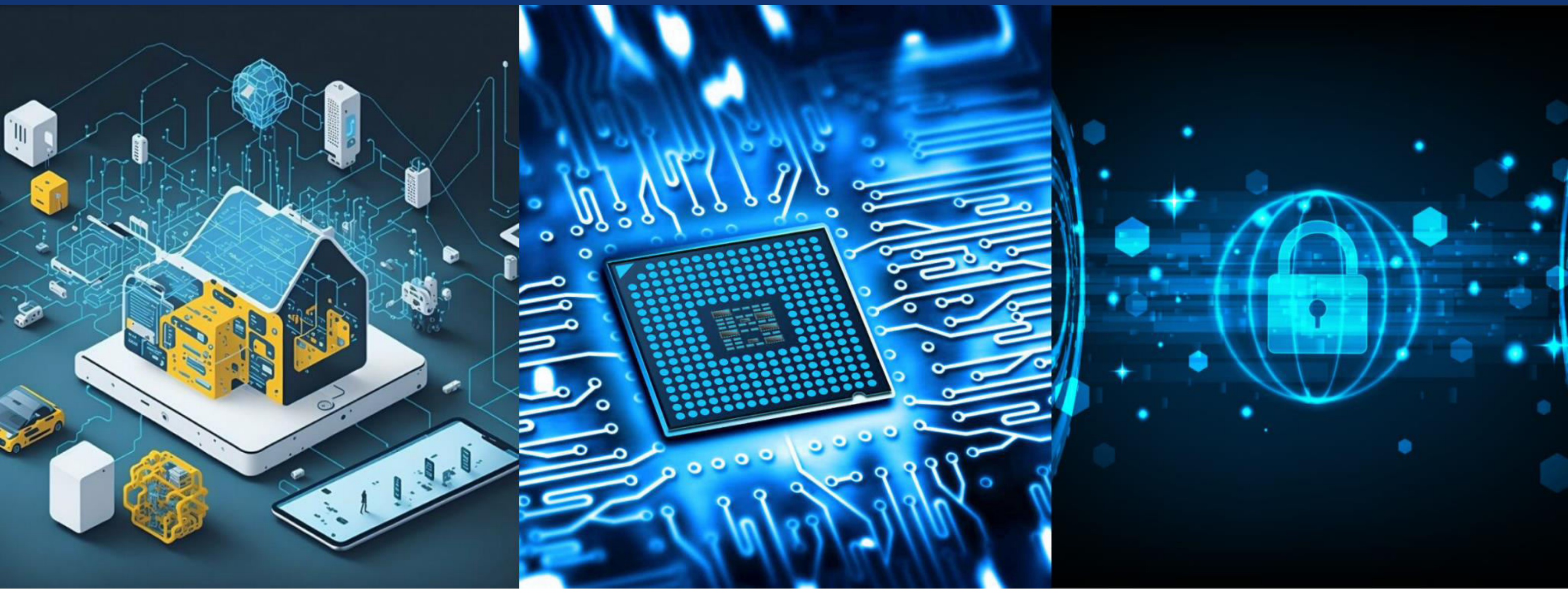
IJIRCCE

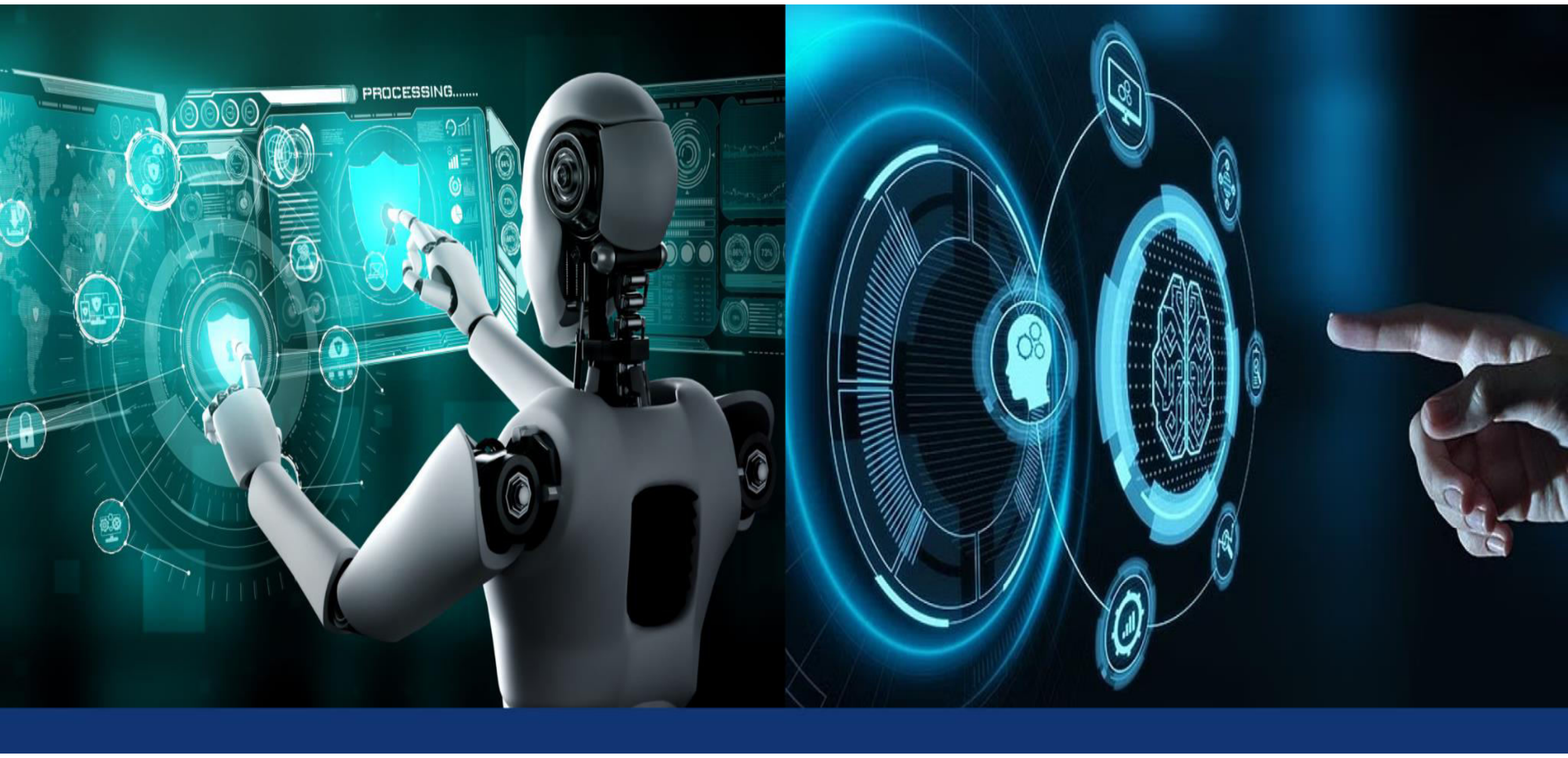
ISSN(O): 2320-9801

ISSN(P): 2320-9798



# International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 3, March 2025

www.ijircce.com ijircce@gmail.com +91-9940572462 +91 63819 07438 www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

# Optimized Ensemble Learning For Advanced Breast Cancer Detection Accuracy

**Aishwarya Patil[1], Manasi Sabale[2], Dr.Pravin Game[3], Dr.Mubin Tamboli[4]**

UG Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India[1,2]

Associate ProfessorDepartment of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India[3,4]

**ABSTRACT**: Early breast cancer detection is crucial for improving patient outcomes, yet remains a challenging task. This research enhances breast cancer classification by analyzing the Wisconsin Breast Cancer Dataset and combining four machine learning models: Logistic Regression, Support Vector Classifier, Random Forest, and XGBoost. Each model was fine-tuned using Bayesian optimization to maximize performance. To improve prediction accuracy and reliability, an ensemble system was created with a weighted voting classifier, assigning more weight to better-performing models. The resulting model achieved 98.25% accuracy, demonstrating the power of ensemble learning and optimization techniques in developing efficient, and reliable tools for breast cancer diagnosis.

**KEYWORDS**: Wisconsin Breast Cancer Dataset (WBCD), XGBoost (eXtreme Gradient Boosting), Support Vector Machines (SVM), Support Vector Classifier (SVC), Bayesian Optimization.

## I. INTRODUCTION

Breast cancer is one of the most common cancers globally and a leading cause of death among women. Early detection significantly improves treatment outcomes. With numerous diagnostic methods available, the complexity of identifying malignancy in breast tissue has led to the increasing use of advanced computational techniques, especially machine learning, to assist in diagnosis. Machine learning models have shown great promise in automating and improving breast cancer detection, helping reduce human error and accelerate diagnosis.

### A. Machine Learning in Breast Cancer Detection

Machine learning has been widely applied in healthcare, particularly for diagnostic purposes, by identifying patterns in complex datasets. In breast cancer, models are trained on datasets containing features related to tumor characteristics, such as size, shape, and texture. The Wisconsin Breast Cancer Dataset (WBCD) is one such dataset, providing valuable features for training classification models. Popular algorithms include Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), and XGBoost. RF is effective in handling large datasets and preventing overfitting, SVM performs well in binary classification tasks, XGBoost is known for its computational efficiency, and LR is useful in ensemble models.

### B. Ensemble Learning for Improved Accuracy

Ensemble learning combines multiple models to improve performance. By leveraging the strengths of individual models and minimizing weaknesses, ensemble methods like voting classifiers enhance predictive power. Soft voting, which averages predicted probabilities, has been particularly successful in improving accuracy in breast cancer detection. Ensemble methods reduce bias and improve precision and recall, which is critical in medical diagnostics to minimize false negatives and positives, ensuring appropriate care.

### C. Bayesian Optimization for Hyperparameter Tuning

Hyperparameter tuning is vital for optimizing machine learning models. Traditional methods like grid search can be computationally expensive and inefficient. Bayesian Optimization (BO) provides a more efficient alternative by predicting optimal hyperparameters based on previous evaluations. This method is particularly useful in fine-tuning models such as SVM, RF, and XGBoost, improving performance while reducing computational resources. This paper applies BO to optimize hyperparameters of various models, aiming to enhance breast cancer detection.

**D. DeepCME for Metastasis Prediction**

The DeepCME framework predicts breast cancer metastasis using gene expression data. By incorporating regularization techniques like L1 regularization, batch normalization, and dropout, DeepCME selects significant genes and reduces overfitting. With an average AUC score of 0.754, it outperforms baseline models. However, its high computational demands may limit broader application. Additionally, the use of Knowledge Distillation for histopathology image analysis demonstrates a balance between accuracy and efficiency, improving the real-time diagnosis of breast cancer.

**E. Specific Contributions**

This paper makes the following contributions:

- Applying an ensemble approach combining SVC, RF, XGBoost, and LR for breast cancer classification.
- Implementing Bayesian Optimization for hyperparameter tuning to enhance performance and reduce computational time.
- Evaluating the ensemble model's performance using accuracy, precision, recall, and F1-score metrics.
- Analyzing the impact of weighted voting on ensemble model performance.

By combining ensemble methods, hyperparameter tuning, and weighted voting, this paper aims to advance breast cancer detection and contribute to the development of more accurate diagnostic tools.

## II. LITERATURE SURVEY

In [2] A novel approach for breast cancer classification integrated transfer learning with attention mechanisms to enhance feature extraction and improve the interpretability of deep learning models for histopathological images. By using modified convolutional neural networks (CNNs) and attention modules, the method achieved test accuracy rates of up to 99.6% on the BreakHis dataset [1]. The strengths of this method included high classification accuracy and effective handling of complex image features. However, its reliance on specialized datasets may have reduced generalizability. Incorporating attention mechanisms significantly improved diagnostic precision and model robustness. An Optimized Stacking Ensemble Learning (OSEL) model was proposed for breast cancer detection and classification. The work leveraged a dataset from the UCI repository, combining multiple machine learning classifiers, including both meta-classifiers and base-classifiers, to improve performance, as mentioned in [2]. The OSEL model achieved an accuracy of 99.45%, significantly outperforming traditional classifiers like AdaBoost and XGBoost. This research highlighted the value of ensemble learning in mitigating biases inherent in individual classifiers. However, the complexity of the model posed challenges, such as overfitting and the need for manual feature engineering. Overall, the work demonstrated the scalability and accuracy of stacking ensembles in predictive medicine, making it a reliable tool for early breast cancer detection and prognosis.

Machine learning algorithms, such as SVC and Logistic Regression, were used for detecting breast cancer, addressing the challenges of accurate tumor classification, as mentioned in [3]. The work employed Python for feature extraction, data visualization, and classifier implementation. Strengths included detailed data preprocessing and visualization techniques, achieving up to 96.5% accuracy with SVC. However, the work’s limitations included a lack of dataset diversity, which may have affected the model’s real-world application. The research emphasized the importance of data cleaning and feature scaling in enhancing model accuracy and classification reliability.

A novel approach for breast cancer detection was introduced by extracting convoluted features from CNNs, which were integrated into an ensemble voting classifier. This model combined Logistic Regression and Stochastic Gradient Descent to classify malignant and benign tumors with improved precision, as noted in [4]. The work achieved 100% accuracy, demonstrating the superiority of using deep learning features over traditional methods. Despite promising results, the model’s reliance on computationally expensive CNNs and potential scalability issues with smaller datasets were notable limitations. The work highlighted the strength of hybrid models that combined handcrafted and deep learning features, setting a benchmark in breast cancer prediction and encouraging future studies to explore similar hybrid methodologies for medical diagnostics.

The work proposed a hybrid approach for breast cancer detection using deep learning (ResNet50V2) for feature extraction and LightGBM for classification. The addressed problem included scalability and accuracy limitations in existing models, as discussed in [5]. The hybrid method improved accuracy (95%) and interpretability through

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

extensive performance analysis. Strengths lay in the model's robustness, use of publicly available datasets, and comprehensive evaluation metrics. However, the reliance on specific datasets may have limited generalization. Observations suggested that combining deep learning with machine learning enhanced diagnostic efficiency and supported informed decision-making in medical contexts.

## III. METHODOLOGY

This implementation utilizes an ensemble learning approach to enhance breast cancer detection using the Wisconsin Breast Cancer Dataset. The methodology involves the following key steps:

- **Ensemble Learning Approach:** We combine four classifiers—Logistic Regression (LR), Support Vector Classifier (SVC), Random Forest Classifier (RFC), and XGBoost—chosen for their proven effectiveness in classification tasks, especially in biomedical applications.
- **Data Preprocessing:**
  - The dataset is cleaned to handle missing values.
  - Feature scaling is applied using StandardScaler to ensure that all features are standardized, preventing algorithms like SVC and LR from being biased by features with larger magnitudes.
- **Hyperparameter Tuning:** Each model's hyperparameters are fine-tuned using Bayesian Optimization, which efficiently identifies the best configurations, reducing the time and effort compared to traditional methods like grid search.
- **Model Training & Evaluation:**
  - The dataset is split into training (80%) and testing (20%) sets using train_test_split.
  - Cross-validation (5-fold) is applied to assess each model's generalization ability and minimize overfitting.
- **Ensemble Model Construction:**
  - The individual models are combined using a Stacking Classifier, with Logistic Regression as the final estimator. A weighted voting mechanism is employed, where models with better performance during cross-validation contribute more to the final decision.
- **Performance Metrics:** The model's performance is evaluated using accuracy, precision, recall, and F1-score to measure its ability to distinguish between malignant and benign tumors.
- **Training and Prediction:**
  - The stacked model is trained on the training data (X_train, y_train) and evaluated on the test data (X_test), with predictions compared using accuracy_score and a classification report.

### A. Algorithms Used

1. Logistic Regression: A statistical method for binary classification, predicting the probability of an input belonging to a given class (output between 0 and 1).
2. Support Vector Classifier : A supervised learning algorithm that finds the optimal hyperplane to separate classes. SVC is particularly effective in high-dimensional spaces and for non-linear problems.
3. Random Forest Classifier : An ensemble method that constructs multiple decision trees and combines their results for improved accuracy and reduced overfitting.
4. XGBoost Classifier: An efficient gradient boosting implementation that builds trees sequentially, with each tree attempting to correct errors made by the previous ones. It includes regularization to prevent overfitting.

### B. Experimental Setup

- Dataset: The Wisconsin Breast Cancer Dataset from sklearn.datasets, with 569 samples and 30 features (e.g., radius, texture, perimeter) for classification (Malignant: 1, Benign: 0).
- Libraries: The implementation uses NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, XGBoost, and Scikit-optimize for model training, evaluation, and hyperparameter tuning.
- Model Selection: Base models include LR, SVC, RFC, and XGBoost with specific configurations (e.g., 10,000 iterations for LR, probability estimates for SVC).
- Stacking Classifier: A stacking model combining base models and evaluated with 5-fold cross-validation, using Logistic Regression as the final estimator.

- Model Evaluation: Accuracy and performance metrics like precision, recall, and F1-score are computed on the test set to assess the effectiveness of the final ensemble model.

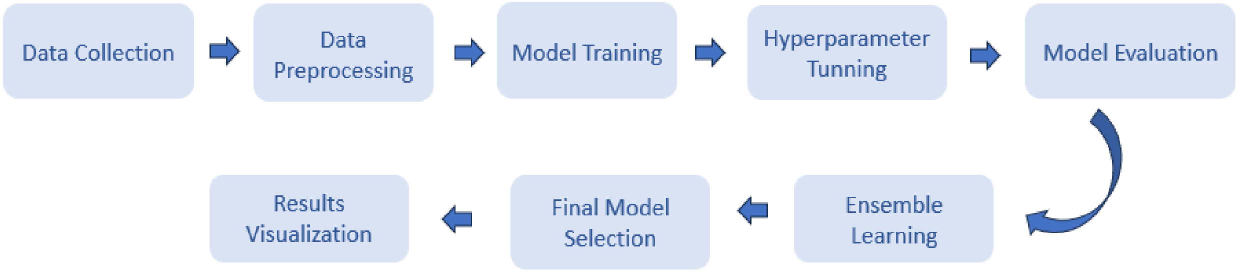


Fig. 1. Flowchart of Implementation

## IV. RESULT

The ensemble model, which combined several machine learning techniques, achieved an impressive overall accuracy of 98.25% when tested on the Wisconsin Breast Cancer Dataset. This high level of accuracy indicates that the model can effectively distinguish between benign and malignant tumors, which is essential for early cancer detection.

Table 1 : Result

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Logistic Regression | 97.37% | 97.37% | 97.37% | 97.36% |
| SVC | 98.25% | 98.29% | 98.25% | 98.24% |
| Random Forest | 96.49% | 96.52% | 96.49% | 96.47% |
| XGBoost | 95.61% | 95.61% | 95.61% | 95.60% |
| Ensemble | 98.25% | 98.29% | 98.25% | 98.24% |

In Table 1. Each individual model was also evaluated for its performance. The results are as follows:

- Logistic Regression achieved an accuracy of 97.37%, with good precision and recall scores of 0.97 each. This means it correctly identified a high percentage of both benign and malignant cases.
- Support Vector Classifier (SVC) was the standout model, reaching the highest accuracy of 98.25%. Its precision and recall were both 0.98, demonstrating its effectiveness in accurately classifying tumors.
- Random Forest achieved an accuracy of 96.49%. It had a precision of 0.97 and a recall of 0.96, indicating it performed well but was slightly less effective than the top models.
- XGBoost showed the lowest accuracy among the four models at 95.61%. Its precision and recall were both 0.96, which, while still solid, were not as high as the other models.

Among these, the Support Vector Classifier proved to be the best model due to its ability to handle complex data and find the best decision boundaries. This capability contributed to its superior accuracy, making it particularly effective in differentiating between the two types of tumors.

Overall, the ensemble model's high accuracy highlights its potential as a valuable tool for breast cancer detection, providing a reliable method that could significantly improve early diagnosis and treatment strategies in clinical practice.

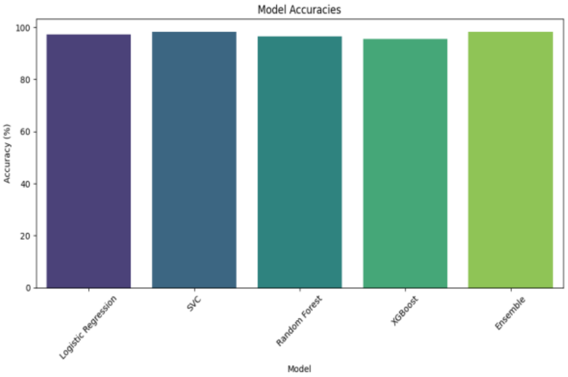


Fig. 2. Model Accuracy Comparison

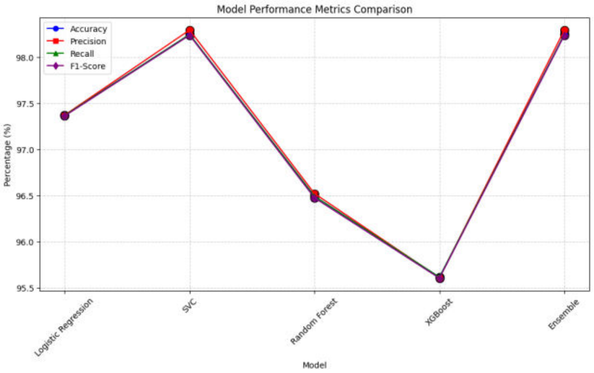


Fig. 3. Model Performance Metrics Comparison

- **Comparison with Previous Work:**

Table 2 : Performance Metrics of Individual Machine Learning Algorithms for Breast Cancer Detection

| Algorithm | Accuracy | Precision | Recall | F1-score | MAE | MSE | RMSE |
|---|---|---|---|---|---|---|---|
| DT | 78 | 77.39 | 79.47 | 77.42 | 22 | 22 | 46.9 |
| RF | 94.5 | 94.45 | 93.63 | 94.01 | 5.5 | 5.5 | 23.45 |
| ET | 94.5 | 94.18 | 93.92 | 94.05 | 5.5 | 5.5 | 23.45 |
| AdB | 90.5 | 90.6 | 88.73 | 89.53 | 9.5 | 9.5 | 30.82 |
| HGB | 94.5 | 94.18 | 93.92 | 94.05 | 5.5 | 5.5 | 23.45 |
| GBC | 94 | 93.77 | 93.24 | 93.49 | 6 | 6 | 24.49 |
| XGB | 94.5 | 93.96 | 94.21 | 94.08 | 5.5 | 5.5 | 23.45 |
| LGB | 95 | 94.86 | 94.32 | 94.57 | 5 | 5 | 22.36 |

In Table 2. comparison to existing models in the literature, our ensemble model's performance outshines several individual classifiers from previous studies. For example, in the work [5], models like LightGBM achieved an accuracy of 95%, while Random Forest and XGBoost both performed at 94.5%. In contrast, our ensemble model demonstrated superior accuracy and consistency across precision, recall, and F1-score, making it more reliable for breast cancer detection.

Overall, the ensemble model's high accuracy highlights its potential as a valuable tool for breast cancer detection. The model's performance not only surpasses individual classifiers but also demonstrates its capacity to combine the strengths of multiple algorithms. This makes it a robust solution for improving early diagnosis and treatment strategies in clinical practice.

## V. CONCLUSION

The ensemble model for breast cancer detection, utilizing the Wisconsin Breast Cancer Dataset, demonstrated exceptional effectiveness, achieving an impressive accuracy of **98.25%**. This high accuracy underscores the model's ability to accurately classify tumors as benign or malignant, which is vital for enhancing early diagnosis and improving treatment outcomes in clinical practice.

Among the individual models, **Support Vector Classifier (SVC)** outperformed others, showing the highest accuracy, highlighting its ability to handle complex datasets effectively. The integration of multiple machine learning techniques within the ensemble model further improved performance, showing that combining the strengths of different algorithms yields better results than relying on a single model.

This research emphasizes the significant role of advanced machine learning methods, particularly ensemble learning, in the early detection of breast cancer. It adds to the growing body of evidence supporting the use of machine learning in healthcare, paving the way for more accurate and efficient diagnostic tools that can improve patient care. Future studies could explore additional datasets, feature engineering, and the integration of other machine learning techniques to further boost diagnostic precision and reliability.

## REFERENCES

1. A. Ashurov et al., "Improved Breast Cancer Classification through Combining Transfer Learning and Attention Mechanism," *Life*, vol. 13, no. 9, p. 1945, Sep. 2023.
2. M. Kumar, S. Singhal, S. Shekhar, B. Sharma, and G. Srivastava, "Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning," *Sustainability*, vol. 14, no. 21, p. 13998, Oct. 2022.
3. P. Talwar, N. Sindhwani, A. Rana, and A. Chaudhary, "Breast Cancer Detection Using Machine Learning Algorithms," in *Proceedings of the 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)*, Noida, India, 2021, pp. 1–5.
4. M. Umer et al., "Breast Cancer Detection Using Convoluted Features and Ensemble Machine Learning Algorithm," *Cancers*, vol. 14, no. 23, p. 6015, Dec. 2022.
5. S. Sharmin, T. Ahammad, M. A. Talukder, and P. Ghose, "A Hybrid Dependable Deep Feature Extraction and Ensemble-Based Machine Learning Approach for Breast Cancer Detection," *IEEE Access*, vol. 11, pp. 87694–87700, Aug. 2023.
6. R. Uppara, S. Yadav, and M. Kavitha, "Voting Classifier on Ensemble Algorithms for Breast Cancer Prediction," in *Proceedings of the International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT 2023)*, IEEE, 2023, pp. 653–655.
7. A. Bilal, A. Imran, T. I. Baig, X. Liu, E. Abouel Nasr, and H. Long, "Breast Cancer Diagnosis Using Support Vector Machine Optimized by Improved Quantum Inspired Grey Wolf Optimization," *Journal of Medical Systems*, vol. 47, no. 8, 2023.
8. H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A Support Vector Machine-Based Ensemble Algorithm for Breast Cancer Diagnosis," *European Journal of Operational Research*, vol. 267, no. 2, pp. 687–699, 2017.
9. J. Chhatwal, O. Alagoz, M. J. Lindstrom, C. E. Kahn Jr., K. A. Shaffer, and E. S. Burnside, "A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis," *American Journal of Roentgenology*, vol. 192, no. 4, pp. 1117–1127, 2009.
10. G. Alfian et al., "Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method," *Journal of Healthcare Informatics Research*, vol. 13, no. 2, pp. 154–167, 2021.
11. J. Kim, M. Lee, and J. Seok, "Deep Learning Model with L1 Penalty for Predicting Breast Cancer Metastasis Using Gene Expression Data," *Machine Learning: Science and Technology*, vol. 4, no. 2, p. 025026, 2023.
12. B. Kolla and V. P., "Breast Cancer Diagnosis through Knowledge Distillation of Swin Transformer-Based Teacher–Student Models," *Machine Learning: Science and Technology*, vol. 4, no. 4, p. 045047, 2023.
13. M. Alshehri, "Breast Cancer Detection and Classification Using Hybrid Feature Selection and DenseXtNet Approach," *Mathematics*, vol. 11, no. 23, p. 4725, 2023.
14. S. Das and D. Biswas, "Prediction of Breast Cancer Using Ensemble Learning," in *Proceedings of the 2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*, Dhaka, Bangladesh, 2019, pp. 804–808.

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

📱 9940 572 462 ☎ 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details