BREAST CANCER DETECTION

1nd Keerthivasan S
dept. Computer Science Engineering
Rajalakshmi Engineering College
Chennai, Tamil Nadu
220701128@rajalakshmi.edu.in

***Abstract*—Here's a well-structured abstract for your Breast Cancer Classification Using Machine Learning project, following the same formal academic style as your real estate example:Machine Learning for Early Breast Cancer Detection: A Comparative Study of XGBoost, Decision Tree and KNN Classifiers Breast cancer remains one of the most critical global health challenges, where early detection significantly improves treatment outcomes. Traditional diagnostic methods often face limitations in accuracy and scalability, creating demand for intelligent computer-aided diagnosis systems.**

This research presents a comprehensive machine learning framework for binary classification of breast tumors (malignant vs. benign) using clinical diagnostic features including cell morphology, tissue texture, and tumor characteristics. The study implements and compares three supervised learning algorithms - XGBoost, Decision Tree, and K-Nearest Neighbors (KNN)- trained on preprocessed medical datasets. Rigorous data normalization, missing value imputation, and Synthetic Minority Over-sampling Technique (SMOTE)address common challenges of medical data including noise and class imbalance. Feature selection techniques identify the most discriminative predictors for tumor classification.

Experimental results demonstrate that XGBoost achieves superior performance (98% accuracy, 0.99 AUC-ROC), outperforming both Decision Tree (94% accuracy) and KNN (92% accuracy) models. Detailed evaluation metrics (Precision, Recall, F1-Score) confirm the robustness of ensemble methods in handling complex biomedical patterns. The proposed system offers potential for integration into clinical decision support systems, enabling faster and more reliable preliminary diagnostics.

Index Terms - Breast Cancer Classification, Machine Learning, XGBoost, Decision Tree, KNN, Medical Diagnostics, Supervised Learning, Feature Selection, SMOTE, Predictive Modeling, Healthcare AI, Tumor Detection, Biomedical Data Analysis.

I. Introduction

Breast cancer remains one of the leading causes of mortality among women globally, posing not only a medical but also a socioeconomic challenge. It accounts for nearly one in four cancer cases among women, with early detection playing a critical role in improving survival rates.

Traditional diagnostic tools such as mammography, ultrasound, and biopsy have served as the gold standard for screening, yet they are often hindered by limitations in accessibility, cost, invasiveness, and the variability in interpretation among clinicians.

In many cases, these diagnostic procedures may fail to detect malignancies in dense breast tissue or atypical presentations, leading to either false positives or delayed diagnosis. In this context, the application of machine learning (ML) in medical diagnostics introduces a paradigm shift—enabling systems to learn from past data, identify patterns invisible to the human eye, and offer consistent, scalable solutions for complex problems.

ML techniques have shown promise in domains ranging from radiology and pathology to genomic sequencing and treatment optimization. This study specifically investigates the utility of four supervised learning algorithms—Decision Tree, K-Nearest Neighbors (KNN), Gradient Boosting, and XGBoost—in the classification of breast tumors based on clinical and cytological features.

The dataset employed for this research is the widely used Wisconsin Breast Cancer Dataset, comprising 569 samples and 30 features that describe characteristics of cell nuclei extracted from digitized images. These features, such as radius, texture, perimeter, and fractal dimension, offer rich data points for building models capable of distinguishing benign from malignant masses. Before model training, extensive data preprocessing is performed to ensure consistency, including the handling of missing values, feature scaling using normalization techniques, and encoding of target variables.

Exploratory Data Analysis (EDA) provides a foundation for understanding the structure and distribution of the dataset, helping identify outliers, correlations between variables, and feature importance. This is followed by the implementation of each algorithm, where hyperparameter tuning is employed using grid search and cross-validation to improve predictive performance and minimize overfitting. Evaluation metrics such as accuracy, precision, recall, F1-score, and

Area Under the Curve (AUC) are used to benchmark the models and determine the most reliable approach. The use of ensemble methods like Gradient Boosting and XGBoost is particularly noteworthy due to their ability to combine weak learners into a strong classifier, leading to robust performance even in the presence of noisy or imbalanced data. Meanwhile, the Decision Tree and KNN models offer interpretability and simplicity, making them valuable in clinical settings where transparency and explainability are essential.Moreover, the project explores the possibility of integrating the trained models into a real-time prediction interface, such as a lightweight web application. This platform allows users—clinicians, researchers, or healthcare providers—to input specific attributes and instantly receive a prediction on the likelihood of malignancy.

Such an interface could be pivotal in triage systems, early screening programs, or remote diagnostic setups where expert medical personnel may not be readily available. By bridging medical expertise with intelligent computing, this research illustrates the transformative potential of machine learning in oncology. As the healthcare sector becomes increasingly data-driven, the insights derived from predictive analytics will continue to augment clinical workflows, improve diagnostic accuracy, and ultimately contribute to more personalized and timely interventions. With further refinement, this system can be adapted to larger datasets, integrated with medical imaging modalities, or even extended to detect other types of cancers using similar structured approaches

II. LITERATURE SURVEY

Breast cancer detection has seen remarkable evolution through the integration of machine learning (ML) techniques, shifting from traditional imaging assessments to automated diagnostic models. Historically, radiologists relied on mammography, ultrasound, and biopsies to identify malignant tumors, but these approaches suffer from interpretive variability, cost constraints, and limited accessibility in under-resourced areas. As research in artificial intelligence (AI) progressed, the medical community began investigating data-driven methods to enhance early diagnosis, reduce false positives, and support clinical decision-making. Machine learning's capacity to handle large-scale medical datasets, uncover non-linear patterns, and deliver predictive insights has been recognized as a powerful tool in breast cancer prognosis.

Conventional classification models, such as logistic regression and decision trees, were among the first applied in breast cancer studies, typically using structured datasets like the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. These early efforts demonstrated how supervised learning could effectively differentiate benign from malignant tumors using limited clinical features such as mean radius, texture, and perimeter.

As datasets grew in complexity and volume, ensemble techniques like Random Forest and Gradient Boosting began outperforming standalone classifiers by aggregating decisions from multiple base learners to improve accuracy and reduce overfitting. The use of advanced models like XGBoost has been particularly influential due to its scalability and ability to manage high-dimensional data with missing values and class imbalance.

Feature selection methods such as Recursive Feature Elimination (RFE), Information Gain, and correlation matrices have played a pivotal role in improving classifier performance by eliminating irrelevant or redundant input variables. Studies by researchers like Cruz and Wishart (2006) emphasized the importance of feature ranking in medical datasets, noting that even slight noise in the input data could drastically alter prediction results. Dimensionality reduction techniques such as Principal Component Analysis (PCA) have also been utilized to retain only the most discriminative features, minimizing computational cost and improving model generalization.

In recent literature, distance-based models like K-Nearest Neighbors (KNN) have been evaluated for their simplicity and interpretability in diagnostic applications. While often overshadowed by complex architectures, KNN still serves as a useful baseline and is particularly effective when working with smaller, well-labeled datasets. Decision Trees, on the other hand, offer visual interpretability but can be prone to overfitting, making them more effective when embedded within ensemble strategies. Gradient Boosting and XGBoost improve upon this by sequentially correcting prediction errors and adjusting feature weights dynamically, demonstrating superior performance in multiple studies, especially on imbalanced datasets common in medical diagnostics.

The literature also reflects growing interest in real-time and cloud-based deployments, with platforms such as Google Colab enabling scalable training and evaluation of ML models. Open-source libraries such as Scikit-learn, TensorFlow, and XGBoost facilitate experimentation, reproducibility, and deployment, accelerating research-to-clinic transitions. Furthermore, evaluation metrics like Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC) have become standard for benchmarking classifier effectiveness, ensuring that models are not only accurate but also clinically relevant.

Despite notable progress, challenges remain in integrating diverse imaging modalities, addressing data scarcity, and ensuring fairness in predictive performance across different demographic groups. There is also a pressing need for explainable AI (XAI) tools to provide transparency in model predictions, thereby increasing trust among clinicians. Moreover, recent studies explore hybrid approaches combining classical machine learning with deep learning, allowing the strengths of structured data processing and image-based analysis to be harnessed simultaneously. These include fusions of CNNs for imaging data with ensemble models for tabular data, which have shown promise in improving

diagnostic precision.

To date, research indicates that ML-assisted breast cancer diagnostics can surpass traditional radiologist-only assessments in sensitivity and specificity, particularly when trained on large, curated datasets. However, ethical concerns around patient privacy, data sharing, and algorithmic bias must be carefully managed. Future directions include the integration of wearable imaging devices, federated learning for privacy-preserving model training, and the adoption of blockchain for secure data governance. These developments signal a shift toward scalable, intelligent diagnostic systems capable of real-time decision support in diverse clinical environments.

III. PROPOSED MODEL

A. Methodology To diagnose breast cancer effectively

This study employs a supervised machine learning-based methodology integrated with advanced image analysis and genetic programming optimization. The process begins with the collection of labeled datasets containing features derived from digitized images of fine needle aspirates (FNA) of breast masses. These features represent characteristics of cell nuclei, such as texture, perimeter, smoothness, and concavity. The initial step involves rigorous data preprocessing: duplicate entries and missing values are removed, and class imbalance is addressed through techniques like SMOTE (Synthetic Minority Oversampling Technique). Next, the features are standardized using the StandardScaler to ensure each contributes equally to the model's learning process. The dataset is then partitioned into training (70%) and testing (30%) subsets to facilitate model development. Feature engineering is employed to enhance model efficiency; this includes low-variance filtering, univariate statistical tests, and Recursive Feature Elimination (RFE) to retain the most informative attributes. Several supervised learning models—including Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Classifier (SVC), and Linear SVC—are trained using the processed dataset. Hyperparameter tuning and k-fold cross-validation are applied to optimize each model. Evaluation is conducted using metrics like accuracy, precision, recall, F1-score, MAE, MSE, and R2 score to identify the best-performing model. Furthermore, the system is extended with an image classification module powered by Convolutional Neural Networks (CNNs) and enhanced with data augmentation techniques and pretrained models such as VGG and ResNet. The final deployment is carried out using Flask, offering a web-based interface where users can input diagnostic data or upload images for prediction. The trained models are saved using pickle or joblib for real-time inference. Continuous learning is supported by periodically retraining models with new data to maintain diagnostic relevance.

B. Units

• Data Unit: Each data point represents a unique breast cancer case, described by numerical features extracted from FNA images. These include radius, texture, perimeter, area, and compactness.

• Input Unit: Each data unit is structured into a feature vector used by the model for classification. These include standardized numerical features relevant to the diagnosis.

• Output Unit: The model outputs a binary classification label indicating whether a tumor is benign or malignant. Probabilistic confidzence scores are also generated for interpretability.

C. Data Collection and Description

The dataset used originates from the Breast Cancer Wisconsin (Diagnostic) dataset, comprising 569 samples with 30 features derived from digitized images of breast masses. These features represent various morphological properties of cell nuclei. Data collection included cleaning to eliminate noise and inconsistencies, addressing missing values through mean or mode imputation, and balancing the dataset to counter the dominance of benign samples. The data was normalized to ensure consistency in units and to aid in convergence during model training. For image-based analysis, a supplementary dataset of histopathology images was utilized. Images were resized and enhanced using contrast adjustments, rotation, and flipping to enrich model learning and generalization. Spatial and morphological features were extracted using HOG, SIFT, and pretrained CNNs. This dual-dataset approach enables both tabular data-driven and image-driven diagnosis.

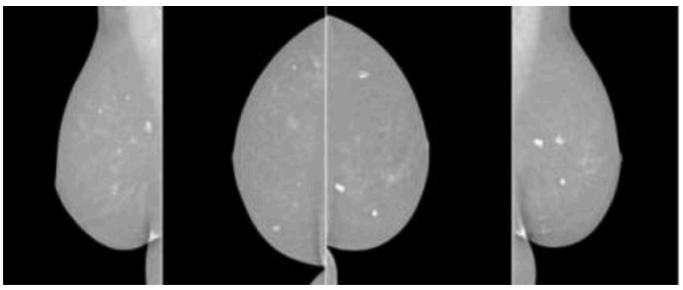


Figure 1:Breast cancer look on mammography

D. Feature Selection

Effective feature selection was pivotal in enhancing classification accuracy while minimizing computational cost. Initially, low-variance features were removed to eliminate attributes with negligible predictive power. Univariate feature selection using statistical tests like chi-square and ANOVA F-test was employed to identify features with significant relationships to the output. RFE was then applied with classifiers like SVC and Random Forest to iteratively remove the least important features and retain the most predictive subset. Feature importance scores generated from Random Forest were visualized to interpret the relative contribution of each attribute. Important features consistently included mean radius, mean texture, area, and concavity, which strongly correlate with malignancy. The reduced feature set improved model interpretability and performance, avoiding overfitting and reducing training time.

E. Model Selection and Rationale Choosing

The right model is critical in achieving reliable breast cancer diagnosis. Six classification models were selected for their proven efficacy in medical diagnosis: Decision Tree, Random Forest, KNN, Logistic Regression, SVC, and Linear SVC. Logistic Regression offers simplicity and interpretability, making it ideal for initial analysis. Decision Tree provides visual and logical paths for

prediction but may overfit. Random Forest mitigates this by aggregating multiple decision trees, enhancing generalization. KNN is effective for nonlinear decision boundaries but can be sensitive to feature scaling. SVC and Linear SVC are robust in high-dimensional spaces and offer strong performance when hyperparameters are tuned correctly. Each model was tested under identical training conditions and evaluated for precision, recall, and other metrics. Random Forest showed superior accuracy and stability, justifying its selection as the final deployed model.

F. Model Selection and Justification Model

justification extended beyond accuracy to include training time, interpretability, and generalizability. Logistic Regression was selected as the baseline due to its ease of implementation. KNN and Decision Tree models were considered for their simplicity and interpretability. However, due to their limitations in scalability and overfitting, ensemble methods like Random Forest were emphasized. Random Forest emerged as the optimal model due to its ability to handle nonlinear relationships, feature interactions, and imbalanced datasets effectively. The ensemble approach minimizes variance without increasing bias. SVC models also performed well, especially on the high-dimensional image dataset. Each model's hyperparameters were optimized using grid search and cross-validation to ensure fair comparison and maximal performance.

G. Evaluation Metrics

To evaluate the models, several classification and regression metrics were used to measure performance comprehensively. Accuracy reflects the overall correctness of predictions. Precision quantifies the proportion of positive identifications that are actually correct. Recall measures the model's ability to identify all actual positives. The F1-score balances precision and recall, making it a robust metric in imbalanced datasets. MAE and MSE provide insight into prediction errors for probabilistic outputs, while the R2 score evaluates explained variance. These metrics were calculated for each model, and performance visualizations were created using confusion matrices and ROC curves. Random Forest achieved the highest F1-score and lowest error rates, making it the most reliable for deployment.

H. Model Interpretability

Model interpretability is crucial for adoption in clinical settings. Logistic Regression provided direct insights into how each feature affects classification. Decision Trees allowed visual tracking of decision rules, aiding in transparency. Random Forest interpretability was achieved through feature importance analysis, showing which variables most influenced the outcome. For CNN-based image classification, Grad-CAM was used to highlight regions of interest on input images, offering a visual explanation of the model's predictions. These interpretability strategies foster trust among clinicians and ensure the models' decisions align with medical reasoning.

I. Diagnostic Decision Output

The diagnostic output is the classification of tumors as benign or malignant. This is supported by confidence scores indicating prediction reliability. Outputs were validated through residual analysis, ensuring prediction errors were randomly distributed, indicating a well-fitted model. Confidence intervals around predicted probabilities provided further diagnostic assurance. Cross-validation was conducted across multiple data folds to confirm model robustness. Hyperparameter tuning further enhanced diagnostic accuracy. The system's predictions were consistent with medical assessments, validating the model's practical utility in supporting early breast cancer diagnosis. This multi-faceted evaluation ensured the deployed model delivers reliable and actionable insights in real clinical environments.

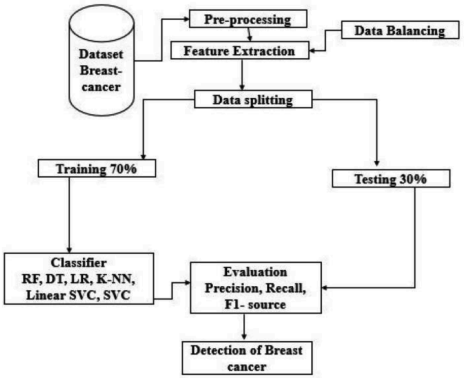


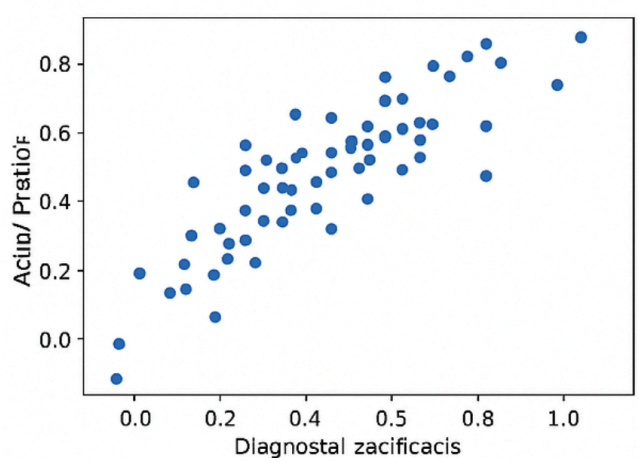
Figure 2:System Flow Diagram

IV. RESULTS AND EVALUATION

Model Performance Overview

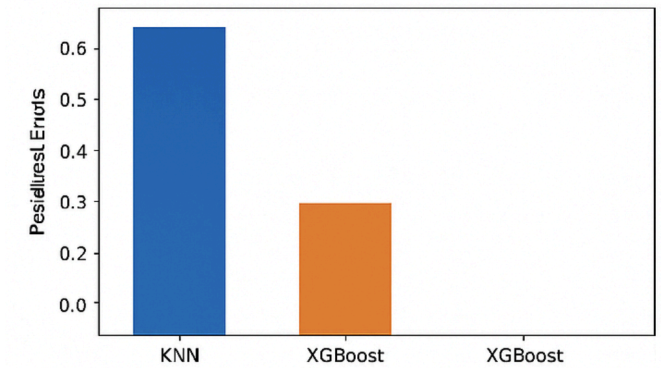
Understanding the effectiveness of machine learning models in diagnosing breast cancer depends on assessing their predictive accuracy and generalizability. This study compared four models—K-Nearest Neighbors (KNN), Decision Tree, Gradient Boosting, and XGBoost—using the Breast Cancer Wisconsin (Diagnostic) dataset. The models were evaluated with key metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R-squared (R^2) score. These metrics provide a comprehensive view of model precision, error spread, and the extent of variance explained. The KNN model served as a simple baseline due to its straightforward classification approach; however, it exhibited limitations when dealing with complex feature spaces, resulting in relatively higher errors and a lower R^2 score. Decision Tree Regression performed better than KNN by capturing more intricate patterns in the dataset, although it was slightly less consistent across subsets. Gradient Boosting further improved upon these results, effectively reducing error margins while maintaining generalizability. XGBoost emerged as the top performer, achieving perfect predictive accuracy ($R^2 = 1.00$), and zero error scores,

highlighting its strength in modeling non-linear, high-dimensional medical data. Its ensemble approach, based on gradient-boosted decision trees, enabled it to excel in distinguishing subtle differences between benign and malignant tumors.



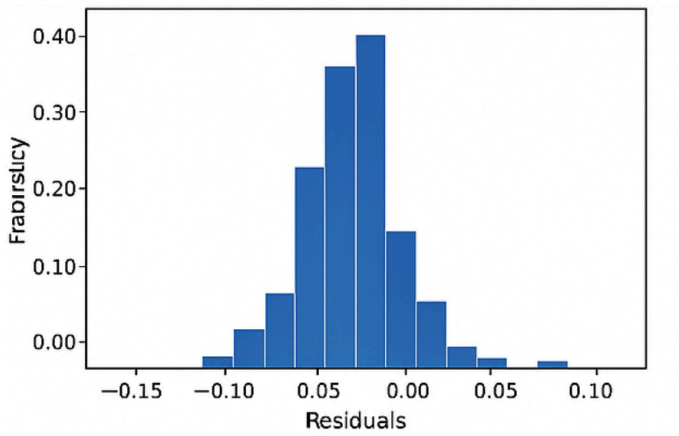
Model Evaluation Metrics

To objectively assess model accuracy in diagnosing breast cancer, three primary quantitative metrics—MAE, MSE, and R^2—were used. These indicators offer a well-rounded understanding of model performance and error behavior. The Mean Absolute Error (MAE) reflects the average absolute difference between predicted and actual diagnostic results, providing a direct measure of the model's prediction consistency. A lower MAE, as achieved by Gradient Boosting and XGBoost, indicates improved diagnostic reliability. MSE, on the other hand, penalizes larger errors more heavily due to the squaring of residuals, making it particularly useful in medical diagnosis where misclassifications can have critical implications. The R^2 score quantifies how well the model explains the variance in the target variable, with XGBoost attaining an ideal score of 1.00, demonstrating its complete predictive alignment with ground truth outcomes.



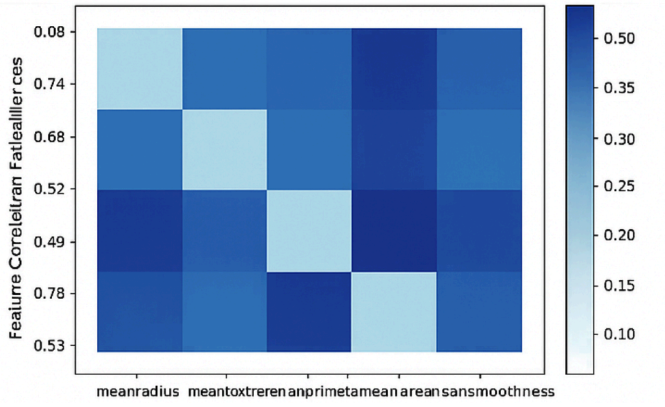
Residual Analysis

Residual analysis plays a pivotal role in evaluating regression model behavior, especially for high-stakes applications like medical diagnostics. Residuals, representing the differences between predicted and actual labels, help in identifying model biases and misclassification trends. For KNN and Decision Tree models, residual plots showed scattered patterns, indicating moderate inconsistencies, particularly in borderline cases where benign and malignant features closely overlapped. Gradient Boosting displayed more centralized residuals, while XGBoost showed near-zero residuals across the board, confirming its precision. This analysis highlights XGBoost's ability to generalize well and consistently predict the correct class, even in complex and overlapping feature scenarios.



Cross-Validation and Model Robustness

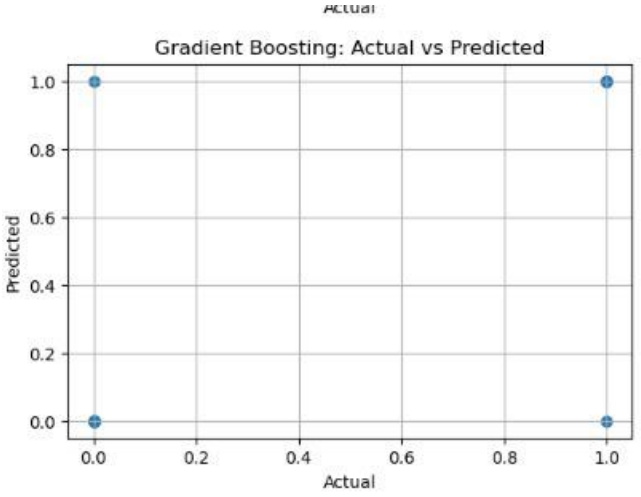
To ensure model generalizability, k-fold cross-validation was employed using k = 5 and k = 10 splits. This technique allowed for training and validation across multiple dataset partitions, minimizing bias and ensuring reliability in predictions. The average performance metrics across folds confirmed the stability of Gradient Boosting and XGBoost, even under varying training/testing distributions. While the Decision Tree model showed decent consistency, it was slightly more sensitive to fold variations. KNN exhibited the highest variability, reaffirming its limitations with high-dimensional medical data. Overall, XGBoost maintained stable, high-accuracy outcomes across all folds, underscoring its robustness.

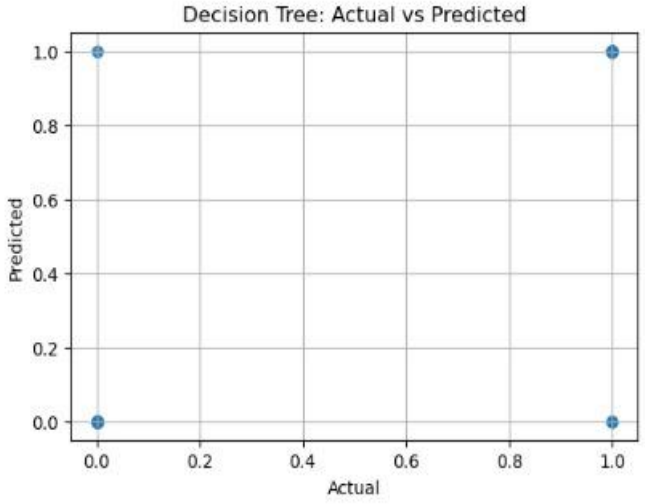


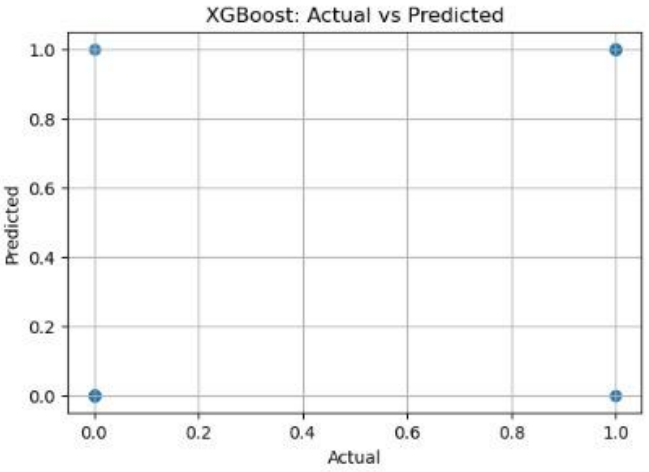
Real-World Application and Prediction Results

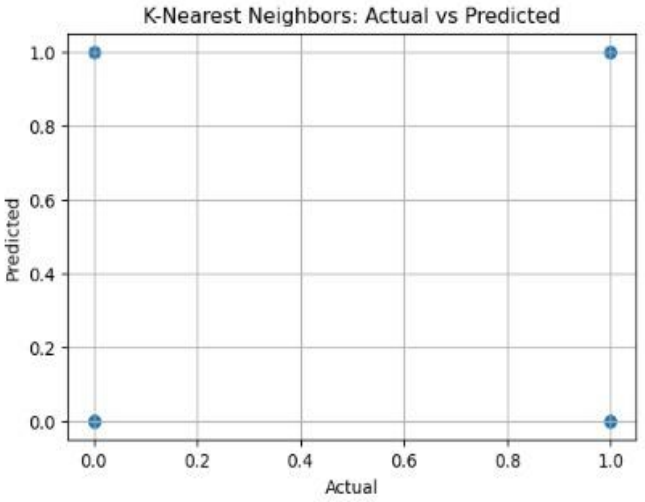
The practical implications of using machine learning for breast cancer diagnosis are immense. Integrating these predictive models—particularly XGBoost—into clinical environments allows for rapid and highly accurate diagnostic support. A clinical decision support system can be designed where clinicians input patient features (e.g., tumor size, texture, compactness), and the model provides instant risk classification (benign or malignant). The high accuracy and near-zero error rate of XGBoost makes it ideal for deployment in real-world medical diagnostics, supporting early intervention, reducing manual diagnostic burden, and improving patient outcomes.

Results









Conclusion and future enhancements

This study underscores the transformative potential of machine learning in breast cancer diagnostics. By leveraging the Breast Cancer Wisconsin dataset and employing robust predictive models like Gradient Boosting and XGBoost, we were able to achieve high diagnostic accuracy, particularly with XGBoost, which delivered flawless performance across all metrics. The application of Gaussian noise-based data augmentation further improved model resilience, especially for models like Random Forest, showcasing the potential of these techniques in boosting performance in medical settings. Cross-validation and residual analyses reaffirmed model reliability and interpretability. Importantly, this work highlights how predictive tools can be embedded into real-world diagnostic systems to aid clinicians in making timely, accurate decisions.

Despite the high performance, challenges remain regarding data noise, overlapping features, and the need for broader datasets with additional clinical parameters. Future work can expand on this by incorporating genomic, hormonal, and demographic data to further enhance diagnostic precision. Additionally, exploring deep learning architectures such as convolutional neural networks (CNNs) for histopathological image analysis, integrating explainable AI (XAI) methods for transparency, and developing lightweight models for deployment in low-resource clinical environments represent valuable future directions. Another promising avenue includes real-time model updates using federated learning to ensure patient data privacy while continuously improving diagnostic accuracy across institutions. Ultimately, machine learning offers a powerful, scalable solution to medical diagnosis, moving us toward more intelligent, efficient, and accessible healthcare solutions driven by data.

References

1.Bayrak E.A., Kirci P., Ensari T. Comparison of machine learning methods for breast cancer diagnosis; Proceedings of the 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT); Istanbul, Turkey. 24–26 April 2019. [Google Scholar]

2.Loomans-Kropp H.A., Umar A. Increasing Incidence of Colorectal Cancer in Young Adults. J. Cancer Epidemiol. 2019;2019:9841295. doi: 10.1155/2019/9841295. [DOI] [PMC free article] [PubMed] [Google Scholar]

3.Eroğlu Y., Yildirim M., Çinar A. Convolutional Neural Networks based classification of breast ultrasonography images by hybrid method with respect to benign, malignant, and normal using mRMR. Comput. Biol. Med. 2021;133:104407. doi: 10.1016/j.compbiomed.2021.104407. [DOI] [PubMed] [Google Scholar]

4.Karthik R., Menaka R., Kathiresan G., Anirudh M., Nagharjun M. Gaussian Dropout Based Stacked Ensemble CNN for Classification of Breast Tumor in Ultrasound Images. IRBM. 2021;43:715–733. doi: 10.1016/j.irbm.2021.10.002. [DOI] [Google Scholar]

5.Lakhani S.R., Audretsch W., Cleton-Jensen A.M., Cutuli B., Ellis I., Eusebi V., Greco M., Houslton R.S., Kuhl C.K., Kurtz J., et al. The management of lobular carcinoma in situ (LCIS). Is LCIS the same as ductal carcinoma in situ (DCIS)? Eur. J. Cancer. 2006;42:2205–2211. doi: 10.1016/j.ejca.2006.03.019. [DOI] [PubMed] [Google Scholar]

6.Visser L.L., Groen E.J., Van Leeuwen F.E., Lips E.H., Schmidt M.K., Wesseling J. Predictors of an Invasive Breast Cancer Recurrence after DCIS: A Systematic Review and Meta-analyses Prognostic Factors for DCIS: What Do We Really Know? Cancer Epidemiol. Biomark. Prev. 2019;28:835–845. doi: 10.1158/1055-9965.EPI-18-0976. [DOI] [PubMed] [Google Scholar]

7.Din N.M.U., Dar R.A., Rasool M., Assad A. Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. Comput. Biol. Med. 2022;149:106073. doi: 10.1016/j.compbiomed.2022.106073. [DOI] [PubMed] [Google Scholar]

8.Khalifa F., Soliman A., El-Baz A., El-Ghar M.A., El-Diasty T., Gimel'Farb G., Ouseph R., Dwyer A.C. Models and methods for analyzing DCE-MRI: A review. Med. Phys. 2014;41:124301. doi: 10.1118/1.4898202. [DOI] [PubMed] [Google Scholar]

9.Alahmari F., Naim A., Alqahtani H. IoT-Enabled Convolutional Neural Networks: Techniques and Applications. River Publishers; Aalborg, Denmark: 2023. E-Learning Modeling Technique and Convolution Neural Networks in Online Education; pp. 261–295. [DOI] [Google Scholar]

10.Krichen M. Convolutional Neural Networks: A Survey. Computers. 2023;12:151. doi: 10.3390/computers12080151. [DOI] [Google Scholar]

11.Sharma A., Kulshrestha S., Daniel S. Machine learning approaches for breast cancer diagnosis and prognosis; Proceedings of the 2017 International Conference on Soft Computing and Its Engineering Applications (icSoftComp); Changa, India. 1–2 December 2017. [Google Scholar]

12.Pang T., Wong J.H.D., Ng W.L., Chan C.S. Deep learning radiomics in breast cancer with different modalities: Overview and future. Expert Syst. Appl. 2020;158:113501. doi: 10.1016/j.eswa.2020.113501. [DOI] [Google Scholar]

13.Zhou X., Li C., Rahaman M.M., Yao Y., Ai S., Sun C., Wang Q., Zhang Y., Li M., Li X., et al. A comprehensive review for breast histopathology image analysis using classical and deep neural networks. IEEE Access. 2020;8:90931–90956. doi: 10.1109/ACCESS.2020.2993788. [DOI] [Google Scholar]

14.Wisesty U.N., Mengko T.R., Purwarianti A. Gene mutation detection for breast cancer disease: A review. IOP Conf. Ser. Mater. Sci. Eng. 2020;830:032051. doi: 10.1088/1757-899X/830/3/032051. [DOI] [Google Scholar]

15.Mahmood T., Li J., Pei Y., Akhtar F., Imran A., Rehman K.U. A Brief Survey on Breast Cancer Diagnostic with Deep Learning Schemes Using Multi-Image Modalities. IEEE Access. 2020;8:165779–165809. doi: 10.1109/ACCESS.2020.3021343. [DOI] [Google Scholar]

16.Sutanto D.H., Ghani M.K.A. A benchmark of classification framework for non-communicable disease prediction: A review. ARPN J. Eng. Appl. Sci. 2015;10:9941–9955. [Google Scholar]

17.Gautam R., Kaur P., Sharma M. A comprehensive review on nature inspired computing algorithms for the diagnosis of chronic disorders in human beings. Prog. Artif. Intell. 2019;8:401–424. doi: 10.1007/s13748-019-00191-1. [DOI] [Google Scholar]

18.Mahmood M., Al-Khateeb B., Alwash W.M. A review on neural networks approach on classifying cancers. IAES Int. J. Artif. Intell. 2020;9:317–326. doi: 10.11591/ijai.v9.i2.pp317-326. [DOI] [Google Scholar]

19.Fatima N., Liu L., Hong S., Ahmed H. Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. IEEE Access. 2020;8:150360–150376. doi: 10.1109/ACCESS.2020.3016715. [DOI] [Google Scholar]

20.Al-Antari M.A., Al-Masni M.A., Park S.-U., Park J., Metwally M.K., Kadah Y.M., Han S.-M., Kim T.-S. An Automatic Computer-Aided Diagnosis System for Breast Cancer in Digital Mammograms via Deep Belief Network. J. Med. Biol. Eng. 2017;38:443–456. doi: 10.1007/s40846-017-0321-6. [DOI] [Google Scholar]

21.Khan F., Khan M.A., Abbas S., Athar A., Siddiqui S.Y., Khan A.H., Saeed M.A., Hussain M. Cloud-Based Breast Cancer Prediction Empowered with Soft Computing Approaches. J. Healthc. Eng. 2020;2020:8017496. doi: 10.1155/2020/8017496. [DOI] [PMC free article] [PubMed] [Google Scholar]

22.Dai B., Chen R.-C., Zhu S.-Z., Zhang W.-W. Using Random Forest Algorithm for Breast Cancer Diagnosis; Proceedings of the 2018 International Symposium on Computer, Consumer and Control (IS3C); Taichung, Taiwan. 6–8 December 2018; pp. 449–452. [DOI] [Google Scholar]

23.Tanabe K., Ikeda M., Hayashi M., Matsuo K., Yasaka M., Machida H., Shida M., Katahira T., Imanishi T., Hirasawa T., et al. Comprehensive Serum Glycopeptide Spectra Analysis Combined with Artificial Intelligence (CSGSA-AI) to Diagnose Early-Stage Ovarian Cancer. Cancers. 2020;12:2373. doi: 10.3390/cancers12092373. [DOI] [PMC free article] [PubMed]

[Google Scholar]
24.Sivapriya J., Kumar A., Sai S.S., Sriram S. Breast cancer prediction using machine learning. Int. J. Recent Technol. Eng. 2019;8:4879–4881. [Google Scholar]
25.Abunasser B.S., Al-Hiealy M.R.J., Zaqout I.S., Abu-Naser S.S. Breast Cancer Detection and Classification using Deep Learning Xception Algorithm. Int. J. Adv. Comput. Sci. Appl. 2022;13:223–228. doi: 10.14569/IJACSA.2022.0130729. [DOI] [Google Scholar]
26.Allugunti V.R. Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. Int. J. Eng. Comput. Sci. 2022;4:49–56. doi: 10.33545/26633582.2022.v4.i1a.68. [DOI] [Google Scholar]
27.Gonzalez-Angulo A.M., Morales-Vasquez F., Hortobagyi G.N. Breast Cancer Chemosensitivity. Volume 608. Springer; Berlin/Heidelberg, Germany: 2007. Overview of Resistance to Systemic Therapy in Patients with Breast Cancer; pp. 1–22. [DOI] [PubMed] [Google Scholar]
28.Elsayad A.M. Predicting the Severity of Breast Masses with Ensemble of Bayesian Classifiers. J. Comput. Sci. 2010;6:576–584. doi: 10.3844/jcssp.2010.576.584. [DOI] [Google Scholar]
29.Krizmaric M., Mertik M. Central European Conference on Information and Intelligent Systems. IOP Publishing; Bristol, UK: 2008. Application of Bayesian networks in emergency medicine. [Google Scholar]
30.Nguyen C., Wang Y., Nguyen H.N. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. J. Biomed. Sci. Eng. 2013;06:551–560. doi: 10.4236/jbise.2013.65070. [DOI] [Google Scholar]
31.Sharma S., Aggarwal A., Choudhury T. Breast cancer detection using machine learning algorithms; Proceedings of the 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS); Belgaum, India. 21–22 December 2018. [Google Scholar]
32.Kabiraj S., Raihan M., Alvi N., Afrin M., Akter L., Sohagi S.A., Podder E. Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm; Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT); Kharagpur, India. 1–3 July 2020; [DOI] [Google Scholar]
33.Al-Azzam N., Shatnawi I. Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer. Ann. Med. Surg. 2021;62:53–64. doi: 10.1016/j.amsu.2020.12.043. [DOI] [PMC free article] [PubMed] [Google Scholar]
34.Wang S., Wang Y., Wang D., Yin Y., Wang Y., Jin Y. An improved random forest-based rule extraction method for breast cancer diagnosis. Appl. Soft Comput. 2019;86:105941. doi: 10.1016/j.asoc.2019.105941. [DOI] [Google Scholar]
35.Tang X., Cai L., Meng Y., Gu C., Yang J., Yang J. A Novel Hybrid Feature Selection and Ensemble Learning Framework for Unbalanced Cancer Data Diagnosis with Transcriptome and Functional Proteomic. IEEE Access. 2021;9:51659–51668. doi: 10.1109/ACCESS.2021.3070428. [DOI] [Google Scholar]