International Journal of Research in Computer & Information Technology (IJRCIT),
Vol.10, Issue 3, June 2024
E-ISSN: 2455-3743
Available online at www.ijrcit.co.in

Ensemble-Driven Machine Learning Models for Robust Breast Cancer Recurrence Prediction

[1]Prachi Damodhar Shahare, [2]Dr. Abha Mahalwar

Abstract— Breast cancer recurrence remains a major clinical challenge despite advances in early diagnosis and treatment. Accurate prediction of recurrence can support personalized follow-up strategies and improve patient outcomes. This study investigates the effectiveness of multiple machine learning models for breast cancer recurrence prediction using clinical and tumor-related features. Decision Tree, Logistic Regression, Support Vector Machine (RBF), Random Forest, and a soft-voting Ensemble model are developed and evaluated. Model performance is assessed using several evaluation parameters. Experimental results demonstrate that while Logistic Regression achieves the highest ROC–AUC (0.702), the proposed Ensemble model provides the most balanced performance, yielding superior recall (0.6118), F1-score (0.5828), and MCC (0.6257). Calibration analysis further confirms improved probability reliability for the ensemble approach. The findings highlight the benefit of ensemble learning for robust and clinically reliable breast cancer recurrence prediction.

Keywords— about four key words separated by commas

I. Introduction

Breast cancer is one of the most commonly diagnosed malignancies among women worldwide and remains a leading cause of cancer-related mortality [1]. Although advancements in screening, surgery, radiotherapy, chemotherapy, and targeted therapies have significantly improved survival rates, a substantial proportion of patients experience disease recurrence after initial treatment [2]. Recurrence can occur locally, regionally, or at distant metastatic sites and is often associated with poor prognosis and increased healthcare burden. Therefore, early and accurate prediction of breast cancer recurrence is crucial for optimizing post-treatment surveillance and personalized clinical decision-making [3].

Traditional prognostic models primarily rely on clinic pathological factors such as tumor size, lymph node involvement, hormone receptor status, and histological grade [4]. While these factors are informative, they often fail to capture complex, non-linear relationships among variables that influence recurrence risk. In recent years, machine learning (ML) techniques have emerged as powerful tools for modeling such relationships, enabling improved prediction accuracy and enhanced risk stratification [5].

Several studies have demonstrated the potential of ML-based approaches for breast cancer recurrence prediction, employing algorithms such as Logistic Regression, Support Vector Machines, Random Forests, gradient boosting methods, and deep learning models [6]. However, individual models may suffer from bias, instability, or limited generalization. Ensemble learning, which combines multiple classifiers, has been shown to improve robustness and predictive performance by leveraging complementary strengths of individual models [7].

Motivated by these observations, this study presents a comparative evaluation of multiple machine learning models and proposes an ensemble-based approach for breast cancer recurrence prediction. The contributions of this work are threefold:

(i)a comprehensive comparative analysis of classical ML models using multiple evaluation metrics beyond accuracy,

(ii) (ii) the development of an ensemble model to enhance predictive robustness, and

(iii) (iii) an in-depth assessment of model reliability through ROC, Precision–Recall, and calibration curve analyses.

The remainder of this paper is organized as follows. Section II reviews recent and relevant literature on breast cancer recurrence prediction using machine learning techniques, highlighting existing approaches and their limitations. Section III present the complete methodology. Section IV presents result analysis of the proposed ensemble model. Section V concludes the paper and outlining potential directions for future research

II. Related Work

Lauritzen, A. D et al. (2023) included patient data from the Danish Breast Cancer Group (DBCG), the National Pathology Database, and the National Patient Registry for patients diagnosed with invasive breast cancer after 1999. A total of 79,483 patients with definitive surgery were analysed to extract relevant features for the machine learning model. The machine learning model was trained on a development sample consisting of 5,333 patients with known recurrence and three times as many non-recurrent patients. The model's performance was validated using a separate validation sample of 1,006 patients with unknown

Manuscript Received April 5, 2025; Revised 25 May, 2025 and published on June 28, 2025
Prachi Damodhar Shahare, Name of *Department of Computer Science, lSBM University, Chhattisgarh, India.* ***Mail Id:*** *prachi_shahare@yahoo.com*
Dr. Abha Mahalwar, Name of *Department of Computer science, lSBM University, Chhattisgarh, India,* ***Mail Id:*** *tamrakar.abha@gmail.com*

recurrence status. The model achieved an AUC-ROC of 0.93 in the development sample and 0.86 in the validation sample, indicating its effectiveness in identifying recurrence patients [8].

Noman et al. (2025) developed and validated predictive models for breast cancer recurrence and metastasis using Recurrence-Free Survival Analysis and machine learning techniques. A comprehensive dataset was created by merging data from multiple sources, resulting in 190,789 rows and 23 columns. Key prognostic factors were identified through survival analysis, enhancing the understanding of recurrence risks. The LightGBM model achieved an impressive AUC of 92% in predicting recurrences, while XGBoost and Random Forest models distinguished between recurrence types with up to 86% accuracy [9].

Berg, T. (2023) identifies 10–30% of breast cancer patients in Denmark experience recurrence despite adequate treatment. An automated identification method for recurrent breast cancer patients was developed using machine learning (ML) models. The study utilized data from the Danish Breast Cancer Group, National Pathology Database, and National Patient Registry, involving 79,483 patients with invasive breast cancer diagnosed after 1999. The ML model achieved an AUC-ROC of 0.93 in the development sample and 0.86 in the validation sample, indicating high accuracy in identifying recurrence [10].

Li, J. (2024) study found that Logistic Regression (LR) and Support Vector Machines (SVM) outperformed other machine learning algorithms in predicting breast cancer recurrences, achieving AUC values of 99.77% and 99.74%, respectively. Both LR and SVM demonstrated high accuracy rates of 97.37%, precision of 97.62%, and recall of 95.35%, indicating their effectiveness in diagnosing breast cancer. The F1 score for LR and SVM was reported at 96.47, showcasing a well-balanced precision-recall performance. The algorithms exhibited high Cohen's Kappa scores, with LR and SVM achieving a score of 94.37, reflecting strong agreement between predicted and observed classifications [11].

Mengad, A et al. (2023) compared the performance of four machine learning techniques: logistic regression, decision tree, K-Nearest Neighbors, and artificial neural networks in predicting types of breast cancer recurrence. The artificial neural network algorithm achieved the highest accuracy of 91%, outperforming the other algorithms. The decision tree algorithm followed with an accuracy of 90.10%, while K-Nearest Neighbors and logistic regression had accuracies of 88.20% and 84.60%, respectively. The research highlighted the importance of incorporating various predictive variables, including psychological and behavioral aspects, to improve risk assessment and patient management [12].

Jiang, X. et al. (2025) developed an interpretable machine learning pipeline to predict distant recurrence-free survival in breast cancer patients at 5, 10, and 15 years. The best models achieved area under the curve (AUC) scores of 0.79, 0.83, and 0.89 for the respective timeframes, significantly outperforming traditional models. The Markov blanket and interactive risk factor learner (MBIL) reduced input dimensionality by over 80% while maintaining accuracy, indicating effective feature selection. Features selected by MBIL, such as nodal status and hormone receptor expression, were strongly correlated with top contributors identified by SHAP, enhancing interpretability. The performance improvement due to grid search optimization ranged from 25.3% to 60%, highlighting the effectiveness of hyperparameter tuning in model accuracy [13].

González-Castro et al. (2023) study found that machine learning algorithms can effectively predict 5-year breast cancer recurrence using data from electronic health records. The XGB (eXtreme Gradient Boosting) model demonstrated the best performance among the five evaluated algorithms, achieving a precision of 0.900, recall of 0.907, F1-score of 0.897, and AUROC of 0.807. It was noted that structured data provided the best prediction results, while the combined dataset of structured and unstructured data yielded the poorest performance. The research highlighted the potential of machine learning tools to improve patient risk stratification and assist in personalized treatment planning [14].

Cartron, M. (2022) research focused on predicting breast cancer recurrence within five years using various machine learning models, including logistic regression, support vector machine (SVM), decision tree, and random forest. The random forest model outperformed other models, achieving a precision of 0.75, accuracy of 0.69, recall of 0.66, C-index of 0.71, and F1-score of 0.70. Shapley additive explanations (SHAP) were utilized to interpret the random forest model, identifying the contribution of each feature to cancer recurrence predictions. Key features influencing cancer recurrence included tumor size, mutation count, positive lymph nodes examined, age at diagnosis, Nottingham prognostic index (NPI), tumor stage, HER2 status, and cancer type [15].

III. METHODOLOGY

Figure 1 shows the systematic machine learning pipeline consisting of data preprocessing, model training, hyperparameter optimization, and performance evaluation. Initially, the dataset is prepared by encoding categorical variables and scaling numerical features to ensure compatibility with machine learning algorithms. Multiple baseline models—Decision Tree (DT), Logistic Regression (LR), Support Vector Machine with RBF kernel (SVM-RBF), and Random Forest (RF)—are trained using the processed data. Hyperparameter tuning is performed using grid search with cross-validation to identify optimal configurations for each model. To improve prediction robustness, a soft-voting Ensemble model is constructed by combining the probabilistic outputs of selected base classifiers. The final predictions are obtained by averaging class probabilities and selecting the class with the highest confidence.

International Journal of Research in Computer & Information Technology (IJRCIT),
Vol. 10, Issue 3, June 2024
E-ISSN: 2455-3743
Available online at www.ijrcit.co.in

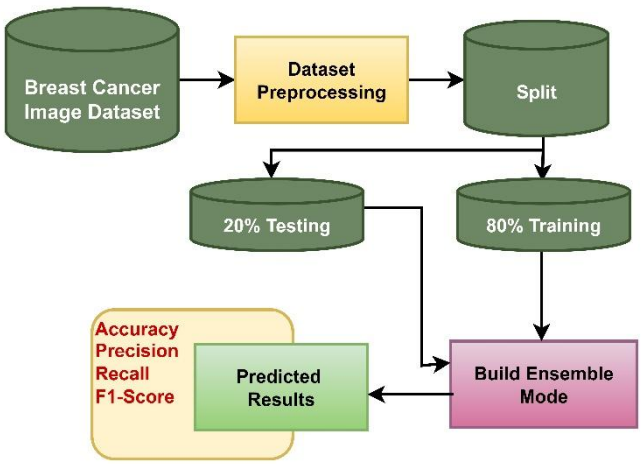

Figure 1: Architecture of Proposed Model

Dataset Description: The dataset used in this study consists of 228 patient records, including 160 non-recurrence cases and 68 recurrence cases, reflecting a naturally imbalanced clinical distribution. Each record contains clinically relevant features related to patient demographics, tumor characteristics, and treatment outcomes. Among these features, tumor size plays a critical role in recurrence risk stratification, as demonstrated through exploratory analysis [16].

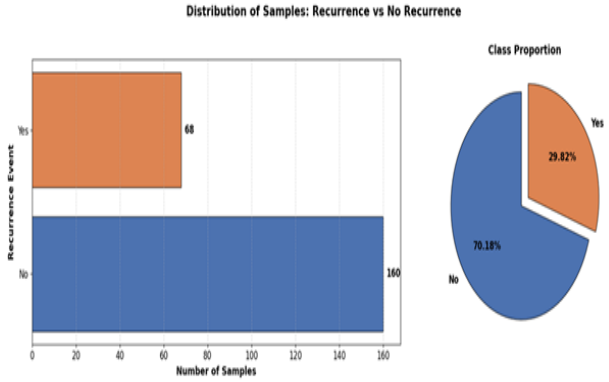

Figure 2: Distribution of samples showing class imbalance between recurrence and non-recurrence cases in the breast cancer dataset.

Figure 2 shows the distribution of samples across recurrence and non-recurrence classes in the breast cancer dataset. The horizontal bar chart shows that 160 samples belong to the No Recurrence class, while 68 samples correspond to the Recurrence class, indicating a clear class imbalance. This imbalance is further highlighted in the pie chart, where the No Recurrence category accounts for 70.18% of the total samples, whereas the Recurrence category represents only 29.82%.

Data preprocessing: Data preprocessing is a crucial step to ensure model reliability and consistency. Missing values, if present, are handled using appropriate imputation strategies based on feature type. Categorical variables are transformed into numerical representations using label encoding, while numerical features are scaled to a standardized range to prevent dominance of high-magnitude variables. The dataset is then split into training and testing subsets to ensure unbiased evaluation. All preprocessing steps are applied consistently across training and testing data to prevent information leakage.

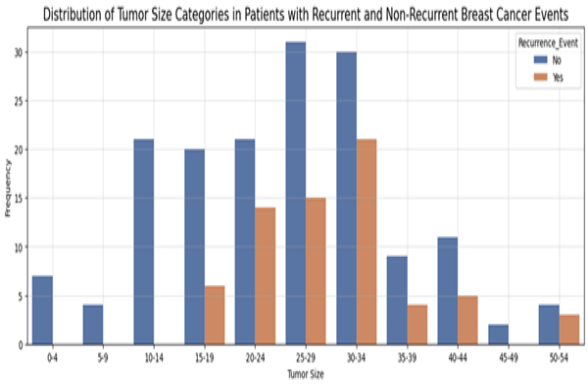

Figure 3: Distribution of tumor size categories for recurrent and non-recurrent breast cancer cases

Figure 3 depicts the distribution of tumor size categories among patients with recurrent and non-recurrent breast cancer events. Non-recurrence cases dominate across most tumor size ranges, with the highest frequencies observed in the 25–29 mm (31 cases) and 30–34 mm (30 cases) categories, followed by 10–14 mm (21 cases), 15–19 mm (20 cases), and 20–24 mm (21 cases). Recurrence cases show a different pattern, peaking in the 30–34 mm (21 cases) and 25–29 mm (15 cases) ranges, with moderate counts in 20–24 mm (14 cases) and 15–19 mm (6 cases). Very small tumors (0–9 mm) and larger tumors above 40 mm exhibit comparatively fewer cases in both groups, although recurrence events remain present even at higher tumor sizes (5 cases in 40–44 mm and 3 cases in 50–54 mm).

Building the Ensemble Model:

To enhance prediction robustness and reduce the bias associated with individual classifiers, the outputs of all baseline models are combined using a soft-voting ensemble strategy. In this approach, the predicted class probabilities from DT, LR, SVM (RBF), and RF are aggregated by computing their average, and the final class label is determined based on the highest combined probability. This fusion mechanism allows the ensemble model to leverage the strengths of diverse learning paradigms, including linear, non-linear, and tree-based methods, thereby improving generalization performance. The final ensemble prediction categorizes patients into either Recurrence or Non-Recurrence, as depicted in the figure. By integrating multiple predictive perspectives, the ensemble model achieves more stable, reliable, and clinically meaningful predictions compared to individual baseline models.

International Journal of Research in Computer & Information Technology (IJRCIT),
Vol. 10, Issue 3, June 2024
E-ISSN: 2455-3743
Available online at www.ijrcit.co.in

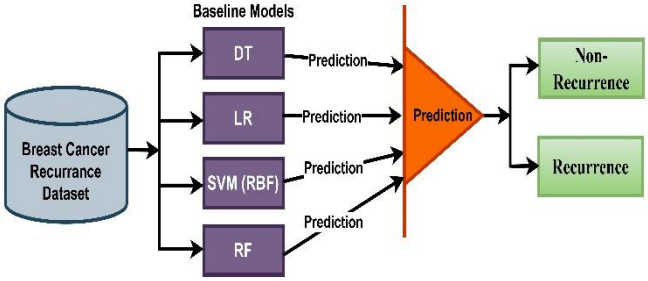

Figure 4: Ensemble Model

Figure 4 shows the architecture of the proposed ensemble learning framework developed for breast cancer recurrence prediction. The process begins with the breast cancer recurrence dataset, which is supplied in parallel to multiple baseline machine learning classifiers, namely Decision Tree (DT), Logistic Regression (LR), Support Vector Machine with RBF kernel (SVM-RBF), and Random Forest (RF). Each of these models is independently trained using the same preprocessed feature set and optimized hyperparameters to learn complementary decision boundaries from the data. During inference, each baseline model generates a probabilistic prediction indicating the likelihood of recurrence or non-recurrence for a given patient instance.

IV. Result Analysis

The experimental evaluation of the proposed breast cancer recurrence prediction models was conducted using Google Colab, using its cloud-based computational environment for efficient model training and reproducibility. All experiments were implemented in Python using standard machine learning libraries, including scikit-learn and NumPy. The dataset was partitioned into training and testing subsets using a stratified split to preserve the original class distribution, with 80% of the samples used for training and 20% reserved for testing. Hyperparameter optimization for individual classifiers, such as DT, LR, SVM (RBF), and RF, was performed using GridSearchCV with 5-fold cross-validation, ensuring robust parameter selection and reduced overfitting. The ensemble model was constructed using a soft-voting strategy, combining probabilistic outputs of optimized base classifiers. Model performance was evaluated on the unseen test set using multiple metrics to provide a comprehensive and unbiased assessment of predictive accuracy, class discrimination, and reliability.

$$Accuracy = \frac{T\_P + T\_N}{T\_P + T\_N + F\_P + F\_N} \quad (1)$$

$$Precision = \frac{T\_P}{T\_P + F\_P} \quad (2)$$

$$Recall = \frac{T\_P}{T\_P + F\_N} \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Table 1: Performance Analysis of proposed Models

Models	Accuracy	Precision	Recall	F1-score	MCC	Kappa
DT	0.6724	0.4286	0.3529	0.3871	0.1679	0.1664
LR	0.7414	0.6000	0.3529	0.4444	0.3077	0.2904
SVM (RBF)	0.7241	1.0000	0.5088	0.5111	0.5057	0.5012
RF	0.7069	0.5000	0.5941	0.5704	0.5075	0.4258
Ensemble	**0.7414**	**0.5833**	**0.6118**	**0.5828**	**0.6257**	**0.3171**

Table 1 presents a comparative performance analysis of the proposed machine learning and ensemble models for breast cancer recurrence prediction. The DT model exhibits the weakest performance across all metrics, indicating limited discriminative capability. The LR improves overall accuracy but suffers from low recall, reflecting its inability to effectively identify recurrence cases. The SVM (RBF) model achieves perfect precision (1.0000), demonstrating strong confidence in positive predictions; however, its moderate recall limits overall sensitivity. The RF model provides a more balanced trade-off between precision and recall, resulting in improved F1-score and agreement metrics.The proposed Ensemble model achieves the best overall performance, yielding the highest recall (0.6118), F1-score (0.5828), and MCC (0.6257), indicating superior robustness and class-wise agreement.

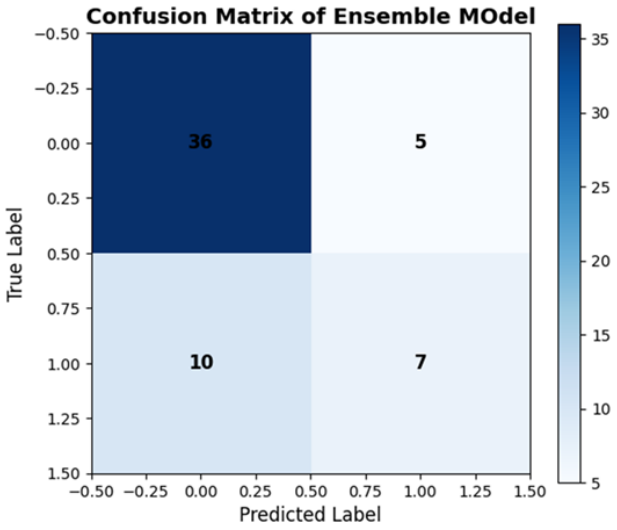

Figure 5: Confusion Matrix of Ensemble Model

Figure 5 shows the confusion matrix of the proposed Ensemble model and its classification behavior for breast cancer recurrence prediction. Out of the total samples, the model correctly classified 36 non-recurrence cases (true negatives) and 7 recurrence cases (true positives), indicating reliable performance for both classes. However, 5 non-recurrence cases were misclassified as recurrence (false positives), while 10 recurrence cases were incorrectly predicted as non-recurrence (false negatives), highlighting a

International Journal of Research in Computer & Information Technology (IJRCIT),
Vol. 10, Issue 3, June 2024
E-ISSN: 2455-3743
Available online at www.ijrcit.co.in

moderate tendency to miss some recurrence instances.

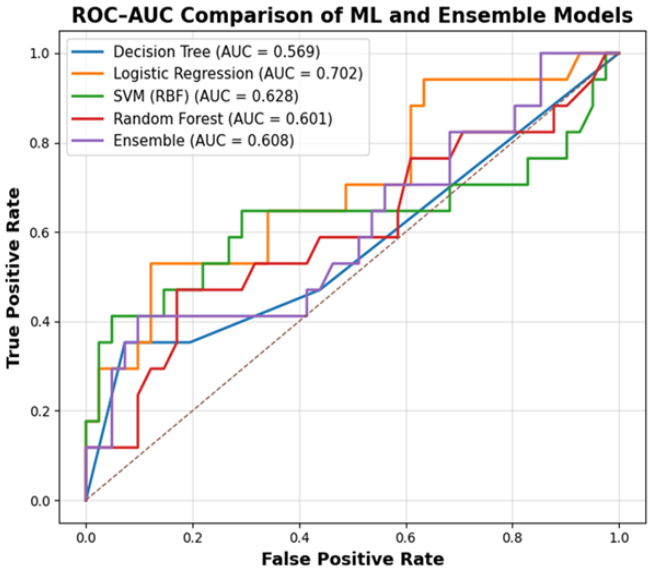

Figure 6: ROC-AUC Curve of proposed Models

Figure 6 shows the ROC–AUC comparison performance of different ML and ensemble models for breast cancer recurrence prediction. Among all evaluated approaches, LR achieves the highest area under the curve with an AUC of 0.702, indicating superior overall class separation capability. The SVM (RBF) model follows with an AUC of 0.628, showing moderate discrimination, while the Ensemble model attains an AUC of 0.608, reflecting improved robustness over single weak learners but limited gain in overall separability. The RF and DT models record lower AUC values of 0.601 and 0.569, respectively, suggesting comparatively weaker performance.

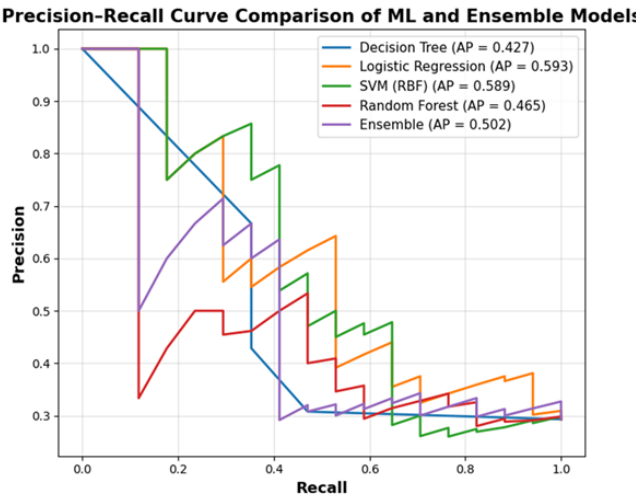

Figure 7: PR Curve of proposed Models

Figure 7 shows the PR curve comparison highlights the class-wise detection capability of the evaluated machine learning and ensemble models for breast cancer recurrence prediction. Logistic Regression achieves the highest average precision with an AP of 0.593, indicating the most stable precision–recall trade-off across varying thresholds. The SVM (RBF) model closely follows with an AP of 0.589, demonstrating competitive performance in identifying recurrence cases at moderate recall levels. The proposed Ensemble model attains an AP of 0.502, reflecting improved robustness over individual tree-based models by balancing precision and recall more effectively. In contrast, the Random Forest and Decision Tree models yield lower AP values of 0.465 and 0.427, respectively, indicating weaker performance under class imbalance conditions.

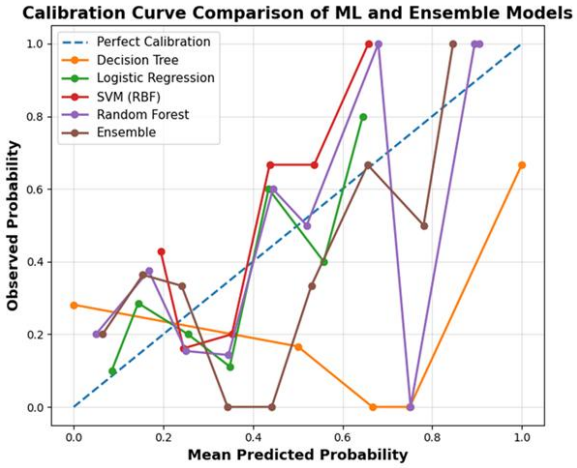

Figure 8: Calibration Curve of proposed Models

Figure 8 shows the calibration curve comparison illustrates the reliability of predicted probabilities produced by different machine learning and ensemble models for breast cancer recurrence prediction. The dashed diagonal represents perfect calibration, against which model behaviors can be assessed. Logistic Regression shows relatively stable calibration in the low-to-mid probability range, with observed probabilities increasing from approximately 0.10–0.30 to around 0.80 as predicted probabilities rise toward 0.60, indicating reasonable probability alignment. The SVM (RBF) model exhibits sharper transitions, with observed probabilities reaching 1.00 at predicted values near 0.65, but showing overconfidence in some mid-range bins. The Random Forest displays notable miscalibration, with abrupt drops to 0.00 around predicted probabilities of 0.70–0.75, followed by perfect outcomes (1.00) at higher probabilities, reflecting unstable probability estimates. The Decision Tree model is poorly calibrated, with observed probabilities fluctuating between 0.00 and 0.67, including underestimation at high predicted probabilities. In contrast, the proposed Ensemble model demonstrates improved reliability, with observed probabilities progressing from 0.20 at low confidence to 0.65–1.00 at higher predicted probabilities, aligning more closely with the ideal calibration line and indicating more trustworthy probability estimates overall

CONCLUSION & FUTURE SCOPE

This study demonstrates the effectiveness of machine

learning techniques for breast cancer recurrence prediction through a comprehensive comparative and ensemble-based analysis. Experimental results indicate that while individual models such as Logistic Regression and SVM exhibit strong discriminative capability, the proposed Ensemble model achieves the most balanced performance in terms of recall, F1-score, and Matthews Correlation Coefficient, along with improved probability calibration, making it more suitable for clinical decision support under class imbalance conditions. The integration of multiple evaluation metrics, ROC–AUC, Precision–Recall, and calibration analysis further strengthens the reliability of the findings. Future research will focus on incorporating larger multi-institutional datasets, survival-based time-to-event modeling, advanced feature selection and explainability techniques such as SHAP, and deep learning or hybrid models to enhance predictive accuracy and clinical interpretability, ultimately supporting personalized breast cancer management and follow-up strategies.

REFERENCES

[1] Wilkinson L., and Gathani T. (2022). Understanding breast cancer as a global health concern. The British Journal of Radiology, 95(1130), 20211033. https://doi.org/10.1259/bjr.20211033

[2] Shahi A., Mishra S., and Rath M. (2025). Early breast cancer—Recent advances in imaging and management. International Journal of Medical and Biomedical Studies, 9, 37–57. https://doi.org/10.32553/ijmbs.v9i4.3097

[3] Shi, Y., Olsson, L. T., Hoadley, K. A., Calhoun, B. C., Marron, J. S., Geradts, J., Niethammer, M., & Troester, M. A. (2023). Predicting early breast cancer recurrence from histopathological images in the Carolina Breast Cancer Study. NPJ Breast Cancer, 9. https://doi.org/10.1038/s41523-023-00597-0

[4] Goel H., and Kumar D. (2023). Data mining in healthcare using machine learning techniques. In 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS) (pp. 25–29). https://doi.org/10.1109/ICSCSS57650.2023.10169802

[5] El Haji, H., Souadka, A., Patel, B. N., Sbihi, N., Ramasamy, G., Patel, B. K., Ghogho, M., and Banerjee, I. (2023). Evolution of breast cancer recurrence risk prediction: A systematic review of statistical and machine learning-based models. JCO Clinical Cancer Informatics, 7, e2300049. https://doi.org/10.1200/CCI.23.00049

[6] Rajendran, T., Aswini, J., Rajathi, S. A., Prakash, R. B., Balakrishnan, C., & Subramanian, S. (2023). Risk Prediction Modeling for Breast Cancer using Supervised Machine Learning Approaches. 702–708. https://doi.org/10.1109/icacrs58579.2023.10404482

[7] Almuhaidib, D. A., Shaiba, H., Alharbi, N. G., Alotaibi, S. M., Albusayyis, F. M., Alzaid, M. A., & Almadhi, R. M. (2018). Ensemble Learning Method for the Prediction of Breast Cancer Recurrence. https://doi.org/10.1109/CAIS.2018.8442017

[8] Lauritzen, A. D., Jensen, M.-B., Lillholm, M., & Knoop, A. (2023). Identifying recurrent breast cancer patients in national health registries using machine learning. Acta Oncologica, 62, 350–357. https://doi.org/10.1080/0284186X.2023.2201687

[9] Noman, S. M., Fadel, Y. M., Attia, N., Tarek, M., Essam, M., ElMasarawi, S. M., Fouad, F. A., GEltasawi, E., & Al-Atabany, W. (2025). Leveraging Survival Analysis and Machine Learning for Accurate Prediction of Breast Cancer Recurrence and Metastasis. https://doi.org/10.21203/rs.3.rs-5059228/v1

[10] Berg, T. (2023). Identifying recurrent breast cancer patients in national health registries using machine learning. Acta Oncologica, 62(4), 350–357. https://doi.org/10.1080/0284186x.2023.2201687

[11] Li, J. (2024). Evaluative Comparison of Machine Learning Algorithms for Precision Diagnosis in Breast Cancer. Highlights in Science Engineering and Technology, 85, 354–362. https://doi.org/10.54097/40fmfw48

[12] Mengad, A., & Amali, S. M. J. (2023). A Comparative Study of Machine Learning Techniques to Predict Types of Breast Cancer Recurrence. International Journal of Advanced Computer Science and Applications, 14(5). https://doi.org/10.14569/ijacsa.2023.0140531

[13] Jiang, X., Zhou, Y., Wells, A., & Brufsky, A. (2025). Leveraging Deep Learning, Grid Search, and Bayesian Networks to Predict Distant Recurrence of Breast Cancer. Cancers, 17(15), 2515. https://doi.org/10.3390/cancers17152515

[14] González-Castro, L., Chavez, M., Duflot, P., Bleret, V., Martin, A., Zobel, M., Nateqi, J., Lin, S., Pazos-Arias, J. J., Del Fiol, G., & López-Nores, M. (2023). Machine Learning Algorithms to Predict Breast Cancer Recurrence Using Structured and Unstructured Sources from Electronic Health Records. Cancers, 15(10), 2741. https://doi.org/10.3390/cancers15102741

[15] Cartron, M. (2022). Prediction of Breast Cancer Recurrence in Five Years using Machine Learning Techniques and SHAP. Lecture Notes in Electrical Engineering, 441–453. https://doi.org/10.1007/978-981-19-0252-9_40

[16] Teng, J. (2019). SEER breast cancer data [Data set]. IEEE DataPort. https://doi.org/10.21227/a9qy-ph35