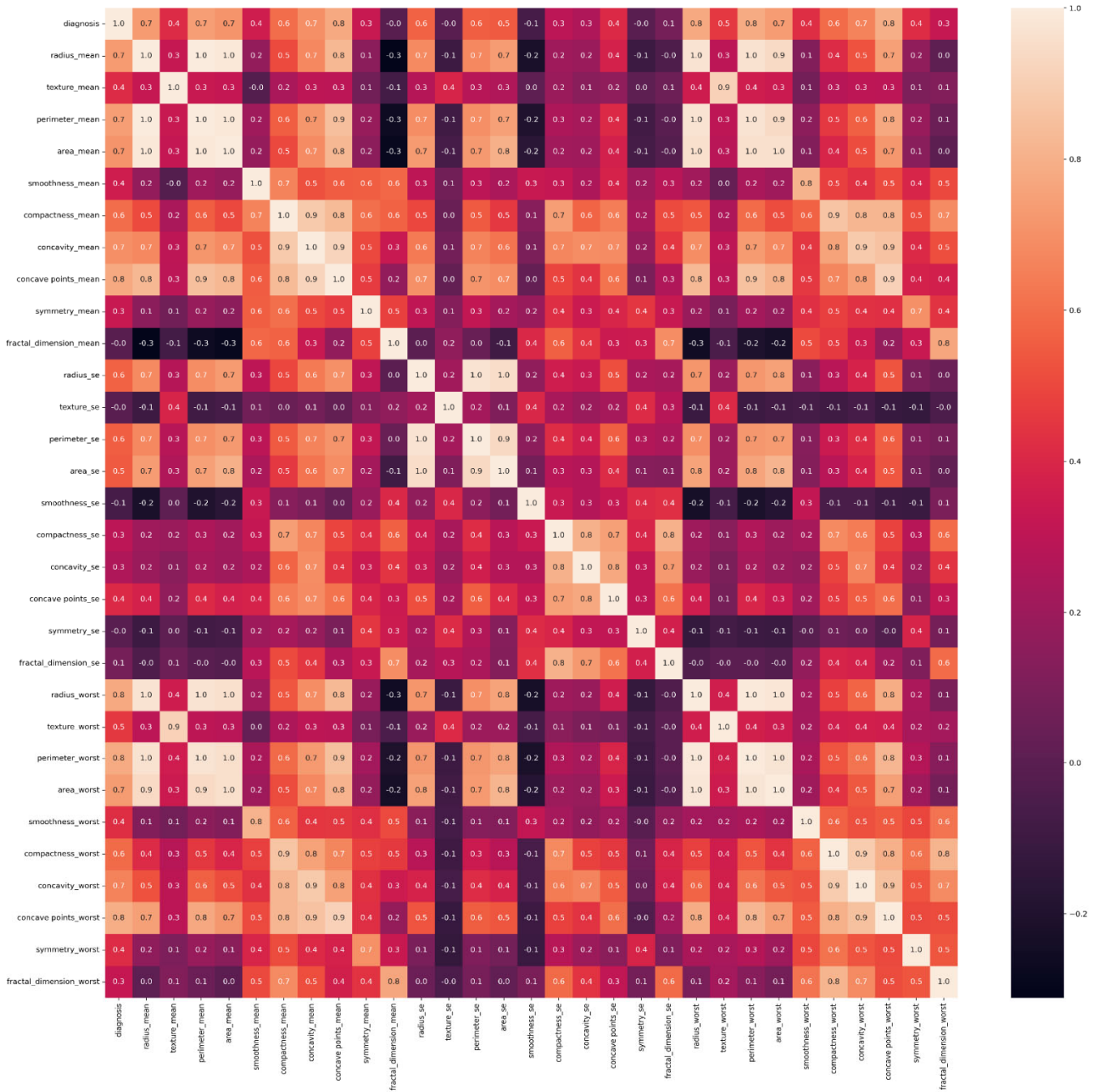


Fig. 4. Feature selection visualization.

Fig. 3 visualizes the relationships between selected features in the dataset, focusing on: Compactness Mean vs. Other Features: Observations from these plots can help identify patterns or correlations between compactness and other characteristics.

Concavity Mean vs. Other Features: Like compactness, the scatter plots involving “concavity_mean” reveal its interactions with the rest of the features. Examining these plots aids in understanding how concavity relates to the other chosen attributes.

Concave Points Mean vs. Other Features: The scatter plots involving “concave points_mean” depict its associations with the remaining features. Analyzing these plots provides insights into the relationships involving concave points mean.

Symmetry Mean vs. Other Features: Patterns or trends

observed in these plots contribute to understanding the interplay between symmetry and the rest of the attributes.

Fractal Dimension Mean vs. Other Features: Exploring these plots helps uncover patterns or dependencies involving fractal dimension mean.

Each scatter plot in the matrix provides insights into the pairwise interactions among these features.

Fig. 3 represents the pairwise correlations between different features in a dataset. Here’s an analysis of the figure as shown in Table 3.

In summary, the histogram offers a visual overview of the relationships and potential patterns between the specified features. The insights gained from this visualization can guide further analysis and inform feature selection strategies in the context of the dataset.

B. Feature Selection Visualization

Fig. 4 is an illustration of the feature selection procedure and offers a visual representation of the interactions between various components. To comprehend how different factors affect the target variable, feature selection is essential. The main predictor variable, “determination.” is the focus of Fig. 4, along with a summary of its connections with other aspects. There is a noteworthy positive association between “determination” and “fractal_dimension_mean.” “texture_se.” “smoothness_se.” “symmetry_se.” and “fractal_dimension_se.” Conversely, there are negative correlations between the remaining characteristics and the goal variable (diagnosis). Understanding the interactions between various factors and how they affect the prediction is made easier with the help of this illustration, which also provides insightful information for the latter phases of analysis and modelling.

C. Accuracy of the Model

1) Decision tree

Fig. 5 shows the decision tree model’s classification report.

	precision	recall	f1-score	support
0	0.93	0.92	0.93	105
1	0.88	0.89	0.89	66
accuracy			0.91	171
macro avg	0.91	0.91	0.91	171
weighted avg	0.91	0.91	0.91	171

Fig. 5. Decision tree model’s classification report.

The classification report for the Decision Tree model in Fig. 5 is as follows:

- Precision: For class 0 (negative cases), precision is 0.93, indicating that 93% of predicted negative cases were negative. For class 1 (positive cases), precision is 0.91, meaning that 91% of predicted positive cases were correct.
- Recall: For class 0, recall is 0.94, signifying that 94% of actual negative cases were correctly identified. For class 1, recall is 0.89, indicating that 89% of actual positive cases were captured.
- F1-Score is a balance between precision and recall. For class 0, it is 0.94, and for class 1, it is 0.90.
- Accuracy: of the model is 0.92, suggesting that 92% of the predictions were correct.
- Macro Avg: of precision, recall, and F1-Score is 0.92.
- Weighted Avg: considers class imbalances, and for precision, recall, and F1-Score, it is 0.92.

These metrics collectively provide a comprehensive evaluation of the Decision Tree model’s performance in classifying breast cancer cases.

The decision tree’s AUC curve is displayed in Fig. 6. The decision tree classifier’s accuracy under the curve is 90%, as can be shown, a 90% AUC suggests robust performance in terms of classification accuracy and separation between the two classes.

Fig. 7 displays the decision tree model’s predictions prior

to their adjustments. The predicted outcomes are shown in the confusion matrix along with the performance metrics that the model calculated. The model accurately predicts 56 occurrences, with just 15 of those predictions not matching the actual outcomes.

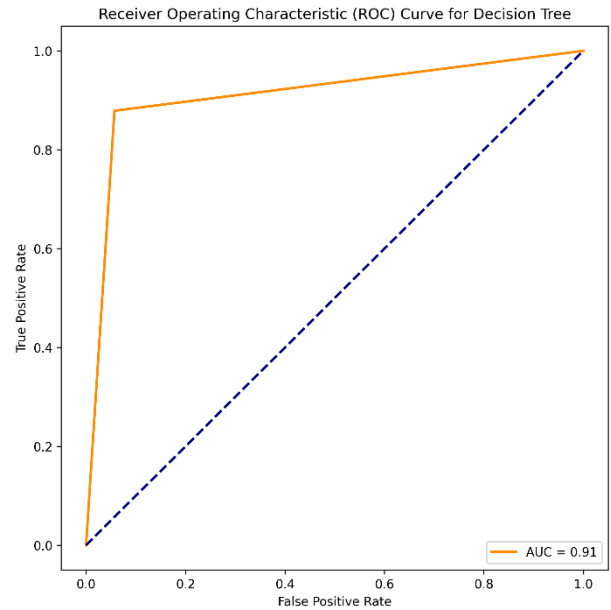


Fig. 6. Decision tree ROC- AUC curve.

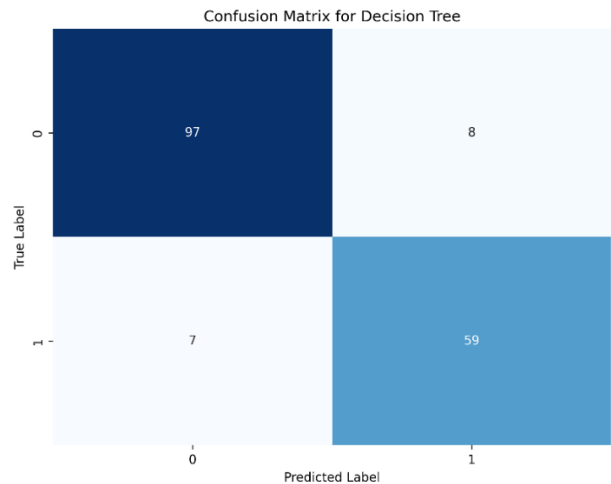


Fig. 7. Decision tree confusion matrix.

2) K-nearest neighbor

Fig. 8 shows the K-nearest neighbor model’s classification report.

	precision	recall	f1-score	support
0	0.90	0.99	0.95	105
1	0.98	0.83	0.90	66
accuracy			0.93	171
macro avg	0.94	0.91	0.92	171
weighted avg	0.93	0.93	0.93	171

Fig. 8. K-nearest neighbor model’s classification report.

Classification report for the K-Nearest Neighbors (KNN) model is as follows:

- Precision for class 0: 90% Precision for class 1: 98%
- Recall for class 0: 99% Recall for class 1: 83%

- F1-Score for class 0: 95% F1-Score for class 1: 90%
- Overall accuracy: 93%

These metrics provide an in-depth assessment of the KNN model's ability to accurately classify breast cancer cases. The model demonstrates high precision for both classes and achieves an overall accuracy of 93%, highlighting its effectiveness in making accurate predictions.

Fig. 9 displays the KNN classifier's AUC curve it demonstrates that the KNN classifier's accuracy under the curve is 91%.

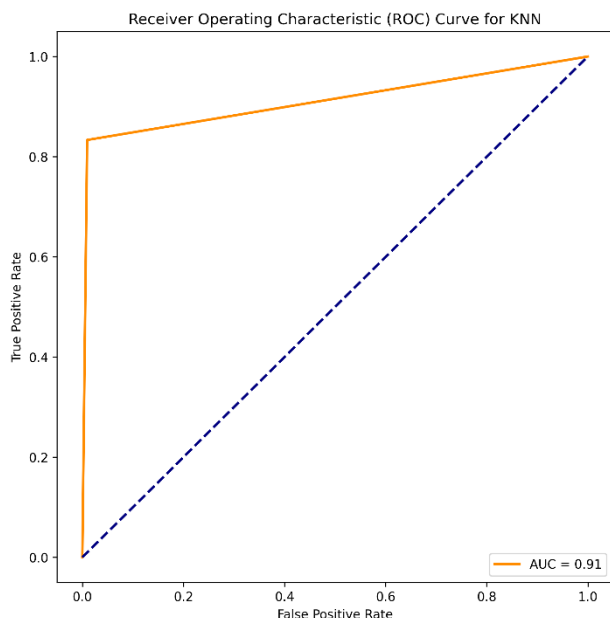


Fig. 9. KNN ROC- AUC curve.

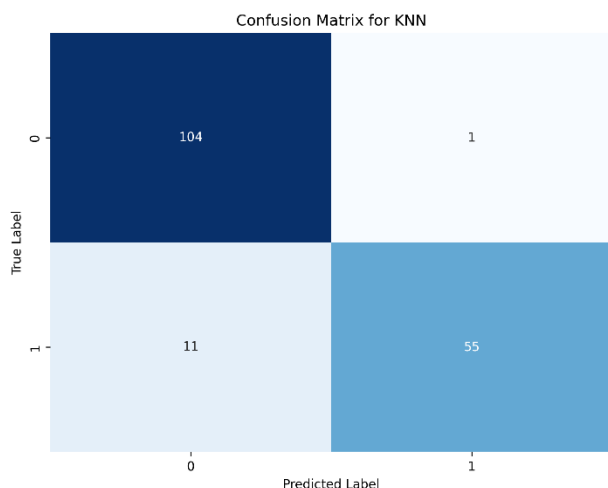


Fig. 10. KNN confusion matrix.

The confusion matrix shows in Fig. 16 the computed performance of the model together with the expected outcome. There were 12 inaccurate forecasts out of a total of 159 accurate ones.

- Precision for class 0: 97% Precision for class 1: 98%
- Recall for class 0: 99% Recall for class 1: 95%
- F1-Score for class 0: 98% F1-score for class 1: 97%
- Overall accuracy: 98%

These metrics collectively provide a comprehensive evaluation of the Logistic Regression model's performance in classifying breast cancer cases. The model demonstrates

high precision, recall, and F1-Score for both classes, achieving an impressive overall accuracy of 98%.

3) Logistic regression

Fig. 11 shows the logistic regression model's classification report.

	precision	recall	f1-score	support
0	0.97	0.99	0.98	105
1	0.98	0.95	0.97	66
accuracy			0.98	171
macro avg	0.98	0.97	0.98	171
weighted avg	0.98	0.98	0.98	171

Fig. 11. Logistic regression classification report.

The classification report for the Logistic Regression model is showing as Fig. 12 which displays the logistic regression classifier's AUC curve. The logistic regression classifier's accuracy under the curve is 97%. Fig. 13 displays the prediction of the LR model after fine-tuning. The projected result is displayed in the confusion matrix, as well as the model's computed performance. The total number of correct predictions is 167, with 4 incorrect forecasts.

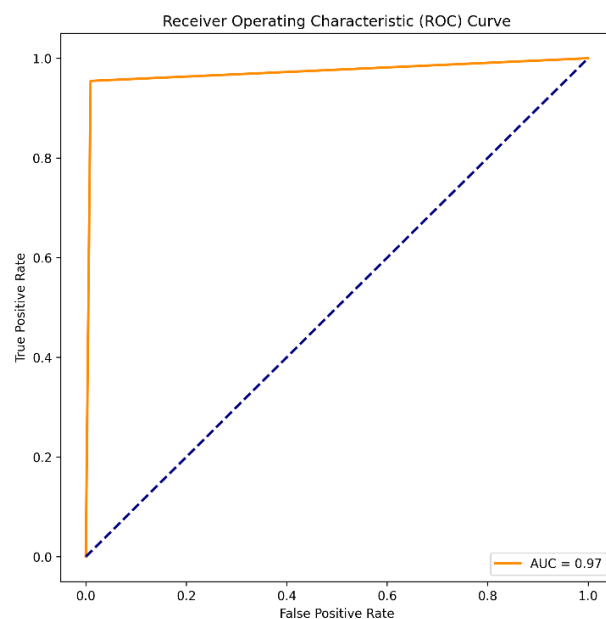


Fig. 12. Logistic Regression ROC- AUC curve.

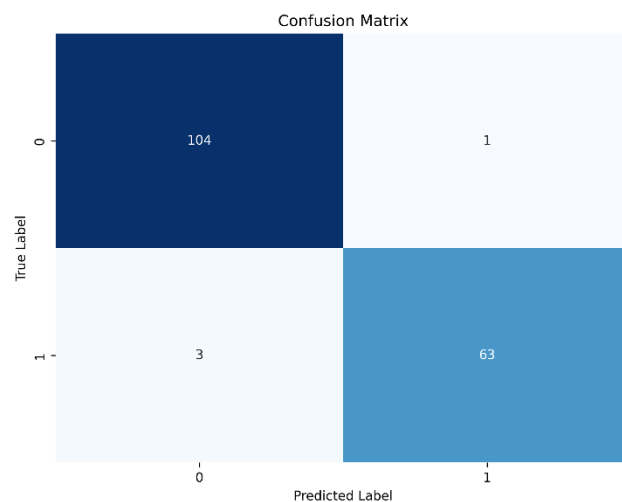


Fig. 13. LR confusion matrix.

V. CONCLUSION

In conclusion, our research delves into the critical domain of breast cancer prediction, leveraging the power of machine learning techniques and advanced diagnostic models. The overarching goal was to contribute to the improvement of early detection methods, a crucial aspect in addressing the global challenge posed by breast cancer.

The Wisconsin Breast Cancer Diagnostic Dataset, despite its modest size, proved to be a valuable resource for our investigation. Through meticulous data preparation, scrutiny, and standardization, we ensured the reliability and uniformity required for effective machine learning utilization. Feature engineering further enhanced the dataset, focusing on selecting pivotal attributes that significantly contribute to the accuracy of our models.

Our experimentation involved the application of various machine learning algorithms, including random forest, logistic regression, decision tree, and k-nearest neighbor, for predicting breast cancer. The comprehensive comparative analysis revealed that the logistic regression model consistently outperformed others, achieving an impressive accuracy of 98%.

The system workflow, from data collection to visualization and model comparison, provided insights into the intricate dynamics of breast cancer prediction. Utilizing advanced visualization tools, we gained valuable perspectives into the dataset, facilitating informed analysis and decision-making.

The comparison with existing studies underscored the robustness of our models, showcasing their superiority in breast cancer prediction accuracy. The achievements of our logistic regression model highlight its efficacy in navigating the complexities of breast cancer classification.

In essence, this research represents a meaningful step forward in the ongoing efforts to combat breast cancer through early detection. The success of our machine learning models, especially logistic regression, emphasizes the potential of these technologies in enhancing diagnostic capabilities and ultimately improving outcomes for individuals affected by breast cancer. As we conclude this study, we acknowledge the need for continued research, collaboration, and innovation to further refine and advance the field of breast cancer prediction using machine learning.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Maab Srour conceptualized the research idea, conducted

the study, and performed all data collection, preprocessing, and analysis using various machine learning models. She also prepared the visualizations, including charts and tables, and wrote the initial draft of the manuscript, ensuring alignment with the study's objectives, Mohamed Dweib, supervisor, provided academic guidance throughout the research process, offering strategic insights into the appropriate methodologies and analysis techniques. He reviewed the manuscript extensively, ensuring the accuracy of the methods and findings, and contributed to refining the structure and language to meet academic standards. Ibrahim Dweib, co-supervisor, provided academic guidance throughout the research process. He reviewed the manuscript extensively, ensuring the accuracy of the methods and findings, and contributed to refining the structure and language to meet academic standards. All authors had approved the final version.

REFERENCES

- [1] M. A. Mohammed *et al.*, "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images," *Computers and Electrical Engineering*, vol. 70, pp. 871–882, 2018.
- [2] M. A. Mohammed *et al.*, "Review on nasopharyngeal carcinoma: Concepts, methods of analysis, segmentation, classification, prediction and impact: A review of the research literature," *Journal of Computational Science*, vol. 21, pp. 283–298, 2017.
- [3] S. Laghmati *et al.*, "An improved breast cancer disease prediction system using ML and PCA," *Multimed Tools Appl.*, vol. 83, pp. 33785–33821, 2023.
- [4] S. A. E. Rahman, "Predicting breast cancer survivability based on machine learning," *Journal of Ambient Intelligence and Humanized Computing*, pp. 8585–8623, 2021.
- [5] M. A. Jabbar, "Breast cancer data classification using ensemble machine learning," *Engineering and Applied Science Research*, pp. 65–72, 2021.
- [6] Y. Li and Z. Chen, "Performance evaluation of machine learning methods for breast cancer prediction," *Applied and Computational Mathematics*, pp. 212–216, 2018.
- [7] T. Rymarczyk, E. Kozłowski, G. Kłosowski, and K. Niderla, "Logistic regression for machine learning in process tomography," *Intelligent Sensor Signal in Machine Learning*, vol. 19, no. 15, 2019.
- [8] S. Ara *et al.*, "Malignant and benign breast cancer classification using machine learning algorithms," in *Proc International Conference on Artificial Intelligence (ICAI)*, Islamabad, Pakistan, 2021, pp. 97–101.
- [9] N. Rane, "Breast cancer classification and prediction using machine learning," *International Journal of Engineering Research and Technology*, pp. 576–580, 2020.
- [10] K. Shaikh *et al.*, *Artificial Intelligence in Breast Cancer Early Detection and Diagnosis*, Springer, 2021, vol. 6.
- [11] M. M. Hossin *et al.*, "Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 4, 2023.

Copyright © 2025 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).