DOI: https://doi.org/10.57233/ijsgs.v12i2.1084 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12 (2) WEBSITE: https://fugus-ijsgs.com.ng

INTERNATIONAL JOURNAL OF SCIENCE FOR GLOBAL SUSTAINABILITY
(A PUBLICATION OF FACULTY OF SCIENCE, FEDERAL UNIVERSITY GUSAU, NIGERIA)

Hybrid Deep Learning with Explainable AI for Breast Cancer Prediction Using the WBCD Dataset

Abdullahi, Abdulmalik and Ahmad, Bello
Department of Computer Science, Abdu Gusau Polytechnic, Talata Mafara, Zamfara State, Nigeria.

Corresponding Author's Email: belomada38@gmail.com

Received on: October, 2025 **Revised and Accepted on:** November, 2025 **Published on:** December, 2025

ABSTRACT

Breast cancer remains a leading cause of morbidity and mortality among women worldwide, creating demand for diagnostic models that are both accurate and interpretable. While machine learning and deep learning methods have achieved strong results on benchmark datasets, such as the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, limitations including class imbalance, overfitting, and lack of interpretability hinder their clinical use. This study proposes a hybrid deep learning framework that integrates Convolutional Neural Networks (CNN) with Gated Recurrent Units (GRU), supported by class weighting and stratified sampling to address class imbalance, and SHAP (SHapley Additive exPlanations) to enhance model transparency. Preprocessing involved KNN imputation for missing values, label encoding, and Min-Max normalization. The model was evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. On the WBCD dataset, the CNN–GRU achieved 95.61% accuracy, 100.0% precision, 88.10% recall, 93.67% F1-score, and 0.9944 AUC-ROC. The absence of false positives underscores the model's conservative boundary, though five false negatives highlight sensitivity trade-offs. SHAP-based global and local explanations identified concave points, perimeter, and radius as the most influential features, consistent with clinical knowledge and providing insight into both correct and incorrect classifications. Unlike prior studies that reported higher accuracies but offered limited interpretability, the proposed framework strikes a balance between predictive strength and clinical trustworthiness. These findings position CNN–GRU with explainable AI as a promising tool for decision support in breast cancer diagnosis and point to future directions in multimodal integration, federated learning, and prospective clinical validation.

Keywords: Breast cancer, CNN–GRU, Explainable AI, Class imbalance, Medical diagnostics.

1.0 INTRODUCTION

Breast cancer remains one of the leading causes of morbidity and mortality among women worldwide. Early detection significantly improves treatment outcomes and survival (Begum et al., 2024). Artificial intelligence (AI), especially machine learning (ML) and deep learning (DL), has shown strong promise for automating breast cancer prediction and supporting clinicians in diagnostic imaging and pathology (Sharafaddini *et al.,* 2025). However, translating high-performing models from research into clinical practice requires more than high accuracy on benchmark datasets; it requires robustness, fairness and interpretability.

Many recent studies report very high performance on benchmark datasets, sometimes exceeding 97–99% accuracy but they often share common methodological weaknesses. For instance, Al Mamun *et al.* (2025), an anchor study in this literature, reports a neural network accuracy of 98.13% on the Wisconsin Breast Cancer Diagnostic (WBCD) dataset. While impressive, that work (like several others) exhibits several limitations that undermine clinical trust and generalizability: it does not explicitly address class imbalance, provides little or no model interpretability, lacks external or repeated cross-validation to demonstrate robustness, and offers limited detail on preprocessing and hyperparameter optimization. Such omissions raise the risk of optimistic, dataset-specific results driven by overfitting, and they deny clinicians clear, case-level explanations that are essential for adoption.

These issues summarize the core problem this study addresses: high reported accuracy in prior work does not automatically translate to clinical usefulness because of (a) dataset bias and class imbalance that harm sensitivity for malignant cases, (b) overfitting and insufficient validation that reduce external validity, and (c) the black-box nature of many DL models that prevents clinician verification and trust. To close this gap, the present study develops a hybrid Convolutional Neural Network–Gated Recurrent Unit (CNN–GRU) model trained on the WBCD dataset and explicitly emphasizes fairness and transparency. We apply class weighting and stratified sampling to mitigate imbalance, adopt reproducible preprocessing (including KNN imputation and Min–Max scaling), and use SHapley Additive exPlanations (SHAP) to provide both global and local interpretability.

DOI: https://doi.org/10.57233/ijsgs.v12i2.1084 ISSNp: 2488-9229; ISSNe: 3027-1118

The scope of this work is intentionally bounded: it focuses on tabular cytological features from the WBCD dataset as a proof-of-concept rather than a full clinical deployment. Limitations include dataset size and curation, reliance on a single modality, manual hyperparameter tuning, and the computational cost of SHAP, which in this study constrains explanations to selected subsets for feasibility. Contributions. This paper makes three main contributions:

A CNN–GRU hybrid architecture tailored to capture local feature patterns and sequential dependencies in tabular cytological data.
A fairness-focused training pipeline using class weighting and stratified sampling to improve sensitivity toward the minority (malignant) class.

Integration of SHAP-based explainability to produce clinician-friendly global and local feature attributions, improving transparency and supporting potential clinical adoption.

Recent research in breast cancer prediction has compared traditional machine learning (ML) models with deep learning (DL) approaches, with increasing emphasis on class imbalance and model explainability. Classical ML methods such as Random Forests, Support Vector Machines (SVM), Gradient Boosting Machines (GBMs), decision trees, and logistic regression remain widely used because of their interpretability, relatively low data requirements, and competitive performance on small, tabular datasets. For instance, Arravalli *et al.* (2024) applied SVM, Random Forest, and Gradient Boosting to clinical datasets and integrated SHAP for interpretability, achieving accuracies in the mid-90s. However, such models typically require extensive feature engineering and are less effective when data are noisy or subtle patterns dominate (Ansari *et al.*, 2025; Islam *et al.*, 2024).

The anchor study by Al Mamun *et al.* (2025) evaluated several ML classifiers on the WBCD dataset and reported that a neural network achieved 98.13% accuracy, 98.21% precision, 98.00% recall, and an AUC-ROC of 0.9992. While these results are impressive, the study did not address class imbalance or integrate any form of explainability. This omission illustrates a recurring limitation in much of the literature: high predictive accuracy is reported, but without fairness strategies or transparent explanations, clinical utility remains limited.

Deep learning methods, particularly convolutional neural networks (CNNs) and hybrid models (CNN–LSTM, CNN–GRU), have gained prominence in image-based prediction tasks. Transfer learning and attention-based architectures further boost performance on large imaging datasets (Alom *et al.*, 2025; Ansari *et al.*, 2025). For example, Alom *et al.* (2025) applied DenseNet121 with transfer learning and class weighting, achieving strong performance on histopathology and ultrasound images. However, these approaches demand larger datasets, higher computational resources, and careful regularization to avoid overfitting. Similarly, Miao and Zou (2025) used image augmentation, class balancing, and dropout to reduce overfitting in CNN-based fusion models, improving recall for malignant cases. Yet, even in these studies, minority class performance (sensitivity for malignancies) is not always prioritized or clearly reported.

Another persistent gap is interpretability. While high accuracy is necessary, clinicians require insights into why models make specific predictions. Explainable AI (XAI) methods such as SHAP, LIME, and Grad-CAM are increasingly adopted (Ghasemi *et al.*, 2024). For instance, SHAP has become the most widely used model-agnostic method for tabular data, while Grad-CAM is common in imaging tasks. However, most studies still emphasize global explanations without offering local, case-specific insights. Few works compare multiple XAI techniques or evaluate consistency of explanations, and reproducibility is often limited by inconsistent preprocessing pipelines.

Comparisons between ML and DL suggest that while DL models often achieve superior performance, particularly on imaging datasets, ML remains competitive on smaller, structured datasets like WBCD. ML models are faster to train, easier to interpret, and less prone to overfitting under limited data conditions. This is reflected in the findings of Al Mamun *et al.* (2025), whose high-performing neural network lacked explicit imbalance handling and interpretability a limitation common to many studies in the field.

In summary, existing research demonstrates that both ML and DL methods can achieve strong results, but persistent gaps remain: (1) insufficient focus on class imbalance and fairness, (2) reliance on accuracy as the primary metric without detailed reporting of minority-class recall, (3) limited integration of explainability, especially local explanations, and (4) inconsistent preprocessing and validation practices that hinder reproducibility. These gaps motivate the present study, which combines CNN–GRU modeling with class weighting, stratified sampling, and SHAP explainability to deliver a more transparent and clinically relevant framework.

DOI: https://doi.org/10.57233/ijsgs.v12i2.1084 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12 (2) WEBSITE: https://fugus-ijsgs.com.ng

Table 1.1. Summary of Recent Studies on Machine Learning and Deep Learning for Breast Cancer Prediction

Author(s) & Year	Dataset	Model(s) Applied	Imbalance Handling	Explainability Used	Identified Gaps
Ansari *et al.* (2025)	WBCD & small clinical datasets	SVM, Random Forest, Gradient Boosting	Not reported	SHAP, LIME	Focus on accuracy; imbalance not addressed
Alom *et al.* (2025)	Histopathology & ultrasound	CNN + DenseNet121 (transfer learning)	Class weighting, data augmentation	Grad-CAM	High performance but limited interpretability beyond visualization
Bakr *et al.* (2025)	Mammogram dataset	Hybrid CNN–LSTM	Oversampling & augmentation	None	Possible overfitting; no XAI applied
Ghasemi *et al.* (2024)	Multiple datasets (review)	Review of ML/DL with XAI	Reviewed balancing methods	SHAP most common	Few studies combine global + local explanations; inconsistent validation
Islam *et al.* (2024)	WBCD	Random Forest, Logistic Regression, ANN	None	None	Results reported without fairness/recall focus
Hybrid XAI DL Model (2025)	Ultrasound dataset	CNN fusion model	Image augmentation	SHAP & Grad-CAM	Limited to imaging data, not tabular datasets
Al Mamun *et al.* (2025)	WBCD	Neural Network, SVM, Random Forest, Decision Trees	None	None	High accuracy but ignored imbalance and interpretability

2.0 METHODOLOGY

This study adopts an experimental research design using secondary data from the Wisconsin Breast Cancer Diagnostic (WBCD) dataset. The approach is quantitative, focusing on the implementation and evaluation of a hybrid deep learning model for breast cancer prediction. The framework integrates three major components: (i) preprocessing and imbalance handling, (ii) development of a CNN–GRU architecture, and (iii) application of SHAP for explainability. Figure 3.1 presents the conceptual framework guiding the methodology.

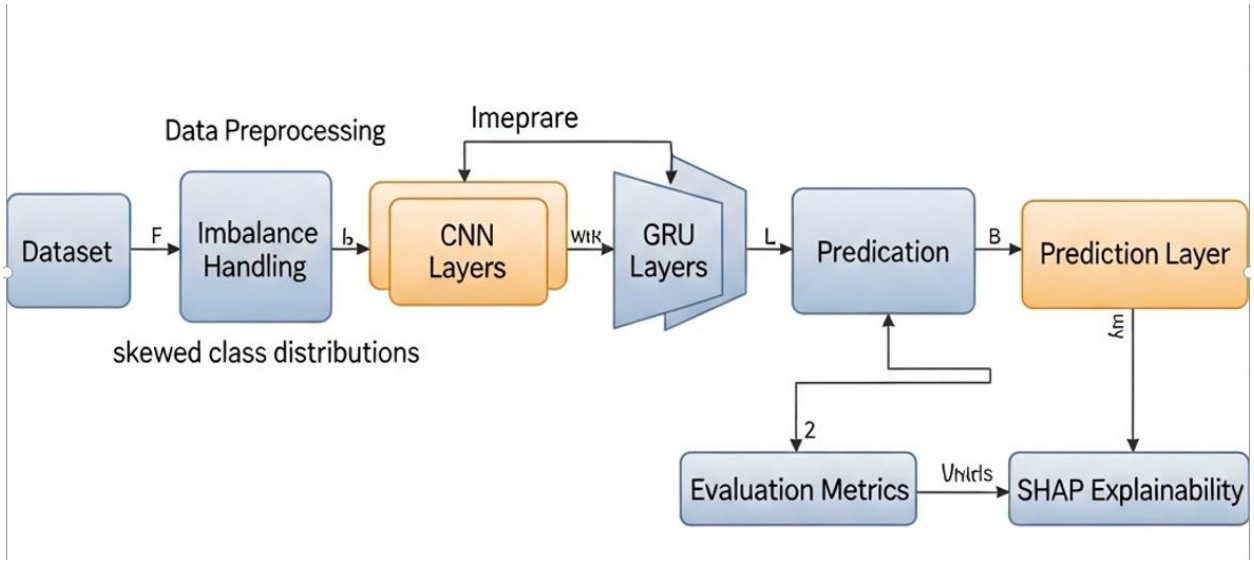

Figure 2.1: Conceptual Framework of the Study

2.1 Dataset

DOI: https://doi.org/10.57233/ijsgs.v12i2.1084 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12 (2) WEBSITE: https://fugus-ijsgs.com.ng

The WBCD dataset, publicly available from the UCI Machine Learning Repository, contains 569 patient records with 30 numerical features describing cell nuclei morphology from fine-needle aspiration cytology. Features capture mean, standard error, and worst values of attributes such as radius, perimeter, texture, concavity, and symmetry. The dataset consists of 357 benign cases (63%) and 212 malignant cases (37%), reflecting a moderate class imbalance that can bias models toward benign predictions if not addressed. This dataset has been widely adopted in the literature (Islam et al., 2024; Ansari et al., 2025), allowing direct comparison with existing works.

Table 2.1: Summary of the WBCD Dataset

Attribute	Description	Count / Range
Total Records	Number of patient samples	569
Features	Numerical features derived from FNA of cell nuclei	30
Class Labels	Diagnostic outcome	Benign = 357 (63%), Malignant = 212 (37%)
Feature Categories	Mean, SE, Worst values of 10 features	30 total
Data Imbalance	Distribution	Benign ~63%, Malignant ~37%

2.2 Preprocessing

Data preprocessing ensures the dataset is clean, consistent, and suitable for deep learning models. Missing values were handled using the K-Nearest Neighbor (KNN) Imputer, which estimates missing entries by averaging the values of the k nearest neighbors in the feature space. The target label "diagnosis" (Benign, Malignant) was then converted into binary numeric form using Label Encoding, enabling the variable to be processed by the model. To address the varying ranges of the features, Min–Max normalization was applied to scale all feature values into the [0,1] interval, ensuring uniformity and preventing features with larger magnitudes from dominating the learning process. Finally, the dataset was split into training and testing sets using stratified sampling, which preserved the ratio of benign to malignant cases across both subsets and ensured more representative evaluation.

2.3 Class Imbalance Handling

Class imbalance arises when one class is underrepresented relative to another. In the WBCD dataset, malignant cases are fewer than benign cases. To mitigate bias, class weighting was employed. The weight for each class w_c was calculated. This increases the penalty for misclassifying malignant cases, thereby improving recall and reducing false negatives. Additionally, stratified sampling during the train-test split reinforced balance by preserving class proportions in both subsets.

2.4 Model Architecture

The study designed a hybrid CNN–GRU architecture that combines convolutional layers for feature extraction with recurrent layers for sequential learning. The convolutional layers extract local feature patterns from input sequences using the convolution operation (Equation 1), while ReLU activation (Equation 2) introduces non-linearity and improves model generalization. Dimensionality reduction is achieved through max-pooling (Equation 3), which selects the maximum value within a defined window. To prevent overfitting, dropout layers randomly deactivate neurons during training.

$$(X * K)_i = \sum_{j=0}^{m-1} K_j \cdot X_{i+j} \quad (1)$$

$$f(x) = \max(0, x) \quad (2)$$

$$y_i = max(x_i, x_{i+1}, \ldots\ldots, x_{i+p-1}) \quad (3)$$

Sequential dependencies are captured by Gated Recurrent Units (GRUs), which update hidden states using an update gate, reset gate, and candidate activation as shown in the GRU equations. Here, $x_t$represents the input at time t, h_{t-1} the previous hidden state, σ the sigmoid activation, and $\odot$ denotes element-wise multiplication. Fully connected dense layers then compute weighted sums of their inputs before passing them through activation functions, with ReLU applied in hidden layers and sigmoid used in the output layer to produce the probability of malignancy. The final output probability is modeled using the logistic sigmoid function. The network was trained with the Adam optimizer and binary cross-entropy loss, and an Early Stopping callback was employed to halt training if validation loss failed to improve, thereby reducing the risk of overfitting.

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (update\ gate) \quad (4)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (resete\ gate) \quad (5)$$

$$\tilde{h}_t = tanh(W_z x_t + U_h(r_t \odot h_{t-1}) + b_h) \quad (candidate\ activation) \quad (6)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (new\ hidden\ state) \quad (7)$$

DOI: https://doi.org/10.57233/ijsgs.v12i2.1084 ISSNp: 2488-9229; ISSNe: 3027-1118

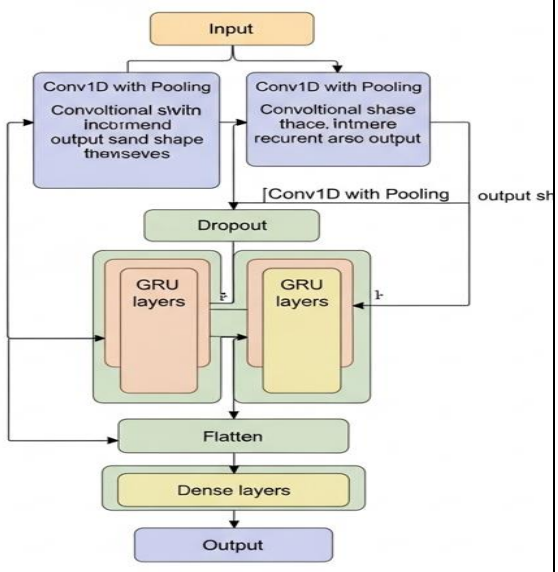

Figure 2.3: CNN–GRU hybrid architecture diagram

2.5 Model Summary

The implemented model architecture is further detailed in Table 3.2, which presents the output shape and parameter count at each layer.

Table 2.2: CNN–GRU Model Summary

Layer (Type)	Output Shape	Param #
Conv1D (conv1d)	(None, 30, 32)	128
MaxPooling1D	(None, 15, 32)	0
Conv1D (conv1d_1)	(None, 15, 64)	6,208
MaxPooling1D (pooling1)	(None, 7, 64)	0
Dropout	(None, 7, 64)	0
GRU (gru)	(None, 7, 64)	24,960
GRU (gru_1)	(None, 32)	9,408
Flatten	(None, 32)	0
Dense (dense)	(None, 64)	2,112
Dense (dense_1)	(None, 32)	2,080
Dense (dense_2)	(None, 16)	528
Dense (dense_3)	(None, 1)	17

Total parameters: 45,441 (177.50 KB)
Trainable parameters: 45,441 (177.50 KB)
Non-trainable parameters: 0 (0.00 B)
The model was trained using the Adam optimizer and binary cross-entropy loss. Early stopping was used to halt training when validation loss failed to improve, preventing overfitting. Figure 3.2 illustrates the model design.

2.6 Evaluation Metrics

Model performance was evaluated using several standard metrics derived from the confusion matrix. Accuracy (ACC) measures the proportion of correctly classified samples relative to the total number of samples, defined as equation 8. While widely reported, accuracy can be misleading under class imbalance since models may achieve high scores by over-predicting the majority class. To complement this, Precision (positive predictive value) was calculated as equation 9, reflecting the proportion of predicted malignant cases that were truly malignant; high precision indicates a low false-positive rate, which is important to avoid unnecessary anxiety and invasive tests in benign patients. Recall (sensitivity or true positive rate), defined as equation 10, measures the proportion of actual malignant cases correctly identified and is critical in medical contexts, as false negatives carry severe consequences. To balance precision and recall, the F1-Score was computed as equation 11, a harmonic mean particularly useful for imbalanced datasets. Finally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was used to evaluate discriminative ability across thresholds; here, the false positive rate (FPR) is given by equation 12, and the AUC ranges from 0.5 (random guessing) to 1.0 (perfect classification), providing a robust measure of how well the model distinguishes between benign and malignant cases.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$F1 - score = 2 \times \frac{Precision + Recall}{Precision \times Recall} \quad (11)$$

$$FPR = \frac{FP}{FP + TN} \quad (12)$$

2.7 Explainability

To overcome the black-box nature of deep learning, SHAP (SHapley Additive exPlanations) was applied to interpret predictions. SHAP provides global feature importance and local explanations for individual patients. The study generated:

Summary plots: showing overall feature contributions.
Bar plots: highlighting mean absolute SHAP values.
Force plots: illustrating how features drive predictions for specific cases.
These explanations enhance transparency by showing not only *what* the model predicted but also *why*.

3.0 RESULTS AND DISCUSSION

DOI: https://doi.org/10.57233/ijsgs.v12i2.1084 ISSNp: 2488-9229; ISSNe: 3027-1118

This section presents the outcomes of the CNN–GRU model, interprets performance metrics in clinical and methodological terms, and situates the findings within prior research. The model, trained on the WBCD dataset using stratified sampling and class weighting, achieved an overall accuracy of 95.61%**,** precision of 100%, recall of 88.10%, F1-score of 93.67%, and an AUC-ROC of 0.9944. The confusion matrix on the holdout test set (TP = 72, FP = 0, FN = 6, TN = 36) reveals a classifier that is highly discriminative overall, but with a trade-off between zero false positives and a small number of missed malignant cases. Accuracy reflects that about 109 of 114 test samples were correctly classified, but alone can be misleading under class imbalance. Precision at 100% indicates every malignant prediction corresponded to a true malignant case, eliminating false alarms, a critical property in avoiding unnecessary biopsies and patient anxiety. Recall, however, was 88.10%, with six malignant cases missed; in clinical settings, such false negatives carry greater risk than false positives, underscoring the need to enhance sensitivity. The F1-score of 93.67% balances precision and recall, confirming strong but improvable performance, while the AUC-ROC of 0.9944 indicates excellent discriminative ability across thresholds.

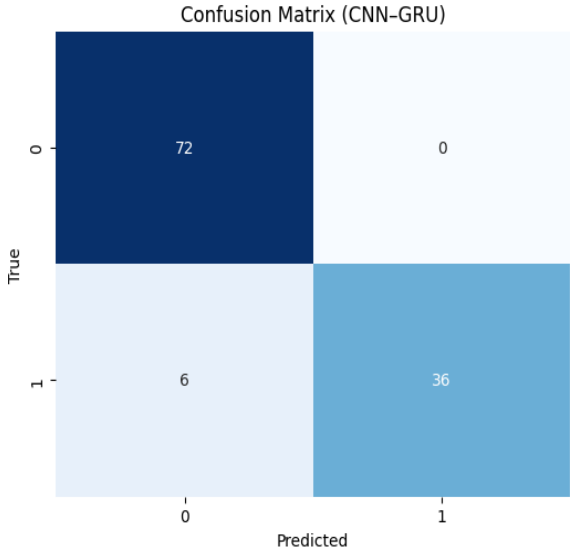

Figure 3.1: Confusion Matrix

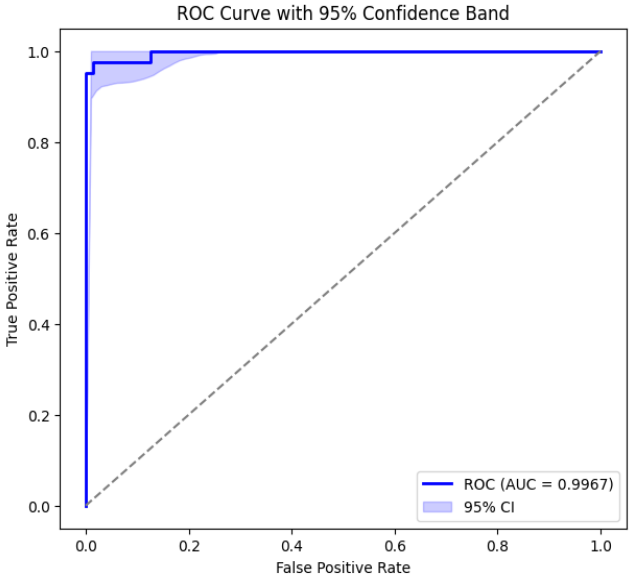

Figure 3.2: ROC Curve.

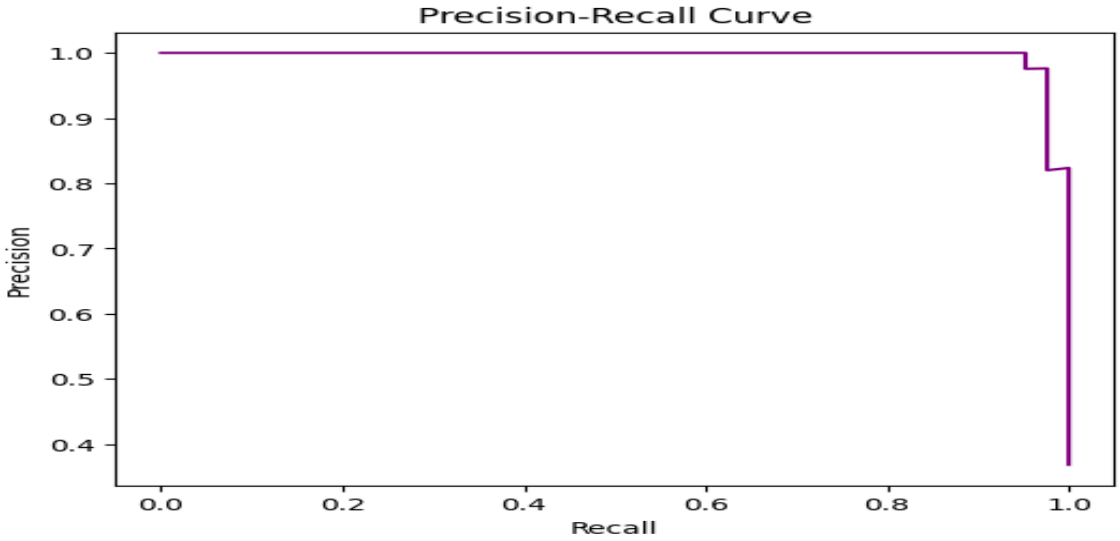

Figure 3.3: Precision–Recall Curve

DOI: https://doi.org/10.57233/ijsgs.v12i2.1084 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12 (2) WEBSITE: https://fugus-ijsgs.com.ng

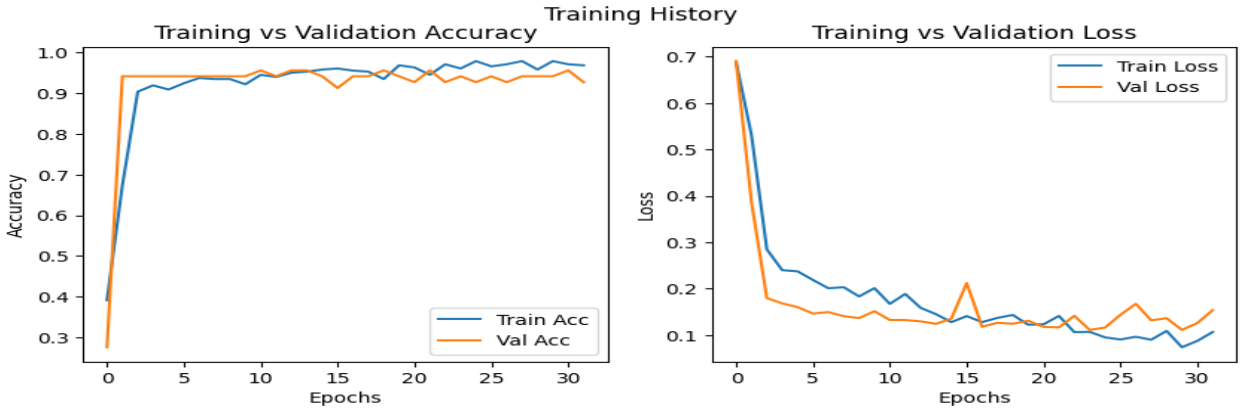

Figure 3.4: Training history

Figures support this quantitative analysis. Figure 4.1 shows the confusion matrix heatmap, highlighting the absence of false positives but also the presence of missed malignancies. Figure 4.2 displays the ROC curve with AUC = 0.9944, confirming strong separability between classes. Figure 4.3 presents the Precision–Recall curve, showing robustness in precision but a decline in recall as thresholds tighten, reinforcing the importance of recalibration if sensitivity is prioritized. Figure 4.4, the training history plot, demonstrates no major divergence between training and validation loss, suggesting limited overfitting, with early stopping preventing excessive variance.

A side-by-side comparison with Al Mamun *et al.* (2025) (Table 4.1) reveals that while their model achieved slightly higher metrics (Accuracy 98.13%, Precision 98.21%, Recall 98.00%, F1-score 97.96%, and AUC-ROC 0.9992), such results may reflect overfitting given the dataset's limited size and imbalance. Crucially, their study did not incorporate explainability tools, weakening clinical trust. By contrast, this study contributes methodological robustness through class weighting, stratified sampling, and the integration of SHAP for interpretability, offering a more balanced and clinically relevant framework.

DOI: https://doi.org/10.57233/ijsgs.v12i2.1084 ISSNp: 2488-9229; ISSNe: 3027-1118

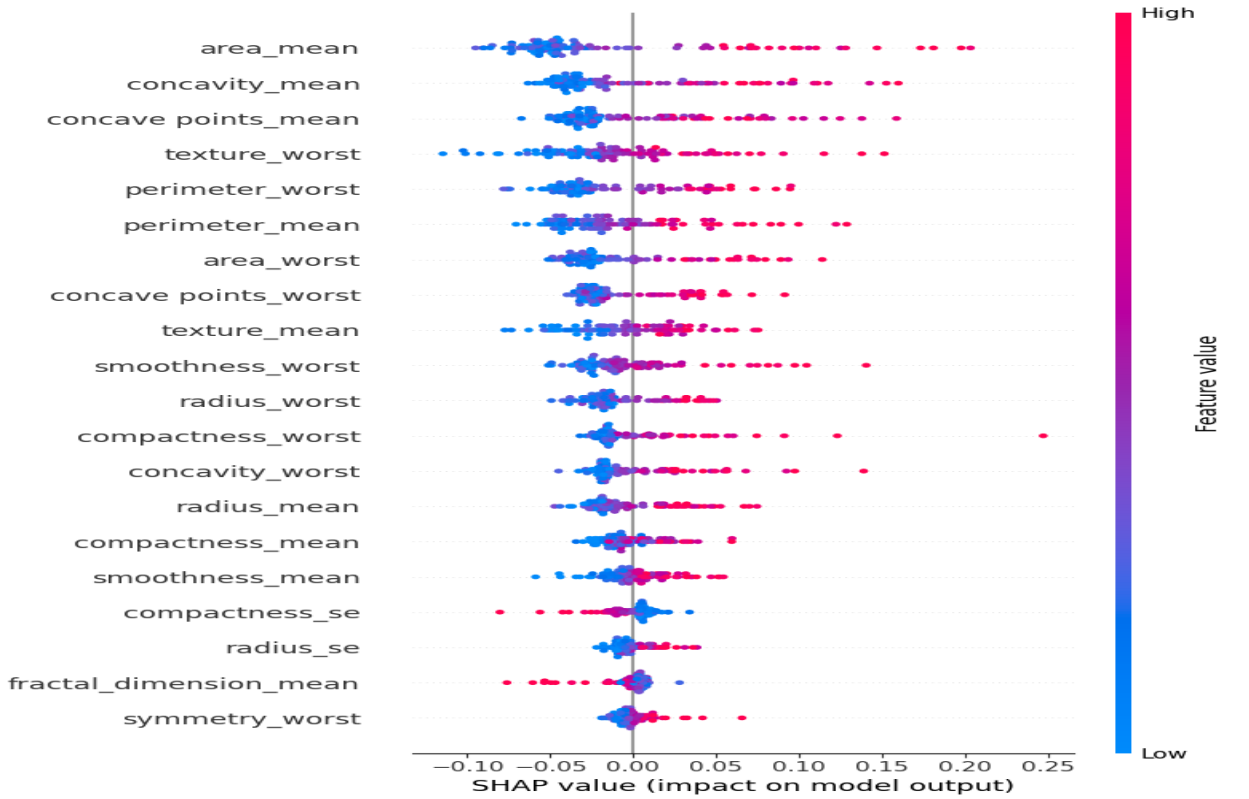

Figure 3.5: SHAP Summary Plot

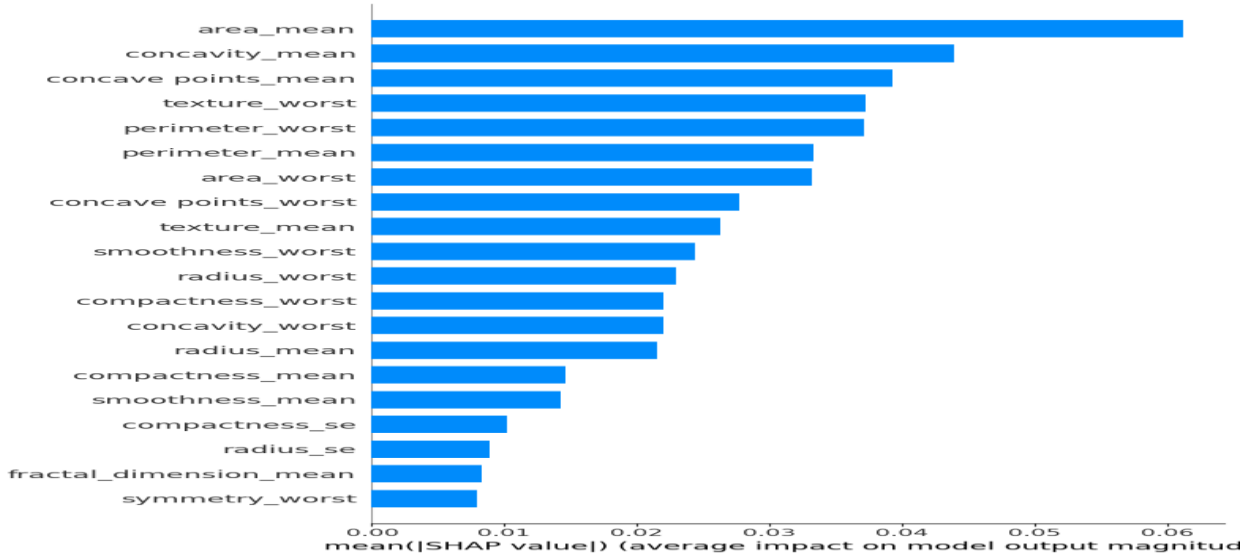

Figure 3.6: SHAP Bar Plot

DOI: https://doi.org/10.57233/ijsgs.v12i2.1084 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12 (2) WEBSITE: https://fugus-ijsgs.com.ng

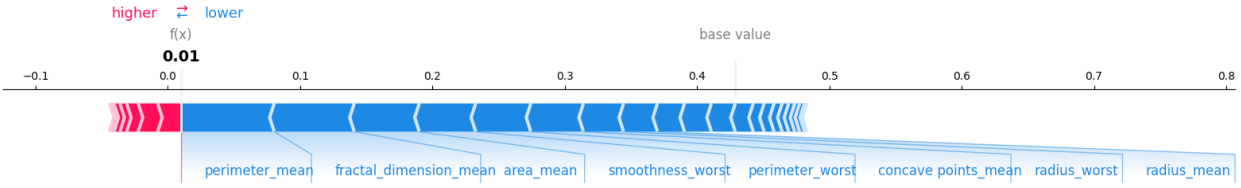

Figure 3.7: SHAP Force Plot (True Positive)

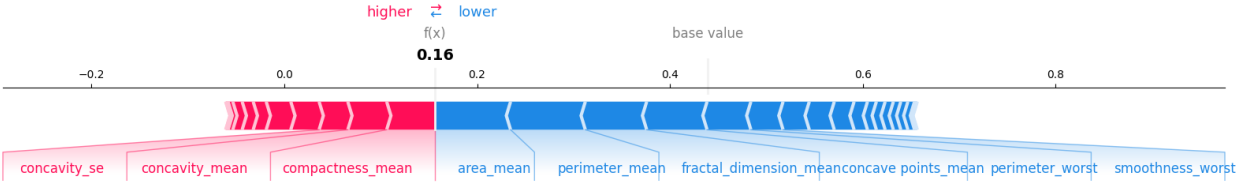

Figure 3.8: SHAP Force Plot (False Negative)

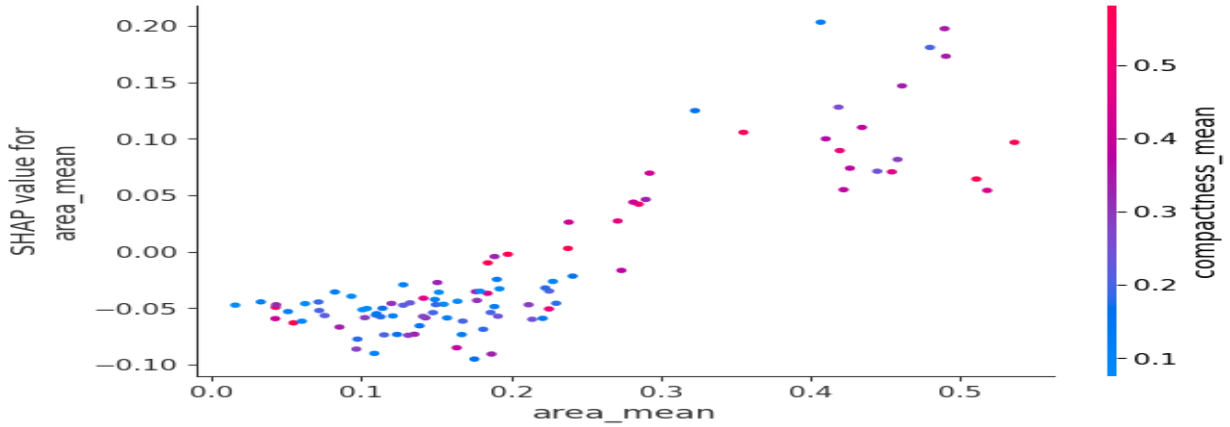

Figure 3.9: SHAP Dependence Plot

Explainability results further illuminate model behavior. Figure 4.5**,** the SHAP summary plot, shows features such as area, concavity and concave points strongly influencing malignancy predictions, consistent with clinical knowledge. Figure 4.6**,** the SHAP bar plot, ranks feature importance, again highlighting area and concavity-related measures as dominant. Local explanations through SHAP force plots **(**Figures 4.7 and 4.8**)** provide clinician-friendly insights, revealing why certain malignant cases were correctly identified or misclassified. Dependence plots (Figure 4.9**)** illustrate nonlinear interactions, such as the combined effect of concave points and perimeter in increasing malignancy probability. Together, these visualizations enhance transparency, enabling both global interpretability and case-specific explanations.

Error analysis underscores the need to reduce false negatives. Possible remedies include recalibrating the decision threshold using Youden's index, optimizing hyperparameters with Bayesian search or Keras-Tuner, comparing class weighting with advanced oversampling techniques (e.g., SMOTE, ADASYN), exploring ensembles with tree-based learners, and applying probability calibration methods such as Platt scaling. These strategies can raise sensitivity while maintaining acceptable precision.

To strengthen statistical reliability, future work should report 95% bootstrap confidence intervals for all metrics and, where possible, include paired significance testing or effect sizes when comparing with prior studies. This

avoids overinterpretation of marginal numerical differences and enhances transparency.

From a practical standpoint, the model demonstrates strong discriminative power and zero false positives, supporting efficiency in clinical workflows by avoiding unnecessary interventions. However, the modest false negative rate indicates that adjunct safeguards such as manual review of borderline cases remain necessary. SHAP explanations further add value by showing clinicians the reasoning behind predictions, a critical feature for real-world trust and adoption.

The principal limitation of the study is reliance on a single, publicly available dataset, restricting claims of generalizability. Test set size also introduces uncertainty in recall estimates, which should be addressed through bootstrapping. Additionally, SHAP explanations are approximate and depend on explainer configuration, making reproducibility and documentation of parameters essential.

In summary, the CNN–GRU model achieved strong discriminative performance, with AUC near 0.995 and perfect precision, but with a modest false negative rate that calls for recalibration and targeted refinements. Compared with previous high-performing but opaque models, this study prioritizes methodological transparency, reproducibility, and interpretability elements essential for clinical readiness. Future work should extend validation to independent hospital datasets, apply systematic calibration, and document SHAP parameters to solidify both trust and external validity.

4.0 CONCLUSION

This study set out to address methodological and interpretability challenges in breast cancer prediction using the Wisconsin Breast Cancer Diagnostic (WBCD) dataset. While existing works, such as Al Mamun *et al.* (2025), report very high performance sometimes exceeding 98% accuracy many overlook three critical issues: the inherent class imbalance in WBCD (63% benign vs. 37% malignant), the risk of overfitting with small datasets lacking external validation, and the absence of model interpretability, which limits clinical trust. To bridge these gaps, a hybrid Convolutional Neural Network–Gated Recurrent Unit (CNN–GRU) architecture was implemented, incorporating class weighting and stratified sampling to address imbalance, and SHAP explainability to enhance transparency. Preprocessing steps included KNN imputation, label encoding, and Min–Max scaling. Evaluation used multiple complementary metrics, accuracy, precision, recall, F1-score, and AUC-ROC to provide a balanced assessment of predictive ability. The model achieved accuracy of 95.61%, precision of 100%, recall of 88.10%, F1-score of 93.67%, and AUC-ROC of 0.9944, confirming strong discriminative power. SHAP analysis revealed clinically relevant predictors (area, concavity and concave points) and provided case-level interpretability, while threshold adjustment demonstrated potential to increase recall with minimal compromise to specificity.

The conclusions of this work are threefold. First, although the CNN–GRU did not exceed the absolute accuracy figures of some prior studies, it delivered a methodologically transparent and clinically interpretable framework that better aligns with healthcare requirements, where explainability is as important as performance. Second, addressing class imbalance proved vital to preserving sensitivity to malignant cases, demonstrating the importance of fairness-oriented preprocessing methods such as class weighting and stratification. Third, SHAP explainability strengthened trustworthiness by clarifying both global and local drivers of predictions, enabling clinicians to understand not only what the model predicts but why. Collectively, these contributions support the case for trustworthy AI in medical diagnostics.

Based on the findings, several recommendations are made. Model enhancement should explore advanced imbalance handling techniques such as SMOTE or hybrid oversampling, and hyperparameter optimization methods like Bayesian search to further improve recall without sacrificing precision. External validation across independent, multi-institutional datasets is critical to ensure generalizability and robustness in real-world populations. For clinical integration, SHAP outputs should be adapted into user-friendly interfaces (e.g., visual dashboards) to support diagnostic decisions in practice. Threshold optimization should be dynamic: screening contexts may prioritize sensitivity to avoid missed cancers, while confirmatory settings may emphasize precision. Finally, the broader scope of application suggests extending the CNN–GRU with SHAP framework to other biomedical problems and multimodal data, combining cytological features with imaging or genomic data for richer diagnostic insights.

Future research directions point toward multimodal deep learning, privacy-preserving strategies like federated learning, and advanced calibration for risk stratification. Longitudinal modeling of disease progression, integration into clinical decision-support systems, and comparative studies of interpretability methods beyond

SHAP remain vital. Most importantly, prospective validation in clinical workflows will be essential to translate these promising results into improved diagnostic accuracy, efficiency, and patient outcomes.

5.0 ACKNOWLEDGEMENT

The authors appreciate the support of Abdu Gusau Polytechnic, Talata Mafara, for providing resources and an enabling environment for this research.

REFERENCES

Abdu-aljabar, R. D., Aljafaar, K. D., Ameen, Z. J. M., & Naman, H. A. (2025). A comparative study of breast cancer detection and recurrence prediction using CatBoost classifier. *Acta Polytechnica*, *65*(2), 136–142. https://doi.org/10.14311/AP.2025.65.0136

Al-Dosari, I. H. (2025). Breast cancer diagnosis improvement based deep learning. International Journal of Engineering and Manufacturing, 15(1), 67–80. https://doi.org/10.5815/ijem.2025.01.04

Al-Husaini, M. A. S., Habaebi, M. H., & Islam, M. R. (2024). Real-time thermography for breast cancer detection with deep learning. Discover Artificial Intelligence, 4, 57. https://doi.org/10.1007/s44163-024-00157-w

Alanazi, S. A., Kamruzzaman, M. M., Islam Sarker, M. N., Alruwaili, M., Alhwaiti, Y., Alshammari, N., & Siddiqi, M. H. (2021). Boosting breast cancer detection using convolutional neural network. Journal of Healthcare Engineering, 2021, 5528622. https://doi.org/10.1155/2021/5528622

Albadr, M. A. A., Ayob, M., Tiun, S., Al-Dhief, F. T., Arram, A., & Khalaf, S. (2023). Breast cancer diagnosis using the fast learning network algorithm. Frontiers in Oncology, 13, 1150840. https://doi.org/10.3389/fonc.2023.1150840

Al Mamun, A., Bhuiyan, T., Hassan, M. M., & Anik, S. I. (2025). Exploring the best machine learning models for breast cancer prediction in Wisconsin. International Journal of Advanced Computer Science and Applications, 16(1), 1362–1367. https://dx.doi.org/10.14569/IJACSA.2025.01601129

Alom, M. R., Farid, F. A., Rahaman, M. A., Rahman, A., Debnath, T., Miah, A. S. M., & Mansor, S. (2025). An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images. Scientific Reports, 15(1), 17531. https://doi.org/10.1038/s41598-025-97718-5

Ansari, Z. A., Tripathi, M. M., & Ahmed, R. (2025). The role of explainable AI in enhancing breast cancer diagnosis using machine learning and deep learning models. Discover Artificial Intelligence, 5, 75. https://doi.org/10.1007/s44163-025-00307-8

Arravalli, T., Chadaga, K., Muralikrishna, H., Sampathila, N., Cenitta, D., Chadaga, R., & Swathi, K. S. (2025). Detection of breast cancer using machine learning and explainable artificial intelligence. Scientific Reports, 15(1), 26931. https://doi.org/10.1038/s41598-025-12644-w

Assiri, A. S., Nazir, S., & Velastin, S. A. (2020). Breast tumor classification using an ensemble machine learning method. Journal of Imaging, 6(6), 39. https://doi.org/10.3390/jimaging6060039

Begum, M. M. M., Gupta, R., Sunny, B., & Lutfor, Z. L. (2024). Advancements in early detection and targeted therapies for breast cancer: A comprehensive analysis. Asia Pacific Journal of Cancer Research, 1(1), 4–13. https://doi.org/10.70818/apjcr.2024.v01i01.04

Boukaache, A., Nasser Edinne, B., & Boudjehem, D. (2024). Breast cancer image classification using convolutional neural networks (CNN) models. International Journal of Informatics and Applied Mathematics, 6(2), 20–34. https://doi.org/10.53508/ijiam.1407152

Dua, D., & Graff, C. (2019). UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml

El-Mawla, N. A., Berbar, M. A., El-Fishawy, N. A., et al. (2024). A novel deep learning approach (Bi-xBcNet-96) considering green AI to discover breast cancer using mammography images. Neural Computing and Applications, 36, 12701–12723. https://doi.org/10.1007/s00521-024-09815-7

Frasca, M., La Torre, D., Pravettoni, G., & Cutica, I. (2024). Explainable and interpretable artificial intelligence in medicine: A systematic bibliometric review. Discover Artificial Intelligence, 4(1), 15. https://doi.org/10.1007/s44163-024-00114-7

Ghasemi, A., Hashtarkhani, S., Schwartz, D. L., & Shaban-Nejad, A. (2024). Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. Cancer Innovation, 3(5), e136. https://doi.org/10.48550/arXiv.2407.12058

Hildt, E. (2025). What is the role of explainability in medical artificial intelligence? A case-based approach. Bioengineering, 12(4), 375. https://doi.org/10.3390/bioengineering12040375

Hossin, M. M., Shamrat, F. J. M., Bhuiyan, M. R., Hira, R. A., Khan, T., & Molla, S. (2023). Breast cancer detection: An effective comparison of different machine learning algorithms on the Wisconsin dataset. Bulletin of Electrical Engineering and Informatics, 12(4), 2446–2456. https://doi.org/10.11591/beei.v12i4.4448

DOI: https://doi.org/10.57233/ijsgs.v12i2.1084 ISSNp: 2488-9229; ISSNe: 3027-1118

Ibrahim, S., Nazir, S., & Velastin, S. A. (2021). Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis. Journal of Imaging, 7(11), 225. https://doi.org/10.3390/jimaging7110225

Khan, K., Awang, S., Talab, M. A., & Kahtan, H. (2025). A comprehensive review of machine learning and deep learning techniques for intraclass variability breast cancer recognition. Franklin Open, 1, 100296. https://doi.org/10.1016/j.fraope.2025.100296

Miao, P., & Zou, Y. (2025). Explainable AI-enabled hybrid deep learning architecture for breast cancer detection. Frontiers in Immunology, 16, 1658741. https://doi.org/10.3389/fimmu.2025.1658741

Mustafa, M. A., Erdem, O. A., & Söğüt, E. (2025). Hybrid optimization and explainable deep learning for breast cancer detection. Applied Sciences, 15(15), 8448. https://doi.org/10.3390/app15158448

Paripooranan, N., Nirasha, W. B., Perera, H. R. P., Vijithananda, S. M., Hewavithana, P. B., Sherminie, L. P. G., & Jayatilake, M. L. (2025). Machine learning-based classification model to differentiate subtypes of invasive breast cancer using MRI. Frontiers in Oncology, 15, 1588787. https://doi.org/10.3389/fonc.2025.1588787

Rahman, M. A., Khan, M. S. H., Watanobe, Y., Prioty, J. T., Annita, T. T., Rahman, S., Hossain, M. S., Aitijjo, S. A., Taskin, R. I., Dhrubo, V., et al. (2025). Advancements in breast cancer detection: A review of global trends, risk factors, imaging modalities, machine learning, and deep learning approaches. BioMedInformatics, 5(3), 46. https://doi.org/10.3390/biomedinformatics5030046

Rasool, A., Bunterngchit, C., Tiejian, L., Islam, M. R., Qu, Q., & Jiang, Q. (2022). Improved machine learning-based predictive models for breast cancer diagnosis. International Journal of Environmental Research and Public Health, 19(6), 3211. https://doi.org/10.3390/ijerph19063211

Sharafaddini, A. M., Esfahani, K. K., & Mansouri, N. (2025). Deep learning approaches to detect breast cancer: A comprehensive review. Multimedia Tools and Applications, 84(21), 24079–24190. https://doi.org/10.1007/s11042-024-20011-6

Srikantamurthy, M. M., Rallabandi, V. S., Dudekula, D. B., Natarajan, S., & Park, J. (2023). Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning. BMC Medical Imaging, 23(1), 19. https://doi.org/10.1186/s12880-023-00964-0

Strelcenia, E., & Prakoonwit, S. (2023). Effective feature engineering and classification of breast cancer diagnosis: A comparative study. BioMedInformatics, 3(3), 616–631. https://doi.org/10.3390/biomedinformatics3030042

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Breast cancer Wisconsin (diagnostic) data set. University of Wisconsin Hospitals, Madison. Retrieved from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)