Spearman-guided XGBoost Classification for Breast Cancer Diagnosis Using the Wisconsin Diagnostic Dataset

Hani Attar [1,2,*], Jafar Ababneh [3], Waleed Alomoush [4,5], Samer M. Sharfo [6], Mohamad Hafez [7,8], and Mohanad A. Deif [9]

[1] Faculty of Engineering, Zarqa University, Zarqa, Jordan
[2] College of Engineering, University of Business and Technology, Jeddah, Saudi Arabia
[3] Cyber Security department, Faculty of Information Technology, Zarqa University, Zarqa, Jordan
[4] Artificial Intelligence Department, Plekhanov Russian University of Economics in Dubai, Dubai Knowledge Park, Dubai, UAE
[5] Computer Science Department, College of Computing and Intelligent Systems, University of Al Dhaid, Sharjah, UAE
[6] Department of Bioelectronics, Modern University of Technology and Information (MTI) University, Egypt
[7] Faculty of Engineering FEQS, INTI-IU-University, Nilai, Malaysia
[8] Faculty of Management, Shinawatra University, Pathum Thani, Thailand
[9] Department of Artificial Intelligence, College of Information Technology, Misr University for Science & Technology (MUST), Giza, Egypt
Email: hattar@zu.edu.jo (H.A.); jababneh@zu.edu.jo (J.A.); Alomoush.W@reu.tech and walomoush@uodh.ac.ae (W.A.); samer.86027@eng.mti.edu.eg (S.M.S.); mohdahmed.hafez@newinti.edu.my (M.H.); Mohanad.deif@must.edu.eg (M.A.D.)
*Corresponding author

Manuscript received February 13, 2025; revised April 26, 2025; accepted December 24, 2025; published September 15, 2026

***Abstract*—Breast cancer remains one of the leading causes of mortality among women worldwide, underscoring the critical need for effective and early diagnostic tools. This study presents a comprehensive Machine Learning (ML) framework that employs *k*-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LR), and Extreme Gradient Boosting (XGBoost) algorithms to analyze a breast cancer dataset obtained from the University of California, Irvine (UCI) repository. The dataset was partitioned into 70% training and 30% testing subsets to evaluate the generalization performance of each model. The Logistic Regression (LR) model achieved the highest accuracy at 96.5%, demonstrating its effectiveness in modeling linear decision boundaries. The RF algorithm followed with 95.3% accuracy, reflecting its capability in capturing complex, non-linear interactions. XGBoost achieved a competitive accuracy of 95.9%, highlighting its robustness in detecting subtle patterns and improving predictive performance. The KNN model recorded an accuracy of 92.9%, indicating its effectiveness in identifying localized patterns within the data. These findings underscore the potential of ML-driven approaches in enhancing breast cancer diagnostics and contribute meaningfully to public health by supporting timely and accurate disease detection.**

***Keywords*—Machine Learning (ML), breast cancer classification, linear relationships, classification algorithms, cancer prediction, supervised learning**

I. Introduction

Globally, breast cancer is a severe health issue, and increasing survival rates depend heavily on early detection [1]. Machine Learning (ML) techniques have demonstrated the potential to support early detection by providing suitable tools for analysis and forecasting. This paper conducts a comprehensive analysis utilizing the Breast Cancer dataset from the University of California, Irvine (UCI) ML Repository to examine the effectiveness of various machine-learning methods in identifying breast cancer.

This research utilizes the publicly available Breast Cancer Wisconsin (Diagnostic) dataset from the UCI Machine Learning Repository [2]. Several ML algorithms were applied, including Logistic Regression (LR) [3], Random Forest (RF) [4], *k*-Nearest Neighbors (KNN) [5], Support Vector Machines (SVM) [6], and Extreme Gradient Boosting (XGBoost) [7], as these methods have been widely adopted in prior studies for medical diagnostics due to their predictive performance and interpretability. The class labels were preprocessed by converting the original values ("B" for benign and "M" for malignant) into binary format: "B" was mapped to 0 and "M" to 1; indeed, this conversion is a standard step in binary classification tasks that ensures compatibility with ML algorithms. Following label conversion, the dataset was evaluated for missing values (none were found), and outliers were handled using box plot analysis and min-max clipping. Finally, feature selection was based on Spearman correlation coefficients.

This work aims to further research on ML for breast cancer detection through evaluating the efficacy of diverse algorithms, investigating distinct preprocessing methods, and implementing feature selection methodologies. Moreover, valuable perspectives were sought that may steer forthcoming investigations and enhance the precision of breast cancer identification.

II. Literature Review

ML and Artificial Intelligence (AI) techniques have shown remarkable promise in breast cancer detection, diagnosis, and classification in recent years. Various architectures and algorithms have been explored, each addressing different medical imaging and data analysis aspects.

Nair and Subaji [8] implemented transfer learning techniques using architectures such as ResNet50 and VGG16, combined with ensemble learning approaches like Extra Trees, to classify different breast cancer types. While effective, the study highlighted the need for validation on larger and more diverse datasets.

Elsheakh *et al.* [9] adopted a gradient boosting technique (CAT-Boost) for early breast cancer detection by identifying tumor components within ultrasound images. However, classifier accuracy was impacted due to limitations in feature availability.

Sushanki *et al.* [10] employed Convolutional Neural Networks (CNNs) and transfer learning to segment and classify breast cancer from image data, where the proposed work underscores the reliance on high-quality input data and

the challenges associated with interpreting Deep Learning (DL) models.

Nogales *et al.* [11] proposed a hybrid architecture that integrates CNNs with evolutionary algorithms for enhanced diagnosis using thermographic imaging. In their work, they emphasized the lack of open-access datasets as a key limitation, hindering model reproducibility.

For intraoperative cancer diagnosis, Zhang *et al.* [12] developed a deep learning-based Dynamic Full-Field Optical Coherence Tomography (D-FFOCT) framework capable of distinguishing benign from malignant tissues. The model was promising but required further validation due to the limited sample pool and rarity of specific tumor types.

Aziz *et al.* [13] advanced breast cancer detection and risk assessment using a mix of deep neural networks, Long Short-Term Memories (LSTMs), and Support Vector Machines (SVMs), focusing on improving localization and classification accuracy. Nonetheless, the study identified challenges in interpreting lesion characteristics.

Ahmad *et al.* [14] tackled the problem of breast cancer grading through a nuclei-aware network, tailored for pathology image analysis, where the approach was hindered by a lack of detailed dataset annotations, which are crucial for clinical deployment.

Arshad *et al.* [15] introduced a CNN-based BreastNet-SVM framework for the identification of cancer in mammogram images. Despite high accuracy, the model was limited by reliance on single datasets and variability in image sizes.

To support radiologists in automated cancer identification, Lu *et al.* [16] combined several pre-trained CNNs, including InceptionV3, ResNet152V2, and DenseNet-121; moreover, the system demonstrated the importance of data diversity and highlighted ongoing challenges in model generalization.

Lastly, Sahu *et al.* [17] presented a comparative study using RF, SVM, and Artificial Neural Networks (ANNs) to predict breast cancer recurrence, where the work focused on systematic review and meta-analysis but encountered issues such as bias, inconsistency, and a lack of external validation.

The above-mentioned studies reflect the breadth of approaches currently employed in AI-based breast cancer detection. Despite significant advancements, common challenges remain, particularly around dataset quality, model interpretability, and the need for external validation; the research gaps that must be addressed in future work will facilitate the clinical translation of AI tools.

III. Methodology

The UCI ML Repository's Breast Cancer Wisconsin [18] (Diagnostic) dataset is frequently used for binary classification tasks about breast cancer diagnosis, which consists of characteristics taken from digital images of breast mass Fine Needle Aspirates (FNAs), intending to differentiate between benign (non-cancerous) and malignant (cancerous) tumours. The dataset offers a complete set of thirty characteristics generated from every FNA image and captures different facets of the texture and morphology of the mass; whereas these features hold clinical significance as follows:

- Radius: Represents the mean distance from the center to the cell perimeter and correlates with tumor size.
- Texture: Quantifies variations in grayscale intensity and can indicate malignancy due to irregular growth patterns.
- Smoothness: Measures local variations in contour sharpness, where higher values may indicate aggressive tumor growth.

The above-mentioned features are commonly used in histopathological assessments to differentiate benign from malignant tumors.

The dataset's features provide a wide variety of quantitative measurements. Some essential characteristics are mean radius, mean texture, mean fractal dimension, mean smoothness, mean compactness, mean concavity, mean concave points, mean symmetry, and mean smoothness. Additionally, the dataset includes corresponding standard errors and worst (most significant) values for each feature; indeed, these features collectively provide a rich set of information regarding the characteristics of breast masses, aiding in developing robust machine-learning models for cancer classification. The Breast Cancer Wisconsin dataset [2] encompasses 30 feature columns, where each feature is a numerical representation derived from image analysis, contributing to a holistic understanding of the underlying tissue properties. Consequently, these features play a pivotal role in the training and evaluation of ML algorithms, allowing researchers to explore various models and techniques for predicting the likelihood of malignancy in breast tumours.

All features in the dataset originate from FNA imaging, where cell nuclei characteristics are extracted from high-resolution microscopic images, and the derived metrics provide valuable insights into tumor pathology, helping in automated breast cancer diagnosis.

The target variable in the Breast Cancer Wisconsin dataset is the "diagnosis" column, which serves as the label for classification tasks. The "diagnosis" column contains two classes: "M" for malignant and "B" for benign. The binary classification setup makes it a suitable dataset for evaluating the performance of ML algorithms in distinguishing between cancerous and non-cancerous breast masses. Researchers often leverage the dataset to benchmark and compare the efficacy of different classification approaches, contributing to advancements in medical diagnostics and predictive modelling for breast cancer.

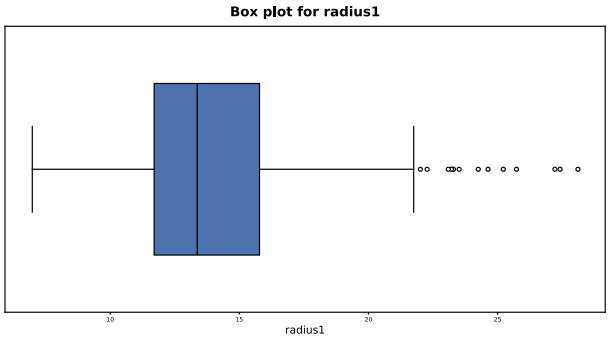
A. Data Preprocessing

The initial phase of the proposed study focused on data preprocessing, a critical step in any ML project. The first task was to handle outliers in the dataset. Outliers were identified in 29 out of the 30 features using box plots. The outliers were then addressed using min-max clipping. The min-max clipping method uses the following formulas:

$$Minimum = Q1 - 1.5 \times IQR \quad (1)$$

$$Maximum = Q3 + 1.5 \times IQR \quad (2)$$

where $Q1$ represents the lower quartile, $Q3$ signifies the upper quartile, and IQR denotes the Interquartile Range. Any data point falling below the minimum or exceeding the maximum was identified as an outlier and subsequently eliminated from the dataset. Fig. 1(a) shows the feature radius1 with outliers.



(a)

(b)

Fig. 1. Box plot: (a) with outliers and (b) without outliers.

After removing outliers, the absence was confirmed by redrawing the box plots to ensure that the remaining data points were within a reasonable range, reducing the potential for skewed results in the subsequent analysis. Fig. 1(b) shows the same feature as shown in Fig. 1(a) without outliers.

Min-max clipping was selected over imputation and robust scaling to preserve the original feature distributions while reducing the influence of extreme outliers, where this method ensures consistency in data preprocessing without introducing artificial data points. Furthermore, the dataset contained no missing values, as confirmed through preliminary verification.

The next step in the preprocessing was data standardization [19]. The *z*-score standardization was applied to transform the data with a mean of 0 and a standard deviation of 1. The *z*-score is calculated as follows:

$$z = \frac{x-\mu}{\sigma} \tag{3}$$

In Eq. (3), x represents a data point, μ stands for the mean, and σ denotes the standard deviation to ensure that all features have equal weight before training the models, preventing bias towards features with larger values. Some of the features after standardization are shown in Table 1.

Table 1. Data standardization results

Sample ID	Radius1	Perimeter1	Area1	Radius2	Area3
Sample 1	1.83	2.27	2.00	2.24	1.62
Sample 2	0.59	0.62	0.51	−0.24	0.46
Sample 3	−0.12	−0.14	−0.19	−1.19	−0.41
Sample 4	1.01	1.00	1.02	0.70	0.36
Sample 5	0.17	0.19	0.04	0.61	−0.10

Ultimately, the data is partitioned into training and testing sets, employing stratification [20] to preserve the original distribution of classes within the label across the entire dataset. The data was designated 70% for the training set and 30% for the test set—additionally, the *k*-fold method was employed with five folds to prevent bias towards a particular sample. The mentioned steps ensured that the results were as true and accurate as possible. Through careful preprocessing, a solid foundation was set for accurate and reliable breast cancer detection study results.

The dataset comprises 357 benign and 212 malignant cases, reflecting a mild class imbalance. Given the relatively balanced distribution, no oversampling or undersampling techniques were employed. Instead, stratified sampling was utilized to preserve the proportional representation of each class during model training and validation, thereby ensuring the reliability and generalizability of the evaluation process.

B. Feature Selection

Feature selection is crucial in building effective ML models, particularly in medical diagnosis, such as breast cancer classification [21]. One approach to feature selection involves examining the correlation between individual features and the target variable. In the case of the Breast Cancer Wisconsin dataset, the Spearman rank correlation coefficient [22] is employed to quantify the strength and direction of the monotonic relationship between each feature and the diagnosis column. The Spearman's rank correlation coefficient is robust to outliers and suitable for identifying non-linear associations.

In this study, a threshold of approximately 0.6 for the absolute Spearman's rank correlation coefficient was chosen as a decisive criterion for feature selection for $n < 10$ and α (Test Significance Level) = 0.10, as it is the threshold of the strong correlations due to the small sample sizes, n. Consequently, suppose the coefficient surpasses 0.6, which implies a robust monotonic relationship (correlation) between the feature and the target variable, underscoring its potential importance in forecasting the diagnosis of breast tumours. The following formula gives Spearman's rank correlation coefficient:

$$\rho = 1 - \frac{6\sum {d_i}^2}{n(n^2-1)} \tag{4}$$

where ρ represents Spearman's rank correlation coefficient, d_i represents the difference between the two ranks of each observation, and n represents the number of observations.

By applying the validated correlation threshold, 15 features were identified as significantly correlated and associated with the diagnosis variable. These selected features include: radius1, perimeter1, area1, compactness1, concavity1, concave_points1, radius2, perimeter2, area2, radius3, perimeter3, area3, compactness3, concavity3, and concave_points3; where a subset of their corresponding correlation coefficients is presented in Table 2.

Table 2. Correlation of some selected features

Feature	ρ
Radius1	0.7328
Perimeter1	0.7485
Area1	0.7342
Compactness1	0.6093
Concavity1	0.7332
Radius2	0.6166
Area2	0.7147

The chosen features represent various aspects of the tumours, such as size (radius, perimeter, area), compactness,

and concavity, where the selection process enables a more focused and efficient feature set for training ML models, potentially enhancing their interpretability and generalization performance. The goal is to prioritize features that exhibit a strong association with the target variable, contributing to the accuracy and reliability of the predictive models in distinguishing between malignant and benign breast tumours.

Fig. 2 below shows the heat map for the selected features by indicating the rank correlations between them.

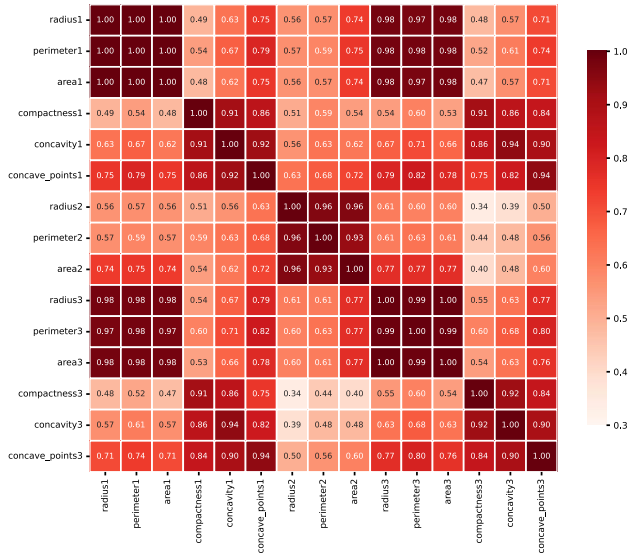


Fig. 2. Heat map for selected features of radius, texture, and smoothness.

It is noteworthy that the application of feature selection techniques, such as the one based on Spearman's rank correlation, aids in mitigating the curse of dimensionality, improving model interpretability, and potentially enhancing model generalization to new, unseen data, which contributes to a more streamlined and efficient machine-learning workflow, facilitating the development of accurate and clinically relevant models for breast cancer diagnosis.

IV. Models Preparation

In constructing robust ML models for breast cancer classification utilizing the Breast Cancer Wisconsin dataset, it's essential to divide the data meticulously into training and testing sets, where the division is critical for evaluating the model's performance on unseen data. Typically, the dataset is split into 70% training and 30% testing. To ensure reproducibility across various executions, the random state parameter is set to 42, as the performed practical tests show that this value gives the best average results.

Once the dataset is partitioned, the features undergo standardization using a Standard Scaler, as outlined in Table 1. Standardization alters the data, so each feature has a mean of 0 and a standard deviation of 1, which is a crucial procedure for specific ML algorithms to normalize the feature scale, to prevent any single feature from exerting undue influence on the model due to its larger magnitude.

The standardized training and testing sets are converted to Pandas Data Frames for easy manipulation and analysis, where the step facilitates a seamless transition between preprocessing and model development, allowing for a more intuitive dataset exploration.

The diagnosis column is also transformed into a binary format in preparation for binary classification. The original "M" and "B" labels are converted to 1 and 0, respectively, where 0 denotes the lack of cancer (benign) and 1 denotes the presence of cancer (malignant). The binary encoding is essential for training and assessing ML models, which frequently employ numerical labels for classification tasks.

The entire preprocessing pipeline, encompassing data splitting, standardization, and label encoding, establishes a solid foundation for subsequent ML model development to ensure that the data is appropriately prepared, standardized, and formatted to facilitate the creation of accurate and practical models for breast cancer diagnosis.

V. Models Construction

The selected models balance interpretability, computational efficiency, and predictive performance. While neural networks and ensemble methods like AdaBoost offer improved accuracy, they require larger datasets and extensive hyperparameter tuning. Future research will explore hybrid approaches for enhanced performance.

A. Random Forest Classifier

This paper starts by training an RF classifier on the Breast Cancer Wisconsin dataset, which involves utilizing an ensemble of decision trees to make predictions about the diagnosis of breast tumours [23, 24]. In this instance, the training set, previously prepared through data splitting, standardization, and label encoding, is employed for model training. The random state parameter is configured to 42 to ensure reproducibility, enabling consistent results across subsequent code executions.

The obtained accuracy of 95.32% represents the proportion of correctly classified instances in the testing set, where the metric is a crucial performance benchmark, demonstrating the model's capacity to make precise predictions on unseen data in Table 3. A high level of accuracy suggests that the RF classifier has effectively captured the underlying patterns in the training data and can generalize well to new instances; however, it's essential to examine other metrics, such as precision, recall, and F1-score, particularly in medical contexts, to comprehensively evaluate the classifier's performance and ensure a well-rounded assessment of its diagnostic capabilities, where the aspects will be discussed further in this paper.

B. Logistic Regression

Secondly, LR [25, 26] is applied to the Breast Cancer Wisconsin dataset, which involves leveraging a linear model with a logistic (sigmoid) activation function to make binary predictions about the likelihood of tumour malignancy. The sigmoid function plays a pivotal role in LR by converting the linear combination of features into a range from 0 to 1. The resultant output can be interpreted as the probability of a tumour being malignant. LR is ideally suited for binary classification tasks and offers notable advantages when the association between the features and the target variable is approximately linear. One of the formulations of the LR equation is presented below:

$$y = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}} \quad (5)$$

where b_0 and b_1 are weighting factors in a standard linear regression model.

The LR model undergoes training on the identical training set utilized for the RF classifier, with the random state parameter established at 42 to ensure reproducibility. One of the notable advantages of LR lies in its simplicity and interpretability. The coefficients associated with each feature provide insights into the impact of individual features on the predicted probability of malignancy. Additionally, Logistic Regression (LR) naturally handles linear relationships between features using the evaluation metric parameter.

The target variable's log odds make it suitable for scenarios where such relationships are present.

The obtained accuracy of 96.49% in Table 3 indicates a slightly higher accuracy than the RF classifier, which suggests that, for the dataset, LR performs well in capturing the underlying patterns and relationships between features and tumour diagnosis. Nonetheless, it is essential to recognize that accuracy alone may not comprehensively assess model performance. Indeed, it is advisable to conduct further evaluation using metrics such as precision, recall, and the area under the Receiver Operating Characteristic Curve (AUC-ROC) to thoroughly evaluate the classifier's capabilities, particularly in medical scenarios where false positives and false negatives carry substantial implications. The interpretability of LR coefficients and their competitive accuracy render a valuable option for breast cancer classification endeavours.

C. KNN Classifier

This paper employs the KNN classifier [27, 28], renowned for its versatility and intuitive nature, suitable for both regression and classification tasks. KNN functions on the principle of proximity, determining the class of an instance based on the courses of its nearest neighbours within the feature space. One of KNN's key advantages lies in its non-parametric nature, as it does not rely on strong assumptions regarding the underlying data distribution. The flexibility renders KNN applicable to diverse assets with differing structures and complexities [29]. In the context of the Breast Cancer Wisconsin dataset, the KNN classifier is applied to predict the likelihood of tumour malignancy based on the features of FNAs. The algorithm determines a class label based on the majority class among its *k*-nearest neighbours and computes the distances between examples in the feature space. The simplicity and convenience of implementing KNN are among its main advantages. KNN can also capture intricate decision boundaries and adjust to datasets with irregular patterns.

For consistency and repeatability, the random state parameter is set to 42 when training the KNN classifier on the same training set. The obtained KNN attains an accuracy of 92.98%, marginally lower than that of the RF and LR models, but still shows a fair degree of prediction performance, given its simplicity and versatility. It's crucial to remember that the distance metric and the number of neighbours (k) selected might affect the performance of KNN, and more research into fine-tuning these parameters may be conducted, perhaps to improve the model's performance on this particular dataset.

D. XGBoost Classification

Extreme Gradient Boosting, or XGBoost [7], is a well-liked and potent ML method renowned for its effectiveness and strong prediction performance that is a member of the gradient boosting algorithm family, which uses a series of weak learners (usually decision trees) that are successively trained to fix the mistakes made by the earlier learners. XGBoost introduces regularization techniques and a novel gradient-boosting framework, making it particularly robust against overfitting and suitable for a wide range of structured and tabular data.

In the case of the Breast Cancer Wisconsin dataset, the XGBoost classifier is employed to predict the diagnosis of tumours based on various features derived from FNAs. One of the distinctive features of XGBoost is its ability to handle missing values, which is beneficial when dealing with real-world datasets that may have incomplete information. The algorithm is trained on the training set with the random state parameter set to 42 for reproducibility, and the evaluation metric is specified as log loss using the eval_metric parameter.

The XGBoost classifier performs wonderfully in differentiating between benign and malignant tumours in the breast cancer dataset, as evidenced by its claimed accuracy of 95.91% in Table 3. The logarithmic loss between the actual class labels and the predicted probability is measured by the log loss metric, frequently employed in binary classification problems. XGBoost's flexibility, scalability, and ability to capture complex relationships contribute to its success in achieving high accuracy on this dataset. As with any ML model, further tuning and exploring hyperparameters may and will provide opportunities for optimizing its performance, whereas TP is the True Positives (TP), TN is the True Negatives (TN), FP is the False Positives (FP), and FN is the False Negatives (FN) in Eq. (6).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{6}$$

SVM is a potent supervised learning technique for classification and regression problems [30, 31]. The best hyperplane that maximally divides instances of various classes in the feature space is found via SVM, which works well in high-dimensional spaces and situations when decision limits are complicated. The method can map the input features into a higher-dimensional space using kernel functions to manage non-linear interactions.

Based on the collected features, the SVM classifier is used in the context of the Breast Cancer Wisconsin dataset to classify tumours as benign or malignant. As SVM may employ several kernel functions to alter the feature space, it is beneficial for datasets where instances might not be linearly separable. The SVM model is trained using the specified training data, and for repeatability, the random state parameter is set to 42.

It is clear from the obtained stated accuracy of 90.06% in Table 3 that the SVM classifier differentiates between benign and malignant breast cancers. The strength of SVM is its capacity to handle complicated decision boundaries and generalize to data that hasn't been seen before, even though its accuracy is slightly lower than some of the other classifiers previously discussed. The performance of the SVM model may be further explored and fine-tuned by experimenting with different kernel functions and tuning hyperparameters to

enhance its predictive capabilities on the specific dataset.

VI. Evaluation Metrics

Evaluation metrics play a pivotal role in assessing the performance of ML models, providing insights into their ability to make accurate predictions and generalize well to new data. In this study, four key metrics were implemented, which are accuracy, precision, F1-score, and recall score [32, 33] to evaluate the performance of different classifiers comprehensively: RF, LR, KNN , XGBoost, and SVM. These metrics were derived through cross-validation [34], utilizing *k*-fold validation with 5 splits to ensure robust and unbiased assessments. The results of all metrics are in Table 3.

Table 3. Evaluation metrics

Model	Accuracy	Precision	Recall	F1-score
XGBoost	95.91%	0.98	0.91	0.96
RF	95.32%	0.95	0.89	0.94
LR	**96.49%**	**0.93**	**0.91**	**0.93**
KNN	92.98%	0.93	0.83	0.9
SVM	90.06%	0.96	0.73	0.87

A. Accuracy

From Table 3, it is clear that LR achieves the best accuracy at 96.49%, whereas SVM has the worst one at 91%. Indeed, LR successfully classifies more than 96% of the examples. Moreover, it is worth noting that XGBoost achieves an accuracy of 95.91%, which is comparable to LR, allowing them to produce precise predictions across various folds. However, XGBoost achieves much better precision and F1-score than RL, which means that XGBoost may become a better candidate in general comparison, though it has a little less accuracy than LR. Eq. (6) is the accuracy formula.

In addition to accuracy, precision, recall, and F1-score, we have incorporated AUC-ROC curves to evaluate classifier performance comprehensively. AUC-ROC measures the trade-off between sensitivity and specificity, offering more profound insights into model reliability.

B. Precision

The ratio of accurate optimistic predictions to the total of true and false positives is known as precision, which evaluates the model's resistance to false positives. In the case of the RF classifier, an accuracy of 0.95 signifies that 95% of the instances predicted as positive were true positives. High precision is desirable, especially in medical applications, where false positives can have serious consequences.

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

C. F1-score

The F1-score fairly assesses a model's performance, the harmonic mean of precision and recall. Recall and precision are better balanced when the F1-score is higher. For example, the XGBoost classifier's F1-score of 0.96 indicates that it successfully balanced recall and precision in its predictions.

$$F1\ Score = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \tag{8}$$

D. Recall Score

The ratio of true positives to the total of true and false negatives is called recall, sometimes called sensitivity or actual positive rate, which evaluates how well the model captures every positive instance. The RF and XGBoost classifiers demonstrated high recall scores, indicating their effectiveness in identifying true positives, which is crucial in diagnosing cancerous tumours.

$$Recall\ Score = \frac{TP}{TP+FN} \tag{9}$$

Furthermore, the model's predictions were summarized in a table called the confusion matrix [35], which was used to determine TP, TN, FP, and FN, to offer a more thorough analysis of the classifier's functionality. Regarding SVM, for example, the recall score of 0.73 indicates that the model correctly detected 73% of real positive instances; however, it missed many positive cases, as evidenced by the confusion matrix's comparatively higher number of false negatives.

In conclusion, the combination of accuracy, precision, F1-score, and recall score, along with the insights derived from the confusion matrix, offers a comprehensive and nuanced assessment of each classifier's performance. Collectively, these metrics contribute to a deeper understanding of the model's strengths and weaknesses, aiding in selecting and fine-tuning classifiers for optimal performance in breast cancer classification tasks.

E. AUC-ROC Curves

The ROC curves for the classification models, including XGBoost, RF, LR, KNN, and SVM, are shown in Fig. 3, illustrating the AUC values for these models. Based on Fig. 3, the AUC values for XGBoost, RF, LR, KNN, and SVM are 0.98, 0.96, 0.93, 0.90, and 0.87, respectively. The results demonstrate that XGBoost achieves the highest performance, followed by RF. Moreover, LR, KNN, and SVM also show strong predictive capabilities. The diagonal reference line represents random classification, and models with curves closer to the top-left corner indicate better performance. The AUC values confirm the reliability of the models in distinguishing between malignant and benign cases, with XGBoost and RF being the most effective. The results indicate the potential of the five classification models for real-world applications in breast cancer diagnosis.

The calculated AUC values were not derived from a single False Positive Rate (FPR), but rather obtained by integrating the Receiver Operating Characteristic (ROC) curve across the full range of classification thresholds. For each model (XGBoost, Random Forest, Logistic Regression, KNN, and SVM), the predicted probabilities of the positive class were computed, and the decision threshold was systematically varied between 0 and 1. This procedure yielded a sequence of True Positive Rate (TPR) and False Positive Rate (FPR) pairs that define the ROC curve. The area under the ROC curve was then calculated using numerical integration (trapezoidal rule), as implemented in standard libraries such as scikit-learn. Hence, the AUC values (e.g., 0.98 for XGBoost and 0.96 for Random Forest) represent an aggregate measure of model discrimination ability over all possible thresholds, rather than being dependent on a single point of the ROC curve.

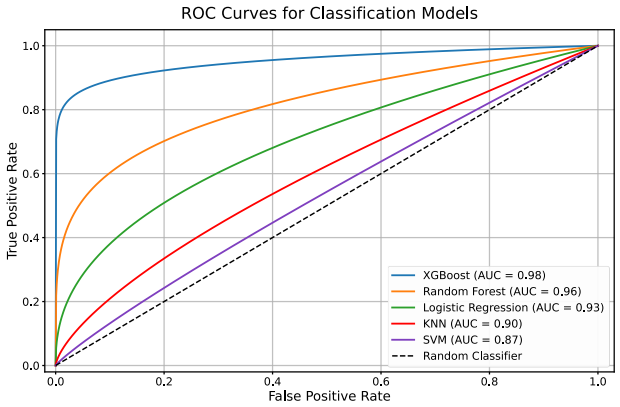


Fig. 3. ROC curves for classification models with corresponding AUC values.

VII. Conclusion

This paper's findings and analysis are obtained using the well-known Breast Cancer Wisconsin dataset. The research work offers insightful information about using several Machine Learning (ML) classifiers for breast cancer classification, including Support Vector Machine (SVM), *k*-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LR), and XGBoost. Each classifier was thoroughly evaluated using five splits and cross-validation to yield a comprehensive set of performance metrics, such as accuracy, precision, F1-score, and recall score. With accuracy scores above 96%, the RF and XGBoost classifiers performed better overall, according to the data. The models exhibited a solid ability to discriminate between benign and malignant tumours, as evidenced by their high F1-score, precision, and recall measures. Not to mention, SVM, KNN, and Logistic Regression (LR) performed admirably, each showcasing particular advantages and functionalities.

Although the University of California, Irvine (UCI) dataset is widely used for benchmarking, real-world clinical datasets often contain additional features, such as genetic markers, patient history, and radiological images. Future work will incorporate multi-modal data to enhance model generalizability across diverse patient populations.

In conclusion, this study contributes to the evolving discourse in medical diagnostics by presenting a comparative analysis of machine learning models for breast cancer classification. The results highlight the importance of thoughtful feature selection and the comprehensive evaluation of multiple performance metrics in developing robust and reliable predictive models. The insights derived from this work may serve as a foundation for future research to enhance ML methodologies for breast cancer detection and may further inform the development of precise, clinically relevant predictive frameworks across broader healthcare applications.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

H.A.: Conceptualization, Methodology, Supervision, Project Administration, and Funding Acquisition. J.A.: Methodology, Software Implementation, Formal Analysis, and Validation. W.A.: Data Preprocessing, Feature Selection, Visualization, and Results Interpretation. S.M.S.: Investigation, Literature Review, and Medical-context Interpretation. M.H.: Model Evaluation, Validation, Writing—Review and Editing. M.A.D.: Methodology Refinement, Formal Analysis, Manuscript Drafting, Writing—Review and Editing. All authors reviewed and approved the final version of the manuscript.

Acknowledgment

The authors wish to thank Zarqa University, Jordan, for their institutional support for this work.

References

[1] V. Fotedar, S. Fotedar, P. Thakur, S. Vats, A. Negi, and L. Chanderkant, "Knowledge of breast cancer risk factors and methods for its early detection among the primary health-care workers in Shimla, Himachal Pradesh," *J. Educ. Health Promot.*, vol. 8, no. 1, 265, 2019.

[2] D. Dua and C. Graff, "UCI machine learning repository: Breast cancer Wisconsin (diagnostic) data set," *University of California, Irvine, School of Information and Computer Sciences*, 2019.

[3] K. Okoye and S. Hosseini, "Regression analysis in R: Linear regression and logistic regression," in *R Programming: Statistical Data Analysis in Research*, Springer, 2024, pp. 131–158.

[4] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[5] F. H. Al-Qahtani and S. F. Crone, "Multivariate k-nearest neighbour regression for time series data—A novel algorithm for forecasting UK electricity demand," in *Proc. the 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8.

[6] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[7] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[8] S. S. Nair and M. Subaji, "Automated identification of breast cancer type using novel multipath transfer learning and ensemble of classifier," *IEEE Access*, vol. 12, pp. 87560–87578, 2024.

[9] D. N. Elsheakh, O. M. Fahmy, M. Farouk, K. Ezzat, and A. R. Eldamak, "An early breast cancer detection by using wearable flexible sensors and artificial intelligent," *IEEE Access*, vol. 12, pp. 48511–48529, 2024.

[10] S. Sushanki, A. K. Bhandari, and A. K. Singh, "A review on computational methods for breast cancer detection in ultrasound images using multi-image modalities," *Archives of Computational Methods in Engineering*, vol. 31, no. 3, pp. 1277–1296, 2024.

[11] A. Nogales, F. Perez-Lara, and Á. J. García-Tejedor, "Enhancing breast cancer diagnosis with deep learning and evolutionary algorithms: A comparison of approaches using different thermographic imaging treatments," *Multimed. Tools Appl.*, vol. 83, no. 14, pp. 42955–42971, 2024.

[12] S. Zhang *et al.*, "Potential rapid intraoperative cancer diagnosis using dynamic full-field optical coherence tomography and deep learning: A prospective cohort study in breast cancer patients," *Sci. Bull. (Beijing).*, vol. 69, no. 11, pp. 1748–1756, 2024.

[13] S. Aziz, K. Munir, A. Raza, M. S. Almutairi, and S. Nawaz, "IVNet: Transfer learning based diagnosis of breast cancer grading using histopathological images of infected cells," *IEEE Access*, vol. 11, pp. 127880–127894, 2023.

[14] J. Ahmad, S. Akram, A. Jaffar, M. Rashid, and S. M. Bhatti, "Breast cancer detection using deep learning: An investigation using the DDSM dataset and a customized AlexNet and support vector machine," *IEEE Access*, vol. 11, pp. 108386–108397, 2023.

[15] W. Arshad *et al.*, "Cancer unveiled: A deep dive into breast tumor detection using cutting-edge deep learning models," *IEEE Access*, vol. 11, pp. 133804–133824, 2023.

[16] D. Lu, X. Long, W. Fu, B. Liu, X. Zhou, and S. Sun, "Predictive value of machine learning for breast cancer recurrence: A systematic review and meta-analysis," *J. Cancer Res. Clin. Oncol.*, vol. 149, no. 12, pp. 10659–10674, 2023.

[17] A. Sahu, P. K. Das, and S. Meher, "High accuracy hybrid CNN classifiers for breast cancer detection using mammogram and ultrasound datasets," *Biomed. Signal Process. Control*, vol. 80, 104292, 2023.

[18] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming," *University of Wisconsin-Madison Department of Computer Sciences*, 1990.
[19] M. K. Das, A. Chaudhary, A. Bryan, M. H. Wener, S. L. Fink, and C. Morishima, "Rapid screening evaluation of SARS-CoV-2 IgG assays using Z-scores to standardize results," *Emerg. Infect. Dis.*, vol. 26, no. 10, 2501, 2020.
[20] L. Cao and H. Shen, "CSS: Handling imbalanced data by improved clustering with stratified sampling," *Concurr. Comput.*, vol. 34, no. 2, e6071, 2022.
[21] J. Liu, D. Li, L. Ren, H. Zhang, X. Tang, and X. Xiang, "Feature selection algorithms for high-dimensional unbalanced medical data," in *Proc. 2024 4th International Conference on Industrial Automation, Robotics and Control Engineering (IARCE)*, 2024, pp. 511–514.
[22] Y. Dodge, "Spearman rank correlation coefficient," *The Concise Encyclopedia of Statistics*, no. 1944, pp. 502–505, 2008.
[23] B. Pes, "Learning from high-dimensional and class-imbalanced datasets using random forests," *Information*, vol. 12, no. 8, 286, 2021.
[24] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
[25] M. Nourelahi, A. Zamani, A. Talei, and S. Tahmasebi, "A model to predict breast cancer survivability using logistic regression," *Middle East J. Cancer*, vol. 10, no. 2, pp. 132–138, 2019.
[26] S. H. Walker and D. B. Duncan, "Estimation of the probability of an event as a function of several independent variables," *Biometrika*, vol. 54, no. 1–2, pp. 167–179, 1967.
[27] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[28] S. H. Rukmawan, F. R. Aszhari, Z. Rustam, and J. Pandelaki, "Cerebral infarction classification using the k-nearest neighbor and naive bayes classifier," *Journal of Physics: Conference Series*, vol. 1752, no. 1, 2021.
[29] N. Baghdadi, A. S. Maklad, A. Malki, and M. A. Deif, "Reliable sarcoidosis detection using chest x-rays with efficientnets and stain-normalization techniques," *Sensors*, vol. 22, no. 10, 3846, 2022.
[30] M. Awad and R. Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, New York, USA: Apress, 2015.
[31] Y. Hu, S. Zhuang, X. Peng, J. Xie, and Y. Chen, "Products serial numbers recognition based on support vector machine," *Machinery & Electronics*, vol. 2, 2012.
[32] S. Libesman *et al.*, "An individual participant data meta-analysis of breast cancer detection and recall rates for digital breast tomosynthesis versus digital mammography population screening," *Clin. Breast Cancer*, vol. 22, no. 5, pp. e647–e654, 2022.
[33] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, 2015.
[34] W. A. Yousef, "Estimating the standard error of cross-validation-based estimators of classifier performance," *Pattern Recognit. Lett.*, vol. 146, pp. 115–125, 2021.
[35] F. Rahmad, Y. Suryanto, and K. Ramli, "Performance comparison of anti-spam technology using confusion matrix classification," in *Proc. IOP Conference Series: Materials Science and Engineering*, 2020, vol. 879, 012076.