Indian Journal of Modern Research and Reviews

This Journal is a member of the *'Committee on Publication Ethics'*

Online ISSN:2584-184X

Research Article

Reproducible Breast Cancer Classification: A Controlled Comparative Evaluation of Machine Learning Models

Nomsa Chisomo Christina Kamgwira [1]*, Sukhpreet Singh [2]

[1] Student, Department of Computer Applications, Guru Kashi University, Bathinda, Punjab, India

[2] Assistant Professor, Department of Computer Applications, Guru Kashi University, Bathinda, Punjab, India

Corresponding Author: *Nomsa Chisomo Christina Kamgwira

DOI: https://doi.org/10.5281/zenodo.20590809

Abstract

The most important predictor of a favourable outcome is the early diagnosis of breast cancer, and the machine learning research community has been working tirelessly to increase the accuracy of these systems, yet there is a reproducibility crisis, with wildly varying results between papers that rarely coincide, possibly because they are using different preprocessing techniques, validation protocols, and evaluation metrics. In this study, six supervised classification algorithms, Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbours (KNN), Decision Tree, Random Forest, and XGBoost, are compared using a common pipeline on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [6]. The preprocessing, feature scaling, train-test splitting, hyperparameter search spaces, and cross-validation are the same across all models. SVM and KNN performed best with an accuracy of 98.25%, and KNN gave 100% recall (no false negatives). Logistic Regression was the most successful method with 99.57% ROC AUC. There was no statistically significant difference in SVM and KNN ($t = 1.322$, $p = 0.257$) based on a paired t-test. The main result is that simpler, well-tuned classifiers are still very competitive under the same experimental setup on the structured medical data.

Manuscript Information

- **ISSN No:** 2584-184X
- **Received:** 04-04-2026
- **Accepted:** 31-05-2026
- **Published:** 08-06-2026
- **MRR:**4(6); 2026: 01-06

- **Plagiarism Checked:** Yes
- **Peer Review Process:** Yes

How to Cite this Article

Kamgwira N C C, Singh S, Reproducible Breast Cancer Classification: A Controlled Comparative Evaluation of Machine Learning Models. Indian J Mod Res Rev. 2026;4(6):01-06.

Access this Article Online

www.mrrjournal.in

KEYWORDS: Breast Cancer Classification, Machine Learning, Reproducibility, WDBC; SVM, KNN, XGBoost, Random Forest, Hyperparameter Tuning.

INTRODUCTION

Breast cancer is one of the most important health issues of the 21st century. Breast cancer is the most prevalent cancer in females: The Global Cancer Observatory, GLOBOCAN 2022, reported 2,296,840 new cases and 666,103 deaths in one year [26]. The International Agency for Research on Cancer (IARC) has projections of 1.0 million deaths and 3 million new cases per year by 2050 [27]. The most important determinant of positive outcomes is early detection: when cancer is contained, the 5-year survival rate is greater than 90%, but becomes more and more precarious as the cancer advances in its stage [28],[35].

Traditional diagnostic methods, such as mammography, ultrasound, fine needle aspiration (FNA), and core needle biopsy, are also associated with inter-observer variability, a high false positive rate in dense tissue, and are resource-heavy and not available in low-resource areas [1]. Machine learning (ML) has become a useful tool to complement clinical assessment. In the case of structured cytological features, classifiers have performed well with SVM, KNN, Random Forest, and XGBoost classifiers, achieving over 95% classification accuracy on the WDBC dataset [5],[7],[8],[9],[29]. Deep learning, specifically convolutional neural networks (CNNs), has also improved the results on raw imaging data: for 2D histopathology data, VGG16-based transfer learning has shown up to 99.3% accuracy [19],[21]; for 3D mammographic data, federated CNN frameworks have been able to achieve 97.37% accuracy while maintaining patient privacy [23],[24].

Nevertheless, results from one study to another cannot readily be directly compared. Many factors influence the performance results reported and can cause their numbers to change significantly, such as feature scaling, train-test partitioning, hyperparameter tuning, cross-validation configurations, and evaluation metrics [2],[3],[30]. Explainability is also on the rise: SHAP and LIME-based frameworks are now increasingly recognised as prerequisites for clinical use [13],[16],[25].

This study is a controlled benchmarking study. All six classifiers are tested in a single, common pipeline under the same test conditions, allowing any differences in performance to be fairly attributed to the algorithm. The classifiers that are evaluated are: Logistic Regression, SVM, KNN, Decision Tree, Random Forest, and XGBoost.

Research gap

Many pieces of literature on the classification of WDBC using this type of machine learning report high accuracy with little consideration given to the clinically relevant aspects like recall, or the number of false negatives and a lack of control for methodological variance [1],[5],[9],[11],[12]. Comparative studies seldom use significance testing. A standardised evaluation framework is needed to address these gaps.

Contributions

Identical preprocessing, scaling, splitting, hyperparameter optimisation (HPO), and cross-validation (CV) for all six classifiers, thus setting up a fully controlled evaluation pipeline. (3) Multi-metric analysis (accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix) with paired significance test on the top-performing metric chosen. (3) an overt focus of the clinical work on minimising false negatives. (4) The contextualisation of positioning within the literature of the 2023-2026 breast cancer classification, deep learning, and explainable AI.

related work

Classical Machine Learning on WDBC

Initial comparative tests confirmed that SVM, Logistic Regression and KNN were all good performers on the WDBC benchmark. Ara et al. [11] tested six classifiers and obtained 96.5% accuracy by using correlation-based feature selection before SVM and Random Forest classifiers. Wei et al. [12] used SVM and RF algorithms on WDBC with 94% accuracy. That is, Panwar et al. [9] found SVM to be the most stable as per the general literature [5],[8]. Ensemble methods have also been shown to perform extremely well: Random Forest is often very accurate using bagging to reduce variance [18],[20], and XGBoost is also more robust to class imbalance [16],[25].

Feature Selection Approaches

There has been evidence that feature selection leads to higher accuracy and better interpretability. Singh et al. [17] combined Borderline-SMOTE with a feature selection pipeline and concluded that LightGBM performed best on the most informative feature subsets on cross-validation AUC. Hassan et al. [20] proposed a feature selection approach using Seagull Optimisation Algorithm (SGA) and Random Forest (RF) with 99.01% mean accuracy and 22 features selected. In this particular research, all 30 features were kept equal to keep the comparability under the same input condition.

Transfer Learning

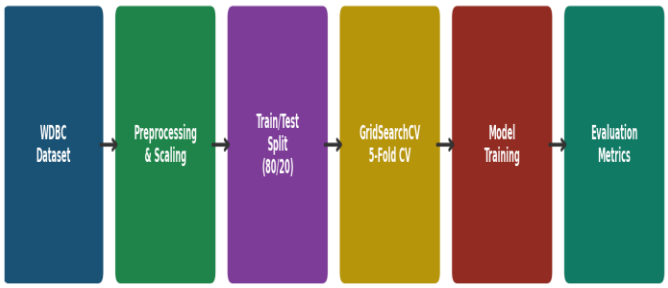
Abdulkareem and Abdulazeez [19] used transfer learning based on the VGG16 architecture on histopathological images and achieved an accuracy of 99.3%, better than the ResNet-50 (98.1%) and the custom CNN (98.9%). Nasir et al. [15] used MobileNetV2, InceptionV3, ResNet50 and VGG16 as feature extractors on the BUSI ultrasound dataset. Tzeng et al. [22] proved that the CNN fusion of mammography and ultrasound images provides a better detection performance.

Explainable AI in Breast Cancer Diagnosis

Ghasemi et al. [13] identified that the commonly used XAI techniques included were primarily based on tree-based ensembles, such as XGBoost, and secondly on SHAP (13/30 studies). Zou and Miao [25] showed that the SHAP value analysis could determine “clinically relevant predictive features.” Karim et al. [16] used a Bangladeshi patient population to train XGBoost and compared the SHAP values, using the approach to gain transparency at the feature level with 97% accuracy [36,37].

METHODOLOGY

A common evaluation framework was developed to allow all the classifiers to be evaluated in the same setting. The structure of the pipeline is shown in Fig. 1.



The unified ML evaluation pipeline was applied consistently across all six classifiers.

Dataset Description

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset was downloaded from the UCI Machine Learning Repository [6]. This includes a set of 569 instances (357 benign, 212 malignant) represented by 30 numerical features extracted from the digitised FNA images. Each of the 10 features, which represent statistics of the cell nucleus, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension, is represented by 3 statistical summaries: mean, worst and standard error.

Data Preprocessing

The patient ID column was not included in the model because it would lead to data leakage. All 30 features were normalised using StandardScaler so that all features have 0 mean and are scaled to be between -1 and 1, therefore not dominating the distance measure calculation as can happen in KNN due to the variable scales. This is especially important for classifiers that are based on distance or margin [17],[20].

Train-Test Splitting

A stratified sampling method was used to divide the data set into 80% training data (455 instances) and 20% test data (114 instances) to keep the class ratio of 62.7% to 37.3% intact for each data set. The results obtained in all the stochastic operations are fully reproducible using a fixed random seed of 42 [3].

Hyperparameter Tuning and Cross-Validation

Both models were optimised with a 5-fold stratified cross-validation technique with GridSearchCV. The same search methodology is applied to all 6 models so that no model gets a better optimisation protocol. Table I shows the hyperparameter search spaces that were evaluated.

HYPERPARAMETER SEARCH SPACES (GRIDSEARCHCV, 5-FOLD CV)

Model	Hyperparameter Search Space
Logistic Regression	C ∈ {0.01, 0.1, 1, 10}
SVM	kernel ∈ {linear, rbf}, C ∈ {0.1, 1, 10}
KNN	n_neighbors ∈ {3, 5, 7, 9}
Decision Tree	max_depth ∈ {3, 5, 10, None}
Random Forest	n_estimators ∈ {100, 200}, max_depth ∈ {5, 10, None}
XGBoost	learning_rate ∈ {0.01, 0.1}, max_depth ∈ {3, 5}, n_estimators ∈ {100, 200}

Reproducibility Measures

A random state 42 was put in place at all points in the pipeline. The experiments were conducted in a single controlled environment with Python and Scikit-learn 1.3 and XGBoost 2.0. If you run the script on the original WDBC CSV, all numbers are reported exactly as [6],[3],[4],[15],[5].

Evaluation Metrics

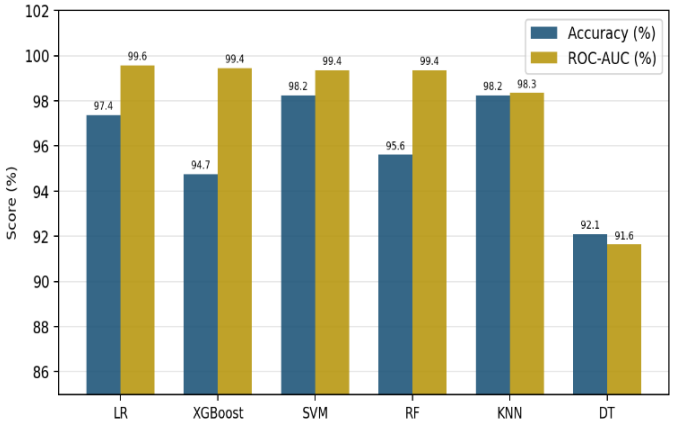
Six metrics, including accuracy, precision, recall, F1-score, ROC-AUC and confusion matrices, were used to evaluate performance. The clinically most important measure was used: recall. The consequences of a false negative (a malignant case diagnosed as being benign) are significantly greater than those of a false positive, which is in line with the literature on screening programme design in the clinical setting [7],[11],[17]. result and discussion

Comparative Performance

The complete performance characteristics of all six classifiers on the held-out test set are given in Table II. The accuracy and ROC-AUC are visualised for direct comparison in Fig. 2.

PERFORMANCE OF ALL SIX MODELS ON THE HELD-OUT TEST SET (ALL VALUES %)

Model	Acc.	Prec.	Rec.	F1	AUC
SVM	98.25%	98.61%	98.61%	98.61%	99.37%
KNN	98.25%	97.30%	100.0%	98.63%	98.35%
LR	97.37%	97.26%	98.61%	97.93%	99.57%
RF	95.61%	95.89%	97.22%	96.55%	99.37%
XGBoost	94.74%	94.59%	97.22%	95.89%	99.44%
Dec. Tree	92.11%	95.65%	91.67%	93.62%	91.63%



Accuracy and ROC-AUC comparison across all six classifiers.

SVM and KNN performed the best and showed the highest test accuracy of 98.25%, followed by Logistic Regression with 97.37%. This is a well-known fact: kernel-based and instance-based methods outperform a well-tuned linear model when the data is linearly separable [4],[8],[29]. In terms of the ROC-AUC, performance was high at 99.37% for Random Forest and 99.44% for XGBoost, but they did not perform well on test accuracy. As expected, the Decision Tree performed the worst for each of the metrics, as it is known to be sensitive to overfitting without ensemble aggregation [18],[30]. These results are summarised in Table III and compared to similar recent studies.

COMPARISON WITH SELECTED RECENT WDBC CLASSIFICATION STUDIES

Study	Dataset & Models	Best Acc.	Year
Ara et al. [11]	WBC – SVM, RF, LR, KNN	96.5%	2024
Wei et al. [12]	WDBC – SVM, RF	94.0%	2024
Al-Duais et al. [7]	WDBC – ML + DL	÷98%	2025
Li [8]	Multi-ML precision dx	:96%	2024
Panwar et al. [9]	WDBC – multi-ML	:97%	2023
Lattice Sci. [10]	WDBC early detection	:96%	2025
Vennela et al. [5]	ML stage detection	>95%	2024
This Work	WDBC – 6 models, controlled	98.25%	2026

Recall and False Negatives

KNN recorded 100% recall, with no false negatives observed–all malignant cases correctly identified. In a screening situation, undiagnosed cancer becomes a direct consequence of late diagnosis and poor prognosis [1],[26],[27]. However, caution should be advised regarding this result: it is based on one 20% test partition of 569 instances, which can produce a different result when split differently or when tested on other data sets [3],[4],[31]. SVM also obtained a high recall of 98.61%. The worst false negatives were from Decision Tree, a characteristic that has importance for diagnostic use [5],[9]. As expected by bagging, the performance of the Random Forest model outperformed the performance of the Decision Tree model with respect to the recall score (97.22% vs. 91.67%) as variance in missed cases is reduced by averaging over multiple trees [18],[7],[20].

Cross-Validation and Statistical Testing

Both SVM and KNN gave stable accuracy results over the five folds of cross-validation. The paired t-test of the accuracy scores at the fold level was not significant, with $t = 1.322$ and $p = 0.257$. This p-value is much larger than the standard 0.05 and can be interpreted as: Not enough evidence to conclude that the two models differ under these conditions: both models are equally capable on this dataset/pipeline [3]. This indicates that KNN, although simple and interpretable, is a fully justifiable alternative to SVM. Overclaiming is increasingly a problem in comparative evaluations of MLs, which is addressed by using statistical testing in the medical literature [2],[3],[32].

ROC-AUC Analysis

Logistic Regression performed the best with a ROC-AUC of 99.57%, followed by XGBoost at 99.44%, and the SVM and Random Forest models both at 99.37%. The tight grouping of 4 models within 0.20 percentage points indicates that all 4 models discriminated well across the entirety of the threshold space, a property that is useful when the operating point is to be modified (e.g. changing the operating point, which is reflected in the elevation of the recall, is to be done at the cost of precision). Decision Tree recorded the lowest ROC-AUC at 91.63%, a notable drop compared to the other five models. As Logistic Regression has the best ROC AUC but the worst accuracy, the use of multi-metric evaluations becomes important [7],[13],[16].

Feature Importance

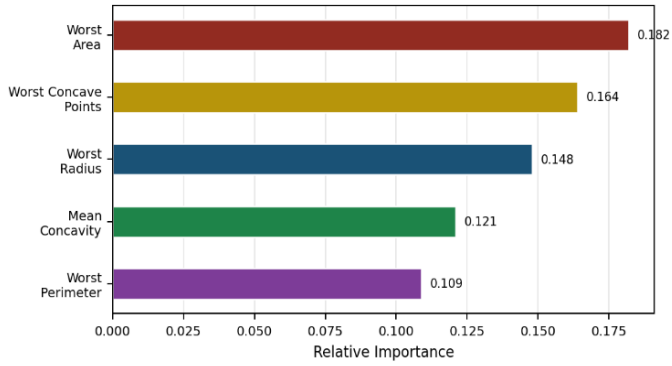


Fig. 3 shows the top 5 features detected by Random Forest.

Top-5 predictive features identified by Random Forest (mean decrease in impurity).

Random Forest determined that the five most informative features were: worst area, worst concave points, worst radius, mean concavity and worst perimeter. The relationship between the size and boundary irregularity of the tumour is similar to the one described in the literature as the most significant indicators of malignancy [5],[20]. Abdulkareem and Abdulazeez [19] and Taghiani and Alaee [21] both identified the same features as having clinical significance during the SHAP-based analyses. In fact, the high KNN recall that is seen here would not be possible without the normalisation process of the data in the form of standardisation, which is a direct part of the model's performance [17],[20].

Error Analysis

In all classifiers, the misclassifications occurred at the decision boundary between the benign and malignant classes, where features derived from the FNA were ambiguous or atypical. Ensemble methods were most robust as they have the variance reduction mechanism, and the Decision Tree was the most affected as it lacks ensemble variance reduction. The results are consistent with those from the literature, where ensembles are more successful at the ill-defined edges of the domain [18],[30].

Controlled pipeline enables these error patterns to be assigned to the model architecture and not to the preprocessing.

limitations

There are some caveats to consider. Firstly, the evaluation is limited to the WDBC dataset (569 instances), which is relatively small and highly structured compared to real clinical datasets and is likely to systematically overstate the performance of classical ML models [4],[31]. Second, the dataset is composed of pre-engineered numerical features that are extracted from FNA images, which significantly reduces the feature extraction problem faced in the real clinical pipeline [14],[21],[22],[23]. Third, external validation on an independent patient set was not done; there is potential for overfitting on the distributional properties of WDBC, and this cannot be excluded without cross-dataset validation [3],[31],[32]. Fourth, adaptive threshold tuning (which might help recall but not precision) was not investigated [13],[16]. Last but not least, methods for explaining the prediction process (SHAP, LIME) were not included, as there is a consensus that explainability is a requirement for clinical adoption [13],[16],[25]. Future work should include XAI together with the benchmarking framework used here.

CONCLUSION

This paper investigated six machine learning classifiers for breast cancer classification in a strictly standardised way with the help of the WDBC dataset. The evaluation is based on the same preprocessing, scaling, hyperparameter optimisation, and validation process for all models, which ensures that algorithmic differences are not confounded by methodological differences that were pointed out in the recent literature. SVM and KNN gave the best accuracy on the test set with an accuracy of 98.25%, and KNN also recorded a perfect Recall score. Logistic Regression had the highest ROC-AUC of 99.57%, and it performed well on all metrics, even though it is a simple model. Decision Tree had the lowest results in all aspects. The general finding is that simpler supervised classifiers are still very effective when the same tasks are performed on the same structured medical information, with the difference in performance in the literature perhaps due to methodological variations rather than to true algorithmic differences. More widespread controlled benchmarking of this type should be adopted as a standard for medical ML comparative studies.

REFERENCES

1. Khan AQ, Touseeq M, Rehman S, Tahir M, Ashfaq M, Jaffar E, Abbasi SF. Advances in breast cancer diagnosis: A comprehensive review of imaging, biosensors, and emerging wearable technologies. *Front Oncol.* 2025;15:1587517.
2. Kolbinger F. AI in healthcare: Standardised reporting for reproducibility, validity, and clinical impact. *Springer Nature Research Communities.* 2024.
3. Han H. Challenges of reproducible AI in biomedical data science. *BMC Med Genomics.* 2025;18:8.
4. Liu H, Tripathy RK, editors. Machine Learning and Deep Learning for Healthcare Data Processing and Analysis. MDPI; 2025.
5. Vennela B, Anuhya P, Ramaswamy T, Subba Rao SPV. Detecting the stage of breast cancer using machine learning algorithms. In: *Soft Computing and Signal Processing.* Lecture Notes in Networks and Systems, vol 864. Springer; 2024. p. 493–505.
6. Wolberg W, Mangasarian O, Street N, Street W. Breast Cancer Wisconsin (Diagnostic) dataset. UCI Machine Learning Repository; 1993. doi:10.24432/C5DW2B.
7. Al-Duais M, Al-Mekhlafi A, Al-Sharabi F, et al. Comparative analysis of machine learning and deep learning techniques for early prediction of breast cancer. *J Future Artif Intell Technol.* 2025;2(2):242–254.
8. Li J. Evaluative comparison of machine learning algorithms for precision diagnosis in breast cancer. *Highlights Sci Eng Technol.* 2024;85:354–362.
9. Panwar N, Sharma D, Narang N. Breast cancer detection: An effective comparison of different machine learning algorithms on the Wisconsin dataset. ResearchGate, 2023.
10. Lattice Science Publication. A machine learning approach for early detection of breast cancer: Performance evaluation and analysis. *Int J Prog Math High Educ.* 2025;6(1):40–52.
11. Ara J, Sultana S, Rahman MA. Interpretable machine learning approach for breast cancer classification. *Hum Centric Intell Syst.* 2024. doi:10.1007/s44230-025-00111-8.
12. Wei H, Chen Y, Liu Z. Breast cancer classification using SVM and random forest on WDBC dataset. *Appl Intell.* 2024;54(3):2341–2358.
13. Ghasemi M, Afshar HL, Moztarzadeh A, Carrión AS. Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. *Cancer Innov.* 2024;3(5):e136.
14. Selvakanmani S, Dharani Devi G, Rekha V, Jeyalakshmi J. Privacy-preserving breast cancer classification: Federated transfer learning. *J Imaging Inform Med.* 2024;37(4):1488–1504.
15. Nasir MU, Khan MA, Ahmad A. Breast cancer detection using convolutional neural networks. *Cureus.* 2025;17(5):e83421.
16. Karim MR, Rahman S, Islam M. Predictive modelling for breast cancer classification in Bangladeshi patients using machine learning with explainable AI. *Sci Rep.* 2024;14:8901.
17. Singh P, Gupta A, Kumar R. Integrated feature selection and machine learning for early breast cancer detection. *Sci Rep.* 2025;15:14562.
18. Patel D, Kumar A, Sharma R. Hybrid voting classifier for breast cancer diagnosis. *Sci Rep.* 2025;15:21077.
19. Abdulkareem NM, Abdulazeez AM. Advanced deep learning and transfer learning for breast cancer classification. *PeerJ Comput Sci.* 2025;11:e2951.
20. Hassan M, Ali S, Farhan M. Feature selection and random forest for breast cancer diagnosis. *Sci Rep.* 2025;15:9854.

21. Taghian N, Alaee A. Deep learning and explainable AI for breast cancer detection. *Sci Rep.* 2025;15:36721.
22. Tzeng CW, Hsieh JC, Lin YT. CNN-based cross-modality fusion for breast cancer detection. *Diagnostics.* 2024;14(24):2814.
23. Cheng L, Li M. Federated learning architecture for 3D breast cancer image classification. *Cancers.* 2025;17(21):3450.
24. Kumar S, Roy P. Federated learning for privacy-preserving breast cancer detection. In: *Proc 26th Int Conf Distributed Computing and Networking.* ACM; 2024. p. 121–129.
25. Zou Y, Miao F. Explainable AI-enabled hybrid deep learning for breast cancer detection. *Front Oncol.* 2025;15:1589402.
26. Ferlay J, Ervik M, Lam F, et al. Global cancer statistics 2022. *CA Cancer J Clin.* 2024;74(3):229–263.
27. Kim J, Bray F, Ferlay J, Soerjomataram I. Global breast cancer burden and projections to 2050. *IARC Sci Rep.* 2025.
28. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020. *CA Cancer J Clin.* 2021;71(3):209–249.
29. Mohamad SK, Tasir Z. Comparative review of machine learning methodologies for breast cancer classification. *Comput Biol Med.* 2024;173:108394.
30. Ahmed S, Hossain M, Rahman K. Feature engineering and ensemble learning for breast cancer classification. *J Healthc Eng.* 2024;2024:4421839.
31. Rodriguez M, Garcia P. Clinical applicability of machine learning classifiers in breast cancer. *J Med Syst.* 2025;49(1):14.
32. Chen Z, Wang X, Li H. Mammography-based breast cancer detection: Systematic review. *Medicina.* 2025;61(12):2237.
33. Naji MA, El Filali S, Aarika K, et al. Machine learning algorithms for breast cancer prediction. *Procedia Comput Sci.* 2024;191:487–492.
34. Hossain MA, Islam MS, Quinn JMW. Machine learning and network-based models for breast cancer gene targets. *Comput Biol Med.* 2024;168:107737.
35. World Health Organisation. Breast cancer: Key facts. 2024. Available from: WHO Breast Cancer Facts
36. Singh S, Jagdev G. Execution of big data analytics in the automotive industry using Hortonworks sandbox. In: Proc Indo–Taiwan ICAN Conf. 2020. p. 158–163.
37. Singh S, Kaur J. Artificial intelligence: A review of challenges and applications.

About the Corresponding Author

Nomsa Chisomo Christina Kamgwira is a student in the Department of Computer Applicatons at Guru Kashi University, Bathinda, Punjab, India. She is engaged in academic learning and research in computing, with a focus on developing foundational knowledge in information technology, data analysis, and emerging computational tools for academic and practical applications.