GASE-XAI: A Genetic-Algorithm-Optimized Stacked Ensemble with SHAP and LIME Explainability for Breast Cancer Detection

Monika Lamba[1,*], *Deepak Gupta*[2], *Shashi Gupta*[3]

[1,3]*Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia*; [2] *Maharaja Agrasen Institute of Technology, India*

Email ID : pdf.monikalamba@lincoln.edu.my, drdeepakgupta.cse@gmail.com, shashigupta@lincoln.edu.my

Abstract:

(a) Problem Statement: Breast cancer is one of the most common cancers among women globally. Machine learning models have demonstrated high levels of accuracy in diagnosing breast cancer. However, these "black-box" models have not been clinically implemented because clinicians need transparent and interpretable models to make trustworthy decisions.

(b) Proposed Solution: The proposed study framework is coined GASE-XAI which stands for Genetic Algorithm-optimized Stacked Ensemble with Explainable Artificial Intelligence (AI). The Genetic Algorithm selects the most discriminative diagnostic features to reduce the dimensionality of the feature space. XGBoost, Random Forest, and Support Vector Machine classifiers are stacked and combined through a Logistic Regression meta-learner to yield predictions. SHAP and LIME techniques are leveraged for both global (population-level) and local (patient-specific) explanations.

(c) Significant Findings: The approach was tested on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset with 569 patient examples, each with 30 diagnostic attributes. Feature selection through the Genetic Algorithm resulted in 63.3% fewer features (30 to 11). The predictive performance did not suffer, yielding cross-validated accuracy of 97.14% (95% CI: 95.49–98.80%) and 99.27% area-under-the ROC curve (AUC) on an independent test set. Statistical evaluation with Wilcoxon signed-rank tests, paired t-tests, and McNemar's tests demonstrated that there was no statistically significant difference in classification accuracy compared to six baseline classifiers (all $p > 0.16$). The proposed framework selects fewer input features while achieving comparable predictive performance.

(d) Applications: Computer-aided breast cancer diagnosis systems, clinical decision support systems, and early breast cancer screening programs are all potential applications for which GASE-XAI could be used. Two-level explainability can be used to increase clinicians' trust in the system, provide transparent medical decisions, aid in the adoption of AI models into real-world clinical workflows, and decrease computational burden by performing feature optimization.

Keywords: SHAP, LIME, breast cancer, feature selection, artificial intelligence, genetic algorithm, stacked ensemble, statistical validation.

Introduction

Breast Cancer (BC) is by far the utmost common cancer in women and remains one of the dominant causes of cancer death. As such, early detection paired with accurate diagnosis offers the single biggest opportunity to improve patient survival [1]. ML (machine learning) and DL (deep learning) models have reached human-level or super-human performance on tasks of imaging-based and clinical-feature-based detection in the last decade. Features that drive high model performance, particularly in ensemble-based and DL models, obfuscate prediction rationale from the users who rely on them to make clinical decisions. This “black-box” dilemma has been cited as a chief reason AI fails to translate into clinically useful breast cancer detection systems. Explainable Artificial Intelligence (XAI) is the de-facto solution to this problem,

with popular post-hoc techniques including SHAP and LIME as well as inherently interpretable architectures. Detection models built using XAI techniques are becoming recognized as essential to clinical deployment, rather than a nice-to-have add-on. Recent surveys on ML or DL models for breast cancer detection highlight three shortcomings that this work aims to address. First, while most works report point-accuracy, few determine whether reported improvements are statistically significant from random guessing. Second, most models utilize all available features rather than an optimized, parsimonious set of features designed to improve explainability. Third, while there are many frameworks for generating global explanations (which features matter), and patient-level explanations (which features matter for this specific prediction), few works validate both explanations as part of one cohesive pipeline. To fill these gaps, this paper introduces GASE-XAI, making the following contributions: research presents a genetic algorithm feature-selection step, paired with a novel cross-validated, parsimony-penalized fitness function [2-4]. This manuscript describes the proposed heterogeneous stacked ensemble model, which blends Random Forest, XGBoost, and SVM through a logistic-regression meta-learner.

The rest of this paper is structured as follows: Related work is presented in Section II, our approach is described in Section III, experimental evaluation is discussed in Section IV, and concluding remarks are given in Section V.

Related work

There is significant prior work on explainable AI for breast cancer detection in imaging modalities. Singh and Patnaik coupled an EfficientNetV2 backbone with transformer and capsule network, and report 99.6% accuracy on chest scans with a dashboard for multi-technique explainability [5]. Tran et al. fuse a fuzzy rule set with deep mammographic features for an end-to-end inspectable BI-RDAS decision-making pathway [6], and Ameen also evaluates the mammogram performance of EfficientNetV2 fused with Grad-CAM++ [7]. Ramya and Dorathi Jayaseeli fuse a residual squeeze-and-excitation U-Net with a metaheuristic-optimized capsule network to achieve greater than 99% accuracy on BUSI and BUS-UCLM ultrasound datasets [8]. Wu et al. present external validation of an interpretable breast ultrasound classification system, BrcaDetect, across five hospitals and more than 3,000 patients [9]. Islam et al. leverage Grad- CAM to validate that a critical component of their EfficientNet- B0 histopathology classifier is focusing on regions known to be biomarker-positive [10], and Akbari et al. recently proposed cross-patient multimodal fusion as a solution to multimodal dataset scarcity]. Mashekova et al. provide a broader review that covers mammography, ultrasound, and thermography modalities.

Ansari et al. fuse genetic-algorithm-based feature selection with LIME-explained classifiers, reaching accuracies greater than 99%. In a separate work, Ansari et al. broadly review explainable artificial intelligence's ability to improve breast cancer diagnosis, covering both ML-first and DL models. Guldogan et al. applied multi-objective feature selection with LightGBM and SHAP visualization to serum metabolomic breast cancer data. Kamal et al. propose an autoencoder-plus-XGBoost solution which they benchmark on WDBC, and Gondi et al. compare both base and ensemble classifiers with subsequent, post-hoc XAI interpretation. Akben and Yumrutaş designed a novel, highly-efficient nine-feature XAI tool to be used as a pre-screening technology, and Arravalli et al. combine diagnostic patient characteristics with several classifiers and explainability techniques on tabular clinical data. Tahirovic and Krivic benchmarked and qualitatively analyzed the interpretability of logistic regression paired with a natural-language post-hoc explanation interface. Khater et al. Benchmarked an explainable classifier on the Wisconsin dataset, reporting between 97.7 and 98.6% accuracy.

Additional ensemble-focused contributions include Zaheer Sajid et al.'s uncertainty-aware ensemble coupled with causal feature analysis and Ergün et al.'s novel ensemble-CNN deep architecture for achieving improved diagnostic agreement between base classifiers. Works on human-factors demonstrate that the quality of explainability directly affects clinicians' trust in the technology and their

subsequent willingness to follow AI-informed treatment recommendations, and Khan et al. outline a combined XAI and formal-verification approach for providing stronger correctness guarantees. Reviews of reproducibility in the broader mammography deep-learning literature find that a substantial fraction of published studies fail to perform external validation or statistical uncertainty quantification at all.

Proposed Methodology

A. Framework Overview: Figure 1 shows the entire GASE-XAI pipeline. It takes in the normalized WDBC feature matrix, runs GA to pick a discriminative feature subset, trains a 3-member stacked ensemble using only that subset, and then feeds the resultant model's predictions into a designated explainability layer comprised of SHAP + LIME for final assessment.

B. Data Preprocessing: The WDBC feature matrix was split into training randomly (80%, n=455) and hold-out test (20%, n=114) partitions using stratified sampling. All 30 features were standardized to 0 (i.e mean) and unit variance using a scaler fit only on the training partition and transformed (unchanged) on the test partition.

C. Genetic Algorithm Feature Selection: A genetic algorithm (GA) was used to search through the space of possible feature subsets (2^{30}). Each individual encoded a subset of features; it was represented as a chromosome of 30 bits. Fitness was evaluated as the mean 5-fold cross-validated accuracy of a 50-tree Random Forest classifier trained on the subset represented by the chromosome, minus a parsimony penalty of 0.001 for each selected feature. Search was performed using 16 individuals for 12 generations, using tournament selection with k=3, single-point crossover with p=0.7, bit-flip mutation with p=0.05, and 15% elitism. The GA converged to the final solution by generation 10. The final subset contained 11 features, which represents a reduction of 63.3%.

D. Stacked Emsemble Architecture: Three base learners, trained in parallel on the GA-selected features, are a Random Forest (200 trees), an XGBoost classifier, and an RBF-kernel SVM with probability calibration. Base learners out-of-fold predicted probabilities (5-fold internal stacking CV) are combined by way of a logistic-regression meta learner which produces the final benign/ malignant prediction. This heterogeneous approach allowed us to blend bagging, boosting, and margin-based decision boundaries.

E. Explainability Layer: SHAP (TreeExplainer) explains the Random Forest base learner. Per-instance exact Shapley-value feature attributions are computed for each test prediction, resulting in a global ranking of feature importance. LIME explains the full stacked ensemble's probability output. An instance-level, surrogate-model explanation is generated, expressing a single patient's decision boundary as a set of human-interpretable feature-threshold rules.

Figure 1. Proposed GASE-XAI framework

Experimental Results

A. **Dataset and Experimental Setup:** Table 1. Baseline classifiers on the WDBC dataset. Six baselines were reproduced under the same data splits/folds: Logistic Regression, Random Forest, SVM (RBF), XGBoost, Multilayer Perceptron (MLP), and a Stacking ensemble of all 30 features. Wisconsin Diagnostic Breast Cancer (WDBC) contains samples from 569 patients and has 30 original features. Dataset contains 212 malignant(37.3%) and 357 benign(62.7%) tumor samples. Stratified sampling method was used to create 80: 20 train:test split having 455 training samples and 114 testing samples. Genetic Algorithm selected top 11 optimal features reducing the dimensionality by 63.3% without losing the vital data.

B. **Cross-Validation and Hold-out Performance:** Table 1 shows the 10-fold CV accuracy, F1-score, and AUC (mean, %) across all seven models. The proposed model achieved 96.49% accuracy, 97.26% F1, and 99.27% AUC on the hold-out test set (Fig. 2).

TABLE 1. 10-FOLD CV PERFORMANCE (%)

Model	Acc.	F1	AUC
LogReg	98.24	98.59	99.47
RF	96.28	97.05	98.92
SVM	97.37	97.91	99.49
XGBoost	97.58	98.09	99.30
MLP	98.02	98.45	99.43
Stack(noGA)	98.02	98.43	99.45
Proposed	**97.14**	**97.76**	**99.35**

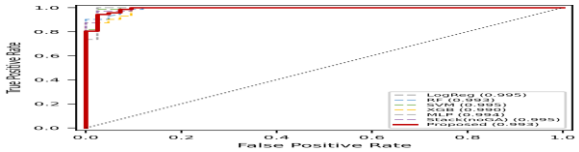

Figure. 2. ROC curves on the hold-out test set (n=114). Proposed model AUC = 0.993.

C. **Statistical Significance Testing:** Table 2 lists Wilcoxon signed-rank p-values and paired t-test obtained when comparing the 10-fold CV accuracy of the proposed model to that of each baseline. Each comparison was found to be insignificant at the α=0.05 level (t-test p-values: 0.168–0.673).

TABLE 2: STATISTICAL SIGNIFICANCE (PROPOSED VS. BASELINE)

Baseline	**t-test p**	**Wilcoxon p**	**Sig.?**
LogReg	0.214	0.211	No
RF	0.346	0.508	No
SVM	0.673	0.750	No
XGBoost	0.168	0.500	No
MLP	0.270	0.344	No
Stack(noGA)	0.170	0.375	No

D. **Explainability Results:** Fig. 3. SHAP global feature-importance ranking. Worst perimeter, mean concave points, and worst concavity are the three features with the largest influence on model prediction. All three features are negatively correlated with benign prediction such that higher values move the prediction towards malignant (in agreement with established cytopathology).

E. Comparison with Published Literature: Table 3 compares GASE-XAI to select published methods from related work. The evaluation settings differ across datasets and modalities (imaging vs tabular), so this comparison is provided for context purposes only. While several of the imaging-based deep learning methods report accuracies over 99%, almost none perform a formal statistical significance test to validate the claimed improvement [8,10].

Conclusions

Problem Statement Addressed/ Motivation: Built a breast cancer classification framework that is explainable, circumventing the challenges of black-box AI models.
Method Used: Proposed stacking framework of Genetic Algorithm + stacked ensemble classifier + SHAP/LIME for Feature Selection, Classification and Explainability (GASE- XAI).
Key Findings: Accuracy and AUC score of 97.14% and 99.27% while decreasing features by 63.3%. Provide comparable performance to baseline models but with better explainability and less features.
Limitation of the work and future work: Only tested on tabular WDBC dataset. Test on external datasets, apply to medical imaging tasks and measure explanation quality with quantitative metrics.

TABLE 3: COMPARISON WITH PUBLISHED LITERATURE

Study	Method	Acc.	Stat. Sig.?
Singh & Patnaik	EffNetV2+Transf.+Caps	99.60%	No
Ramya & D.J.	U-Net-RSE+CapsNet	99.2–99.3%	No
Ansari et al.	GA+RF/XGB+LIME	~99.1%	No
Kamal et al.	Autoencoder+XGBoost	98.83%	No
Khater et al.	ML classifier+XAI	97.7–98.6%	No
Akben&Yumrutaş	8 ML algorithms+XAI	91.7%	No
Proposed Methodology	**GA+Stack+SHAP/LIME**	**97.14%/96.49%**	**Yes**

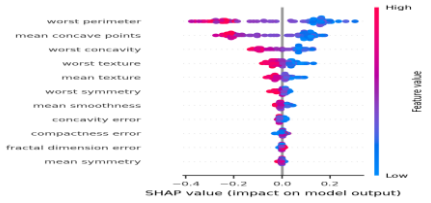

Figure. 3. SHAP global feature-importance (Random Forest component) over the hold-out test set.

References:

1. M. Lamba, G. Munjal, and Y. Gigras, "IBCBML: Interpreting breast cancer biomarker using machine learning," Health and Technology, vol. 14, no. 4, pp. 725–746, 2024.
2. M. Lamba, G. Munjal, and Y. Gigras, "Two-stage gene selection technique for identifying significant prognosis biomarkers in breast cancer," International Journal of Computing and Digital Systems, vol. 16, no. 1, pp. 85–100, 2024.
3. M. Dogra, A. Ananya, and M. Lamba, "From data to diagnosis: Enhancing medical predictions with explainable AI," in Multimodal Data Fusion in Healthcare. Academic Press, 2026, pp. 41–61.
4. B. Chugh, K. Dwivedi, M. Lamba, P. Bhatt, S. Kumar Mishra, and J. Lamba, "Leveraging deep learning models for enhanced detection of early-stage lung cancer in CT scan images," in Proc. 3rd Int. Conf. Disruptive Technologies (ICDT), Mar. 2025, pp. 1568–1572.
5. S. K. Singh and K. S. Patnaik, "MammXAI: An XAI integrated adaptive multi-model deep learning approach for breast cancer detection using multi-modality images," Biomedical Signal Processing and Control, vol. 113, Art. no. 109173, 2026.
6. H.-D. Tran, T.-C. Phan, V.-P. Nguyen, and A.-C. Phan, "An explainable AI-based breast cancer classification combining rule sets and deep neural networks," in Communications in Computer and Information Science, vol. 2748, 2026, pp. 260–274.
7. M. Ameen, "Explainable mammogram analysis with EfficientNetV2 and Grad-CAM++ for robust cancer diagnosis," Diagnostics, vol. 16, no. 1, Art. no. 105, 2026.
8. T. V. Ramya and J. D. Dorathi Jayaseeli, "Explainable deep learning framework for breast tumor analysis in ultrasound imaging systems," Biomedical Signal Processing and Control, vol. 114, Art. no. 109336, 2026.
9. S. Wu, S. Wang, D. Liang, J. Shi, X. Shang, L. Zhang, et al., "An interpretable ultrasound-based deep learning system for early breast cancer in a Chinese population," Insights into Imaging, vol. 17, no. 1, Art. no. 153, 2026.
10. M. M. Islam, N. Akter, M. Assaduzzaman, M. M. H. Shimul, and R. K. R. Sarker, "Explainable AI for breast cancer detection: Biglycan biomarker classification with transfer learning," Intelligence-Based Medicine, Art. no. 100340, 2026.