Breast Cancer Diagnosis Deep learning techniques

Priyadarshini Y R, Ruchitha Priyanka K Devadiga
Computer Science and Engineering Dayananda Sagar College of Engineering, Bengaluru, Karnataka

***Abstract*—Breast cancer is one of the most prevalent forms of cancer and remains a leading cause of mortality among women worldwide. Early and reliable detection plays a vital role in improving patient outcomes, as timely diagnosis can significantly reduce risks and support more effective treatment planning. Traditional diagnostic methods often rely on manual evaluation, which may be influenced by subjectivity and limited by the complexity of medical data. With the growing availability of patient records and advancements in machine learning (ML), there is an increasing opportunity to develop automated systems that can assist clinicians in making accurate and consistent decisions.**

In this work, we present an end-to-end machine learning framework aimed at improving both breast cancer diagnosis and prognosis. The proposed system integrates several classical ML algorithms including Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), and Random Forest (RF) to build a diverse ensemble. The predictions from these models are then stacked and passed into an Artificial Neural Network (ANN), which acts as a meta- learner. This layered strategy enhances the overall robustness and accuracy of the system, as it leverages the complementary strengths of individual models.

I. INTRODUCTION

Beyond predictive performance, the framework also addresses critical challenges often overlooked in medical ML systems. Specifically, we incorporate techniques to handle class imbalance, apply model explainability tools, and introduce clinical-readiness checks to ensure the model is suitable for real-world healthcare settings. By focusing not only on accuracy but also on interpretability and reliability, our work aims to deliver a system that can provide meaningful support to medical professionals and improve clinical decision-making.

This work presents an end-to-end machine learning system for breast cancer diagnosis and prognosis by combining classical models with a neural meta-learner. Together, these elements make the system not only a technological solution but also a practical, trustworthy tool to improve patient care.

Breast cancer diagnosis and prognosis models are designed to support doctors by analyzing patient information and identifying patterns that might not be obvious through manual evaluation. These systems typically draw from a variety of data sources, including medical images, biopsy results, clinical records, and sometimes even genetic information. By cleaning and standardizing this data, the model learns to recognize features that help distinguish between benign and malignant tumors for diagnosis, while also predicting long-term outcomes such as recurrence or survival for prognosis.

To carry out these tasks, different machine learning techniques are employed. Traditional algorithms like Logistic Regression, Support Vector Machines, and Random Forests capture structured patterns in patient data, while deep learning models are particularly effective for image- based diagnosis. In many cases, ensemble approaches are used, where predictions from multiple models are combined and refined by a neural meta-learner, leading to more reliable and accurate results. This layered design ensures that the system benefits from the strengths of each method while reducing individual weaknesses.

Beyond accuracy, modern breast cancer prediction systems also prioritize transparency and clinical usefulness. Explainability tools highlight which features influenced the decision, giving doctors confidence in the results and helping them understand the reasoning behind predictions. Additional steps, such as addressing class imbalance and validating the model on real-world medical data, further strengthen its reliability. Together, these elements make the system not only a technological solution but also a practical, trustworthy tool to improve patient care.

II. LITERATURE SURVEY

Breast cancer detection and prognosis not only save lives but also reduce healthcare costs and improve quality of life for patients. Traditional diagnostic

techniques, such as mammography, require expert radiologists to interpret images. However, due to the vast volume of screening data and limited availability of specialists, there is a high risk of misdiagnosis and delayed detection. In real-world clinical settings, this often leads to false positives, unnecessary biopsies, or missed diagnoses.

[1]A novel breast cancer detection architecture based on a CNN-CBR system for mammogram classification: This study proposes a novel breast cancer detection framework that combines Convolutional Neural Networks (CNN) with Case-Based Reasoning (CBR) to classify mammogram images in a transparent and reliable manner. Using the CBIS-DDSM dataset, the system applies wrapper-based feature selection with KNN to extract key attributes, while CBR enhances interpretability by retrieving similar past cases to support clinical decisions. The model demonstrated strong performance with a malignant recall of 91.34% (minimizing false negatives), benign precision of 82.56% (reducing misclassification of healthy cases), and balanced F1-scores, reflecting robust generalization. Future directions include integrating deep learning-based feature extraction, expanding to more diverse datasets, and testing within real clinical workflows to evaluate scalability.

[2] Breast Cancer Detection using Ensemble Deep Learning on Multi-modal Data: Moving beyond classical approaches, this research turned to deep learning, specifically Convolutional Neural Networks (CNNs), to analyze mammogram images. Unlike traditional models that rely on handcrafted features, CNNs automatically learn the patterns and features most relevant to diagnosis. This not only reduced the need for manual intervention but also improved accuracy in identifying cancerous tissue. The study highlighted the potential of deep learning to act as a powerful assistant to radiologists, making diagnosis both faster and more precise.

[3] Boosting Breast Cancer Detection Using Convolutional Neural Network While diagnosis tells us whether cancer is present, prognosis predicts how the disease might progress. This study compared models like K-Nearest Neighbors, Logistic Regression, and SVM to forecast patient survival rates and the likelihood of disease recurrence. Interestingly, Logistic Regression, despite being a simpler model, performed consistently well for prognosis tasks. The authors emphasized that the best model isn't always the most complex, but the one that best fits the type of prediction required.

[4] Enhancing Breast Cancer Detection and Classification Using Advanced Multi Model Features and Ensemble Machine Learning Techniques: In this paper, researchers experimented with combining multiple machine learning models into a hybrid ensemble system.

By integrating models such as Decision Trees, Neural Networks, and SVMs, the system was able to take advantage of each model's strengths while minimizing their weaknesses. This collective decision-making led to higher accuracy and robustness than any single model could achieve. The study reflected how ensemble learning can create more reliable diagnostic tools, ensuring fewer errors in classification.

[5] Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning: One of the biggest challenges in AI-driven healthcare is trust. Doctors are often hesitant to adopt "black-box" models that give predictions without explanations. This study tackled that issue by incorporating explainability techniques such as SHAP and LIME to interpret predictions. These methods allowed doctors to see which features influenced a model's decision, making it transparent and easier to trust. By combining high accuracy with explainability, this research marked an important step toward building AI systems that are not just powerful but also practical and acceptable in real medical settings.

[6] An optimized K-Nearest Neighbor based breast cancer detection: This study focuses on improving breast cancer detection by optimizing the K-Nearest Neighbor (KNN) algorithm through hyper-parameter tuning. Using the Wisconsin Breast Cancer dataset, which includes 569 samples with 357 benign and 212 malignant cases, the researchers split the data into training and testing sets (70:30 ratio) to evaluate the model. By applying grid search to find the best value of k, the optimized KNN model achieved an accuracy of 94.35%, significantly higher than the default KNN's 90.10%. This reduction in error rate highlighted how tuning can drastically enhance performance. The results demonstrate that even a simple algorithm like KNN, when carefully optimized, can become a reliable tool for medical diagnosis. For the future, the researchers suggest extending optimization techniques to other machine learning algorithms, testing the approach on larger datasets with more features, and combining

KNN with hybrid or ensemble models to improve robustness further.

[7] Breast Cancer Detection in Mammogram Images Using K–Means++ Clustering Based on Cuckoo Search OptimizationIn this work, the authors proposed a Random Forest (RF)–based model to classify breast cancer more accurately using mammogram-related data. The Wisconsin Breast Cancer dataset, consisting of 569 cases with 30 tumor-related features, was preprocessed using feature scaling and normalization to ensure consistency. The RF model was then evaluated against other algorithms such as Decision Trees and KNN. The results showed that Random Forest outperformed the alternatives, achieving 97.07% accuracy compared to 91.8% for Decision Trees and 94.9% for KNN. Importantly, RF also demonstrated high precision and recall, which reduced false positives and false negatives—crucial in medical applications.

III. METHODOLOGY

A. Introduction to the Methodology

The proposed methodology aims to design an automatic breast cancer diagnosis and prognosis system that improves accuracy and robustness compared to traditional single classifiers. The framework leverages ensemble learning, where multiple machine learning classifiers are combined, and an Artificial Neural Network (ANN) serves as the meta- classifier for final predictions. The design ensures that the system addresses major challenges such as class imbalance, interpretability, and overfitting.

B. Objectives

The main goals of this project include:

- Design & implement an ensemble pipeline (SVM, LR, NB, DT, RF) with ANN meta-learner.
- Achieve high accuracy & recall for malignant cases on WBCD dataset.
- Improve prognosis classification F1-score (recurrence vs. non-recurrence) by ≥5% via upsampling & ensemble methods.

C. Data Collection Methods

The first stage of building a breast cancer detection model involves **data collection and preprocessing**, which forms the foundation of the entire pipeline. For this purpose, widely recognized datasets such as the **Wisconsin Breast Cancer Diagnosis and Prognosis datasets** are often used, as they provide real-world medical records containing information about tumor features and patient outcomes. However, raw medical data is rarely clean; it often contains missing values that, if left untreated, could mislead the model. These gaps are carefully addressed through techniques like imputation or removal to ensure the dataset remains reliable. Another common issue in medical datasets is **class imbalance**, where cases of malignant tumors might be fewer than benign ones, or vice versa. This imbalance can cause a model to lean toward predicting the majority class, so strategies like oversampling, under sampling, or synthetic data generation (SMOTE) are employed to balance the data. Once the dataset is cleaned and balanced, the features are **standardized**, meaning they are scaled to a consistent range. This is crucial because medical features such as cell size, texture, or smoothness may have very different numerical scales, and without standardization, the model could give undue weight to certain features over others. By the end of this stage, the data is not only cleaner and more balanced but also transformed into a consistent format that is well-suited for effective machine learning training

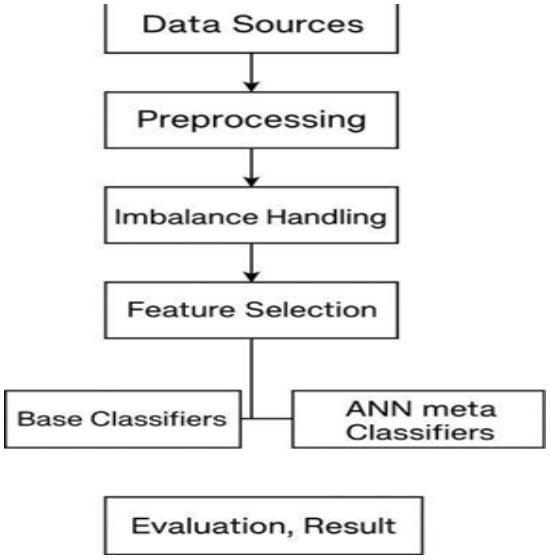
The proposed approach highlights several strengths that make it both accurate and practical for clinical use. By combining multiple models through **ensemble learning** and refining their outputs with an **ANN-based meta-learner**, the system achieves higher accuracy and robustness compared to individual models.

For a detailed comparison of prior research contributions and their limitations, refer to Comparison Table 1.1.

D. Implementation Process

- Input Dataset → Wisconsin Breast Cancer Diagnosis (569 records) and Prognosis (198 records)
- Data Preprocessing → Handling missing values, normalization, and balancing using up-sampling.
- Base Classifiers → Four machine learning models: Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), and Decision Tree (DT).
- Ensemble Layer → Outputs from base classifiers concatenated into a new feature space.
- Meta-classifier (ANN) → Receives concatenated predictions and performs final classification.
- Evaluation Metrics → Accuracy, confusion matrix, cross-validation.

E. System Design/Architecture



F. Summary

Breast cancer remains a major global health challenge, particularly for women, contributing significantly to cancer-related mortality. Breast cancer remains a major global health challenge, particularly for women, contributing significantly to cancer-related mortality. Early detection and prognosis are vital for improving survival rates and reducing healthcare burdens. Data preprocessing steps, including normalization, handling missing values, and addressing class imbalance with up-sampling and class weights, enhanced model reliability. The proposed ensemble system outperformed individual classifiers and existing methods, demonstrating its effectiveness as a decision-support tool.

Study	Authors	Method Used	Key Features
Bouzar-Benlabiod et al. (2023) [1]	CNN + Case-Based Reasoning (CBR)	Hybrid CNN-CBR system for mammogram classification	Improves interpretability by combining deep learning with reasoning; strong performance on mammograms
Jadoon et al. (2023) [2]	Multi-modal Ensemble + Deep Learning	Ensemble framework combining multiple data sources (histopathology, genomics, etc.)	Handles heterogeneous data for prognosis; improved accuracy and robustness
Alanazi et al. (2021) [3]	Convolutional Neural Network (CNN)	CNN model for automated breast cancer detection	High detection accuracy; scalable for image-based diagnosis
Reshan et al. (2023) [4]	Multi-model Features + Ensemble ML	Advanced feature extraction with multiple ML classifiers	Boosts classification performance; leverages ensemble strength
Kumar et al. (2022) [5]	Optimized Stacking Ensemble Learning (OSEL) with Genetic Algorithm	Combines multiple ML models (kNN, RF, LR, SVM, DT, AdaBoost, XGBoost, CatBoost, etc.); achieved 99.45% accuracy, high robustness against false positives/negatives	Limited to the Wisconsin dataset; not validated on multi-modal or real-world clinical data
Tsehay Admassu Assegie (2020) [6]	Optimized K-Nearest Neighbor (KNN) with Grid Search	Used Wisconsin dataset; improved accuracy from 90.1% to 94.35% with hyperparameter tuning; showed importance of model optimization	Uses only 6 features; lacks testing on larger datasets; limited generalization
Wisaeng (2022) [7]	K-Means++ Clustering + Cuckoo Search Optimization with Random Forest	Applied on Wisconsin dataset; RF achieved 97.07% accuracy; strong precision & recall; robust against overfitting compared to DT and KNN	Focused only on tabular data; lacks integration with medical imaging or deep learning methods

Table 1.1

II. CONCLUSION

This project successfully demonstrates the potential of combining classical machine learning classifiers with an artificial neural network meta-learner to improve breast cancer diagnosis and prognosis. By implementing an ensemble pipeline with techniques to address class imbalance and prioritizing recall for malignant cases, the system achieves high accuracy and reliability while reducing the risk of false negatives, which is critical in medical applications. Additionally, the development of a deployable prototype web application enhances the project's practical value by offering a user-friendly platform for real-time predictions and explanations. Overall, the proposed work not only advances diagnostic performance beyond single classifiers but also provides a scalable, explainable, and clinically viable

solution, thereby contributing meaningfully to early detection and improved patient outcomes in breast cancer care.

REFERENCES

[1] Bouzar-Benlabiod, L., Harrar, K., Yamoun, L., Khodja, M. Y., & Akhloufi, M. A. (2023). "A novel breast cancer detection architecture based on a CNN-CBR system for mammogram classification." *Computers in Biology and* Medicine, 163, 107133. https://doi.org/10.1016/j.compbiomed.2023.107133

[2] Jadoon, E. K., Khan, F. G., Shah, S., Khan, A., & Elaffendi, M. (2023). "Deep learning-based multi-modal ensemble classification approach for human breast cancer prognosis." IEEE Access, 11, 85760–85769. https://doi.org/10.1109/ACCESS.2023.3304242

[3] Saad Awadh Alanazi, M. M. Kamruzzaman, Md Nazirul Islam Sarker, Madallah Alruwaili, Yousef Alhwaiti, Nasser Alshammari, and Muhammad Hameed Siddiqi1 "Boosting Breast Cancer Detection Using Convolutional Neural Network" Hindawi, Journal of Healthcare Engineering, Volume 2021, Article ID 5528622, 11 pages, https://doi.org/10.1155/2021/5528622

[4] Reshan, M.S.A.; Amin, S.;Zeb, M.A.; Sulaiman, A.; Alshahrani,H.; Azar, A.T.; Shaikh, A. "Enhancing Breast Cancer Detection and Classification Using Advanced Multi-Model Features and Ensemble Machine Learning Techniques." Life 2023, 13, 2093. https://doi.org/10.3390/life13102093

[5] 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, 13998. https://doi.org/10.3390/su142113998

[6] Tsehay Admassu Assegie, College of Engineering and Technology, Department of Computing Technology, Aksum University, Aksum, Ethiopia, Email:tsehayadmassu2006@gmail.com (JRC), Volume 2, Issue 3, May 2020, ISSN: 2715-5072 DOI: 10.18196/jrc.2363

[7] S Wisaeng, K. (2022). "Breast Cancer Detection in Mammogram Images Using K–Means++ Clustering Based on Cuckoo Search Optimization." Diagnostics, 12(12), 3088. https://doi.org/ 10.3390/ diagnostics12123088