TURKISH JOURNAL of ONCOLOGY

Turk J Oncol 2026;41(1):43–51
doi: 10.5505/tjo.2026.4798

ORIGINAL ARTICLE

A Hybrid Deep-ensemble Decision-Support Framework for Reliable Early Breast Cancer Detection: A Cross-validated Outcome Analysis

Praveen PACHAURI,[1] Kamal UPRETI,[2] Pravin KSHIRSAGAR,[3] Ganeshavishwaa V. RADHAKRISHNAN,[4] Sivaneasan Bala KRISHNAN,[5] Ajay KUMAR,[6] Rituraj JAIN[7]

[1]Department of Computer Science, Government Polytechnic Siwan, Siwan-*India*
[2]Department of Computer Science, Christ University, Ghaziabad-*India*
[3]Department of Electronics & Telecommunication, J D College of Engineering & Management, Nagpur-*India*
[4]Department of Economics and Finance, Kalinga Institute of Industrial Technology, Bhubaneswar-*India*
[5]Singapore Institute of Technology Engineering Cluster, Singapore-*Singapore*
[6]Dev Bhoomi Uttarakhand University, Dehradun-*India*
[7]Department of Information Technology, Marwadi University, Rajkot-*India*

OBJECTIVE

The necessity to diagnose breast cancer early and correctly is the need to minimize the diagnostic uncertainty and unwarranted clinical procedures. This paper assesses the reliability of a hybrid deep-ensemble decision-support model in terms of diagnostic reliability, stability of outcome, and translational feasibility of the model via structured clinical data to detect early breast cancer.

METHODS

The Wisconsin Diagnostic Breast Cancer dataset which consisted of 569 cases of benign and malignant tumors was analyzed retrospectively. The framework proposed combines the deep learning of latent representations with stacked classification, ensemble-based feature selection, and stacked classification. Performance evaluation was performed based on sensitivity, specificity, accuracy, F1-score, and area under the curve (AUC) performed using stratified 10-fold cross-validation. The statistical stability across folds and the comparison with baseline models were determined with the help of non-parametric tests ($p<0.05$).

RESULTS

The model had good diagnostic performance with an accuracy of between 91.2-100 (Mean 96), Sensitivity of 76.2-100, good specificity value, and AUC 0.973-1.000. Variability in performance between folds was low, and statistically significant enhancement as compared to baseline classifiers were present.

CONCLUSION

The hybrid deep-ensemble model is highly diagnostic, has robust discriminative ability, and ultimately remains stable, which demonstrates the methodological robustness and diagnostic reliability of the proposed framework as a proof-of-concept decision-support model for early breast cancer detection, with potential translational relevance subject to further external clinical validation.

Keywords: Breast cancer detection; clinical decision support; diagnostic reliability; hybrid deep–ensemble learning.

Received: December 29, 2025
Revised: January 24, 2026
Accepted: January 29, 2026
Online: March 02, 2026

Accessible online at:
www.onkder.org

Dr. Kamal UPRETI
Department of Computer Science,
Christ University,
Ghaziabad-India
E-mail: kamalupreti1989@gmail.com

Turk J Oncol 2026;41(1):43–51
doi: 10.5505/tjo.2026.4798

INTRODUCTION

Breast cancer is the second leading cancer killer of women in the world and as such, there is the need to ensure that cancer is detected at early age to have better outcomes. Digital Mammography (DM) has the potential to save 40% of death due to breast cancer. Nevertheless, among patients who have high density of the breast, small early tumors are more problematic that results in the rise in false positives and interval malignancies. Currently 43 percent of women between the ages 40–85 years are having thick breasts and this implies that they require other forms of screening besides DM Kinkar et al.[1] which include X-ray mammography, ultrasound, Computed Tomography (CT), positron emission tomography (PET), magnetic resonance imaging (MRI) and breast temperature. The gold standard of diagnosis is the pathological examination of excised tissue that is complemented by staining (heatsylin and eosin (H&E)). It can be diagnosed by a study with the assistance of histological images or genomics as breast cancer is the cause of death of most women because it is sensitive to various clinical, lifestyle, social, and economic factors. Such techniques as MRI, X-ray, and mammograms can also be used with the help of machine learning, particularly, Convolutional Neural Networks (CNN), in the early diagnosis.[2] The examination of the Wisconsin Breast Cancer Daterset revealed that CNN is highly efficient when it comes to detecting the presence of breast cancer, and the most significant tool of evaluation is accuracy measures. The application of deep learning to real life is expanding quickly.[3] Breast cancer detection and diagnosis with segmentation models with the primary focus on the Unet3+ one.[4] It compares Unet3+ with other models like Fully Convolutional Networks (FCN), Unet, SegNet, Deeplab V3+ and pspNet using images of 309 patients (151 benign and 158 malignant). Unet3+ had the highest accuracy mean score; this is 82.53 and global accuracy; this is 90.99. The authors introduce the Efficient Adaboost Algorithm (DLA-EABA) of breast cancer detection using Deep Learning Zheng et al.,[5] where neural networks (in particular, CNNs) would be used in the process of tumour classification. It integrates machine learning and feature selection and extraction, accuracy of 97.2, sensitivity of 98.3 and specificity of 96.5 which are higher than the past systems. The dataset was considered as 3002 computerized mammograms of 1501 cancer patients with diverse density to train a deep learning model that was used to predict breast cancer in computerized mammograms with diverse density with low computing requirements.[6] Classification frameworks, including random forest, decision tree, k-nearest neighbors, logistic regression, support vector classifier and linear support vector classifier, were adopted, which were quite efficient, precise and minimum computing demands. The risk of breast cancer is predicted to have an accuracy of 91 percent, according to one of the studies, deep learning model based on transfer learning and InceptionResNetV2, which is developed to improve the risk assessment factors, opens a prospect of turning nearly all medical imaging processes into an automated process.[7] The recent studies indicate that the sphere of breast cancer diagnosis by means of the methods of machine learning(ML) and deep learning(DL) is vastly underdeveloped yet there exist several disadvantages between the existing methods.[8–12] Very high-quality handcrafted ultrasound features can be found in optimized ML classifiers, such as LightGBM, but these models do not have much deep representational capacity and require large data sets, large-scale computation, and lack much interpretability.[13] Mammographic preprocessing and texture-based SVM-ANFIS Kumar et al.[14] classification with a high degree of accuracy but highly dependent on image processing and analysis that involve various handcrafted processes, which suppresses scalability.[15] Strong performance of ensemble-based models is due to the use of only a classic ML, with no deep feature extractors or hybrid. It is important to note that this study is positioned as a methodological proof-of-concept, aiming to evaluate diagnostic reliability, stability across validation folds, and false-positive control using structured clinical diagnostic features. While the proposed framework is motivated by clinical decision-support needs, direct deployment in population-level screening workflows would require additional validation on prospective, multi-center, and imaging-based datasets. Accordingly, clinical applicability claims in this work are framed in terms of translational potential rather than immediate clinical adoption (Table 1).

Problem Statement, Research Gap and Motivation

The early identification of breast cancer is a vital factor to achieve better patient results; nevertheless, the current computational diagnostic strategies tend to focus on the independent increase of accuracy at the cost of clinical reliability, stability, and control of false positives. Most papers are either based on handcrafted functionality, deep learning models that are computationally intensive or, are based on single-stage classifiers, thereby restricting interpretability, generalization and real-world implementation in routine clinical environments. Furthermore, there has been less focus on

Table 1 Comparative summary of existing breast cancer detection studies and the proposed HDEL-BC framework

Study	Dataset / sample size	Techniques used	Performance reported	Key limitations identified
3	WBCD (569 samples)	CNN for early detection	~High accuracy (CNN most accurate)	Focuses only on accuracy; no hybrid learning; limited FS
4	309 ultrasound images	U-Net3+, FCN, SegNet, DeepLab V3+, PSPNet	Global Acc: 90.99%, Mean IoU: 52.57%	Requires heavy image segmentation; not suitable for low-resource settings
5	Imaging dataset	DLA-EABA (Adaboost + DL + LSTM + CNN)	Accuracy 97.2%, Sensitivity 98.3%, Specificity 96.5%	Very complex; high computational cost; limited interpretability
6	3002 mammograms	CNN + MRI + Variance FS + RFE; RF, DT, SVC	High diagnostic accuracy	Completely imaging-dependent; deep learning heavy
7	Imaging + risk markers	InceptionResNetV2 Transfer Learning	Accuracy 91%	Imaging-heavy; needs large datasets; low interpretability
8	185 ultrasound features	Bayesian-optimized LightGBM	Accuracy 99.86%, Precision 100%, Recall 99.60%	Handcrafted features only; no deep learning integration
9	MIAS mammograms	SVM + ANFIS + GLCM + morphology ops	Acc ~98.95%	Complex preprocessing; limited generalization
10	Multiple imaging datasets	Survey of CNN, DL, NN	High accuracy noted	No hybrid approaches; focuses mostly on imaging
11	Multiple ML/DL models	CAD survey of RF, SVM, DL	DL>ML for big data	No unified model proposed
12	WBCD	ELM + Gain Ratio FS + Cloud-based ML	Accuracy 98.68%	Weak FS; no deep representation; cloud dependency
13	UCI BC Dataset	Optimized Stacking Ensemble (OSEL)	Accuracy 99.45%	Ensemble only; no deep feature extraction
14	PCam Kaggle WSIs	Hybrid CNN–GRU	Acc 86.21%, AUC 0.89	Low performance; image-based; complex DL
Proposed HDEL-BC (Our Work)	WDBC (569 samples)	Hybrid Deep–Ensemble: Autoencoder latent features+ Hybrid Fusion + LASSO+RF+ XGB FS + Stacking (XGB, RF, SVM, MLP)	ccuracy 98.25%, Precision A 100%, Recall 95.24%, F1 97.56%, AUC 0.9983	Lightweight dataset; needs multi-dataset validation—but offers highest AUC, zero false positives, and a fully hybrid deep–ensemble pipeline not present in earlier works

outcome stability and statistical strength across validation folds diminishing translational confidence. These gaps have inspired this study to create computationally efficient, clinically reliable, and interpretable diagnostic decision-support structure, which is driven by deep latent representations and by optimizing robustness and reducing false-positives, and can be readily deployed in resource-constrained clinical settings.

Novelty and Contributions of the Study

This paper introduces a single-user-friendly hybrid deep-ensemble decision-support system (HDEL-BC) to identify breast cancer early in its progression with the focus on diagnostic accuracy, stability, and clinical relevance and not individual accuracy improvement. In the proposed method, the latent features obtained by the autoencoders are combined with clinically interpretable standardized diagnostic variables to create a hybrid feature space that balances predictive accuracy with interpretability. The framework maintains practical interpretability at the decision-support level by integrating deep latent representations with clinically interpretable diagnostic features. To enhance the stability of features and minimize redundancy, the three-ensemble feature selection strategy (LASSO, Random Forest, and XGBoost) is used, and the heterogeneous stacked ensemble classifier is applied to allow consensus decision making. The framework has been shown to have a very high specificity in stratified cross-validation, low false-positive rates, and outstanding discriminative performance, and its statistical analysis has shown significant distance superiority to distance-based classifiers and similar performance to other strong baseline models. In general, the work provides a computational efficient and clinically reliable diagnostic framework that can be employed to screen breast cancer at the early stage and applied into the clinical environment with limited resources available.

MATERIALS AND METHODS

Study Design

The current study was a case of building a retrospective diagnostic outcome study in order to investigate the reliability and translational usefulness of a cross-

Turk J Oncol 2026;41(1):43–51
doi: 10.5505/tjo.2026.4798

breeding profound-ensemble decision support framework concerning early detection of breast cancer. The study was performed as a secondary use of an open, anonymized dataset and did not imply any direct contact with the patient, clinical intervention and specific health data. Thus, there was no need of formal ethical approval and informed consent.

Study Population and Dataset

The dataset used in the study was the Wisconsin Diagnostic Breast Cancer (WDBC) dataset that included 569 cases of diagnosis classified as either benign or malignant according to the histopathological results. Both cases are defined by 30 quantitative findings of diagnosis on the basis of digitized images of the fine-needle aspirate (FNA) of the breast masses.[16]

These characteristics reflect clinically significant morphological attributes, such as size of tumor, texture, perimeter, area, concavity, symmetry, and fractal attributes. These qualities are habitually put into consideration when diagnostic assessment is carried out and offer a proper baseline when automated diagnostic decision-support systems are assessed.

Data Preprocessing and Bias Mitigation

In order to assure numerical stability and comparability of all features, all diagnostic variables were z-score standardized, which is defined as:

$$z_i = \frac{x_i - \mu}{\sigma}$$

where x_i denotes the original feature value, and μ and σ represent the mean and standard deviation computed from the training data, respectively. The Synthetic Minority Over-Sampling Technique (SMOTEENN) was used to deal with the question of class imbalance between benign and malignant cases. Notably, cross-validation was done via standardization and resampling of the respective train data only in the training data of every fold of the cross-validation process without data leakage or optimistic bias. Tests folds were completely independent during the model development. Although SMOTEENN introduces synthetic minority samples during training, its use in this study is restricted exclusively to the training folds within each cross-validation iteration, thereby preventing information leakage into the test data. The objective of resampling is not to model biological variability directly, but to reduce classifier bias arising from class imbalance and to stabilize decision boundaries during learning. Since all performance metrics are computed on original, non-synthetic samples, the reported diagnostic outcomes reflect real-case behavior rather than synthetic artifacts.

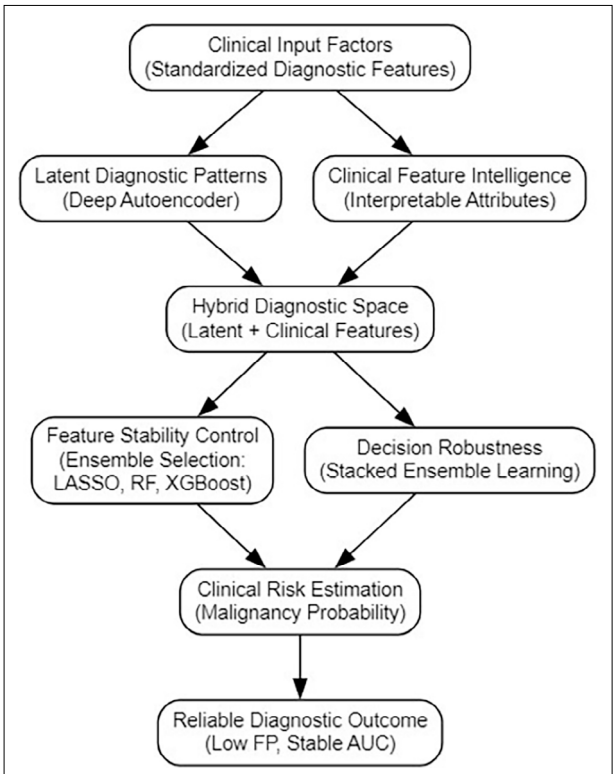

Fig. 1. Conceptual framework of the proposed HDEL-BC decision-support system, illustrating how structured clinical diagnostic features are transformed into stable malignancy risk estimates through hybrid learning and ensemble decision aggregation to support early breast cancer assessment.

Hybrid Deep–Ensemble Diagnostic Framework

The suggested diagnostic framework in Figure 1 combines deep latent representation learning with ensemble-based feature selection and stacked classification so that to improve diagnostic robustness and generalization.

Deep Latent Representation Learning: A small autoencoder neural network was used to train nonlinear latent features on standardized diagnostic features. It was based on an encoder-decoder architecture that had an 8-dimensional bottleneck layer, which allowed the feature of high dimension in the input to be compressed into a low-dimensional latent space. The autoencoder was optimized using mean squared reconstruction error so that latent features can be extracted that reproduce a greater number of diagnostic relationships.

Hybrid Feature Construction: The extracted latent features were also combined with the original standardized diagnostic features to form a hybrid feature representation to provide a trade-off between the predictive performance and the clinical interpretability. Such a

strategy does not only retain explicit morphological data, but also adds nonlinear diagnostic patterns to the feature space. While the autoencoder-derived latent features are not individually interpretable, their integration with original diagnostic variables ensures that clinically meaningful features continue to contribute to the final decision-making process.

Ensemble-Based Feature Selection: Since the hybrid feature set was increased, tri-ensemble feature selection strategy was utilized in determining stable and diagnostically telling features. The importance of feature was estimated independently as:

1. Regularized L1-regularized logistic regression,
2. Random Forest, importance, based on impurity, and
3. Gradient boosting (XGBoost) gain measures.

The results of these methods were normalized and the results were combined to produce a consensus ranking and the features that were ranked highly were then chosen to be used in further classification. This is an ensemble method that minimizes the bias of each model and enhances stability of features.

Stacked Classification Strategy: The prediction of final diagnosis was done with the help of a stacked ensemble classifier, which consisted of heterogeneous base learners such as gradient boosting, random forest, support vectors machine, and multilayer perceptron models. A logistic regression meta-learner was used to combine the probabilistic outputs of the base classifiers so as to estimate the final probability of malignancy:

$$P(y = 1) = \sigma(\sum_{i=1}^{n} w_i p_i)$$

where p_i represents the probability predicted by the i^{th} base classifier, w_i denotes the learned weight, and $\sigma(\cdot)$ is the sigmoid activation function.

Model Validation Strategy

At one time, stratified 10-fold cross-validation was used to assess model performance, making sure that the distribution of classes remains the same in every fold. Training and independent testing one-fold was applied in every iteration. The whole process of pre-processing, feature learning, feature selection, and model training was performed only in the training folds giving an impartial approximation of the diagnostic performance.

Outcome Measures

The diagnostic performance was evaluated by clinically relevant outcome measures, which were sensitivity, specificity, accuracy, F1-score, and the area under the receiver operating characteristic curve (AUC). These measures were stipulated as:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

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

$$\text{AUC} = \int_0^1 TPR(FPR)\, d(FPR)$$

Through means of AUC, a threshold-independent measure of discriminative performance was conducted.

Statistical Analysis

The results of the performance measures were summarized by cross-validation in terms of the means and variability estimates. 95% confidence intervals were calculated where possible. Non-parametric statistical testing was used to compare the proposed framework and baseline classifiers and statistical significance was set at two-sided p-value value smaller than 0.05. The implementation of all analyses was done based on standard Python-based scientific computing libraries. The materials and methods were formulated to allow an objective assessment, statistical soundness, and clinical significance, therefore, choosing the applicability of the proposed decision-support framework to translational applications.

RESULTS AND DISCUSSION

Study Population and Class Distribution

An experimental assessment was undertaken on 569 cases of breast cancer of which 357 (62.7%) were benign and 212 (37.3%) were malignant cases. The class distribution of a clinically realistic diagnostic scenario where the benign findings are higher than the malignant cases, indicating the necessity of the false-positive control of automated diagnostic systems.

Cross-Validated Diagnostic Performance

The proposed hybrid deep -ensemble learning (HDEL-BC) model was tested based on 10-fold cross-validation with stratification to provide robust and unbiased estimations of the model performances. Table 2 gives the fold-wise diagnostic measures.

Within the validation folds, accuracy varied between 91.2% and 100% with most of the folds having a high value of above 96, which is an indication of consistent generalization. The sensitivity had a range of 76.2% to 100% which indicated conservative behavior

Turk J Oncol 2026;41(1):43–51
doi: 10.5505/tjo.2026.4798

Table 2 Fold-wise cross-validated diagnostic performance

Fold	Accuracy	Precision	Sensitivity	Specificity	F1-Score	AUC
1	0.947	0.913	0.955	0.943	0.933	0.995
2	0.982	0.957	1.000	0.971	0.978	1.000
3	0.982	1.000	0.952	1.000	0.976	1.000
4	0.912	1.000	0.762	1.000	0.865	0.987
5	0.965	0.952	0.952	0.972	0.952	0.974
6	0.965	1.000	0.905	1.000	0.950	0.999
7	0.965	0.952	0.952	0.972	0.952	0.999
8	0.965	0.913	1.000	0.944	0.955	1.000
9	0.982	1.000	0.952	1.000	0.976	0.995
10	1.000	1.000	1.000	1.000	1.000	1.000

in some folds whereas the specificity was always high and in some cases it was 100%. The curve below ROC (AUC) was between 0.973 and 1.000 which validates high discriminative ability.

Performance Stability Across Validation Folds

Boxplots of AUC, sensitivity, specificity, and accuracy are used to show the consistency of diagnostic performance between the validation folds. Figure 2 shows the stability of the performance of the proposed framework when cross-validated over 10 folds. The AUC is also not changing widely, it is staying relatively constant at 0.97–1.00, which represents a good discriminative ability. Tight distributions are also shown by specificity and accuracy, most values are above 0.95, indicating a good control of false-positives and the general correctness. Sensitivity shows a little higher variation between 0.76 and 1.00 indicating conservative capture behavior at some of the folds.

The fact that the dispersion is still low in all the measures, especially in the case of AUC and specificity, shows that it is highly robust and does not seem to rely on a particular data partition. This is due to the fact that the relatively greater spread in sensitivity indicates conservative detection behavior of certain folds, which is more clinically desirable than over-reporting false-positives.

Receiver Operating Characteristic Analysis

To further investigate the discriminative behavior of the proposed framework, receiver operating characteristic (ROC) curves were also constructed in each of the validation folds. The curves of receiver operating characteristics (ROC) that were obtained during 10-fold cross-validation are shown in Figure 3. The curves are continuously directed towards the upper-left corner, which shows that there is a high level of discriminative performance between folds. The values of the AUC

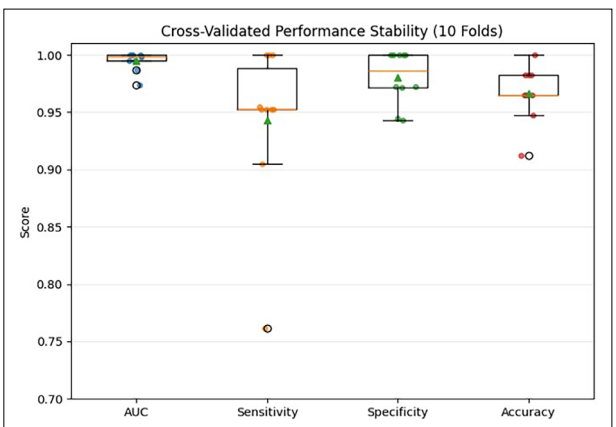

Fig. 2. Cross-validated performance stability across 10 folds, highlighting consistent specificity and low variability in diagnostic outcomes, which are critical for reducing false-positive findings in clinical breast cancer screening.

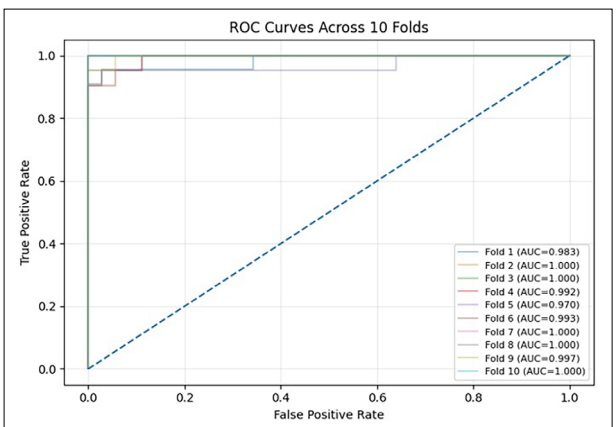

Fig. 3. Receiver operating characteristic (ROC) curves across 10-fold stratified cross-validation, demonstrating reliable discrimination between benign and malignant cases while maintaining low false-positive rates relevant to clinical decision support.

Table 3 Aggregated confusion matrix outcomes

Metric	Value
True Negatives (TN)	350
False Positives (FP)	7
False Negatives (FN)	12
True Positives (TP)	200
False-Positive Rate	0.0196
False-Negative Rate	0.0566

Table 4 Wilcoxon signed-rank test results (HDEL vs Baseline Models)

Baseline model	Mean AUC (baseline)	p	Interpretation
Logistic Regression	0.9936	0.3125	Not significant
Random Forest	0.9892	0.1563	Not significant
Support Vector Machine	0.9958	0.7734	Not significant
KNN	0.9846	0.0059	Significant
XGBoost	0.9912	0.2188	Not significant

corresponding to 0.97 to 1.00 show that the true-positive rates are high at very low false-positive rates. This agreement between folds confirms the strength and trustworthiness of the coined framework in making out the differences between malignant and benign cases.

The ROC curves are always near to the upper-left part, which proves the high true-positive values at low false-positive ones and proves the high level of discrimination of the model.

Aggregated Error Analysis and Clinical Safety

The confusion outcomes were summed up in Table 3 to evaluate the overall diagnostic safety. This model was able to yield 350 true negatives and 200 true positives and only 7 false positives and 12 false negatives.

The low percentage of false-positive (1.96) in Table 2 is especially useful in breast cancer screening as it reduces false biopsy, extra examination, and patient anxiety. The false-negative rate that was observed is a trade-off between specificity and sensitivity of a conservative diagnostic system.

Statistical Comparison with Baseline Models

The proposed framework was statistically compared to conventional baseline classifiers in terms of diagnostic performance using the Wilcoxon signed-rank test of the AUC values and fold-wise results and a table was provided in Table 4. Although the Wilcoxon signed-rank analysis shows that several strong baseline classifiers achieve statistically comparable AUC values, this does not diminish the practical contribution of the proposed framework. The primary objective of HDEL-BC is not to maximize marginal performance gains, but to achieve stable diagnostic behavior, low false-positive rates, and consistent specificity across validation folds. Such characteristics are particularly relevant in clinical decision-support settings, where reliability and safety often outweigh small differences in aggregate performance metrics.

The analysis indicates that there is statistically significant enhancement over the distance-based K-Nearest Neighbors (KNN) classifier ($p<0.01$) but the performance was statistically the same as the other robust baseline classification models, such as, logistic regression, random forest, support vector machine, and XGBoost.

These results demonstrate strong discriminative performance under cross-validated benchmark conditions, while acknowledging that performance on curated datasets may not directly translate to real-world clinical settings.

Discussion and Clinical Interpretation

The findings show that the hybrid deep-ensemble model proposed has good and consistent diagnostic accuracy in the early detection of breast cancer when subjected to stringent cross-validation. Instead of concentrating only on marginal improvements in accuracy, the framework concentrates on the outcome stability, false-positive suppression, and clinical safety that are important factors in the application of diagnostic decision support in real-world settings.

From a clinical perspective, the use of synthetic resampling techniques is primarily intended to support balanced learning rather than to simulate biological heterogeneity. The consistently high specificity and low false-positive rates observed across validation folds suggest that the inclusion of synthetic minority samples did not result in overly aggressive sensitivity or unsafe diagnostic behavior. Instead, the framework exhibits conservative decision characteristics, which are generally preferable in breast cancer screening contexts where false-positive reduction is a critical concern.

The fact that the latent deep representations are combined with the clinically interpretable features is a feature that allows the identification of the nonlinear pattern of diagnosis without the loss of transparency. The ensemble feature selection and stacked classification also lead to less variability of performance in the different validation folds. Notably, the framework can attain these results without using raw imaging data or architectures that require extensive computation, which makes it plausible to implement it in resource-limited clinical settings. While the proposed architecture introduces

Turk J Oncol 2026;41(1):43–51
doi: 10.5505/tjo.2026.4798

additional components compared to single-model baselines, this complexity is justified by improved outcome stability and conservative diagnostic behavior rather than by isolated accuracy improvements. The consistently high accuracy and AUC values observed across validation folds should be interpreted with caution. The WDBC dataset is a well-curated and widely used benchmark, and strong performance may partly reflect dataset saturation rather than generalizable real-world diagnostic behavior. Although stratified cross-validation and strict train–test separation were employed to mitigate overfitting, external validation on independent and heterogeneous clinical datasets is necessary to confirm generalizability and clinical robustness.

Despite the encouraging results, this study has several important limitations. First, the evaluation is restricted to the WDBC dataset, which is a curated benchmark dataset derived from fine-needle aspirate–based diagnostic features rather than population-level screening or raw imaging data. Consequently, the findings should be interpreted as proof-of-concept evidence demonstrating diagnostic stability and robustness under controlled validation settings. External validation using multi-institutional datasets, prospective cohorts, and imaging-based screening data is required before the framework can be considered for real-world clinical deployment. Nevertheless, the present results establish a reliable methodological foundation for such future translational studies.

The hybrid deep-ensemble framework that is proposed in this paper shows a good, consistent, and clinically viable performance in detecting breast cancer in its early stages with a high level of discrimination, low false-positive probability, and statistically stable comparative performance, which adds evidence to the translational viability in the use of the framework as a decision-support system in breast cancer diagnosis. Although the proposed framework emphasizes interpretability at the system and feature-integration level, no explicit feature-attribution or explainability analysis of individual latent dimensions was performed, which represents an important direction for future work.

CONCLUSION

Early and reliable identification of breast malignancy remains central to improving patient outcomes and optimizing downstream diagnostic pathways. In this study, a hybrid deep–ensemble decision-support framework (HDEL-BC) was developed and evaluated with a specific focus on diagnostic safety, stability, and clinical usability rather than isolated performance optimization. By integrating autoencoder-derived latent representations with standardized, clinically interpretable diagnostic features, the proposed approach provides a balanced representation of both hidden morphological patterns and established radiological descriptors relevant to breast cancer assessment.

The framework demonstrated consistently high specificity, low false-positive rates, and strong discriminative ability across stratified cross-validation, supporting its potential role as a screening-support tool aimed at minimizing unnecessary follow-up procedures and patient anxiety. Statistical analysis further indicated robust performance relative to commonly used baseline models, reinforcing the reliability of the framework under repeated validation rather than dependence on single-split performance.

From an oncology perspective, the proposed model is computationally efficient and interpretable, making it a promising decision-support prototype for further translational research rather than an immediately deployable clinical screening system. While the demonstrates consistently strong diagnostic performance under benchmark validation, supporting its value as a proof-of-concept decision-support framework, future studies will involve external datasets and prospective clinical evaluation are essential to fully establish real-world screening applicability.

Informed Consent: Informed consent was no need Informed consent.

Conflict of Interest Statement: No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article. All authors declare no conflict of interest for this article.

Funding: No Funding support is provided for this paper publication.

Use of AI for Writing Assistance: No AI technologies utilized.

Author Contributions: Concept – P.P., K.U., P.K., G.V.R., S.B.K., A.K., R.J.; Design – P.P., K.U., P.K., G.V.R., S.B.K., A.K., R.J.; Materials – P.P., K.U., P.K., G.V.R., S.B.K., A.K., R.J.; Data collection and/or processing – P.P., K.U., P.K., G.V.R., S.B.K., A.K., R.J.; Data analysis and/or interpretation – P.P., K.U., P.K., G.V.R., S.B.K., A.K., R.J.; Literature search – P.P., K.U., P.K., G.V.R., S.B.K., A.K., R.J.; Writing – P.P., K.U., P.K., G.V.R., S.B.K., A.K., R.J.; Critical review – P.P., K.U., P.K., G.V.R., S.B.K., A.K., R.J.

Acknowledgments: The authors gratefully acknowledge all individuals and institutions who provided support and resources for the successful completion of this work.

Peer-review: Externally peer-reviewed.

REFERENCES

1. Kinkar KK, Fields BKK, Yamashita MW, Varghese BA. Empowering breast cancer diagnosis and radiology practice: Advances in artificial intelligence for contrast-enhanced mammography. Front Radiol 2024;3:1326831
2. Nassif AB, Talib MA, Nasir Q, Afadar Y, Elgendy O. Breast cancer detection using artificial intelligence techniques: A systematic literature review. Artif Intell Med 2022;127:102276
3. Akhtar NDN, Pant NDH, Dwivedi NA, Jain NV, Perwej NDY. A breast cancer diagnosis framework based on machine learning. Int J Sci Res Sci Eng Technol 2023;118–32
4. Alam T, Shia W, Hsu F, Hassan T. Improving breast cancer detection and diagnosis through semantic segmentation using the Unet3+ deep learning framework. Biomedicines 2023;11(6):1536
5. Zheng J, Lin D, Gao Z, Wang S, He M, Fan J. Deep learning assisted efficient AdABoost algorithm for breast cancer detection and early diagnosis. IEEE Access 2020;8:96946–54
6. Khalid A, Mehmood A, Alabrah A, Alkhamees BF, Amin F, AlSalman H, et al. Breast cancer detection and prevention using machine learning. Diagnostics 2023;13(19):3113
7. Humayun M, Khalil MI, Almuayqil SN, Jhanjhi NZ. Framework for detecting breast cancer risk presence using deep learning. Electronics 2023;12(2):403
8. Michael E, Ma H, Li H, Qi S. An optimized framework for breast cancer classification using machine learning. Biomed Res Int 2022;2022(1):8482022
9. Mehmood M, Ayub E, Ahmad F, Alruwaili M, Alrowaili ZA, Alanazi S, et al. Machine learning enabled early detection of breast cancer by structural analysis of mammograms. Comput Mater Continua 2021;67(1):641–57
10. Ghorbian M, Ghorbian S. Usefulness of machine learning and deep learning approaches in screening and early detection of breast cancer. Heliyon 2023;9(12):e22427
11. Chugh G, Kumar S, Singh N. Survey on machine learning and deep learning applications in breast cancer diagnosis. Cogn Comput 2021;13(6):1451–70
12. Lahoura V, Singh H, Aggarwal A, Sharma B, Mohammed MA, Damaševičius R, Kadry S, Cengiz K. Cloud computing-based framework for breast cancer diagnosis using extreme learning machine. Diagnostics 2021;11(2):241
13. Naseem U, Rashid J, Ali L, Kim J, Haq QEU, Awan MJ, Imran M. An automatic detection of breast cancer diagnosis and prognosis based on machine learning using ensemble of classifiers. IEEE Access 2022;10:78242–52
14. Kumar M, Singhal S, Shekhar S, Sharma B, Srivastava G. Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning. Sustainability 2022;14(21):13998
15. Wang X, Ahmad I, Javeed D, Zaidi S, Alotaibi F, Ghoneim M, et al. Intelligent hybrid deep learning model for breast cancer detection. Electronics 2022;11(17):2767
16. Kaggle. Breast cancer Wisconsin (diagnostic) data set. Available at: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?resource=download. Accessed Feb 4, 2025.