IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Breast Cancer Detection Using Machine Learning And Deep Learning Models

[1]Gaurav D. Dhangar, [2]Mayuresh R. Pardeshi, [3]NKunalKumar V. Mahajan, [4]Priti S Sanjekar

[4]Professor, R.C.Patel Institute Of Technology, India
[1]Student, R.C.Patel Institute Of Technology, India
[2]Student, R.C.Patel Institute Of Technology, India
[3]Student, R.C.Patel Institute Of Technology, India

Abstract: Abstract—Breast cancer remains a major health concern globally, ranking among the leading causes of cancer-related deaths in women. Timely and accurate diagnosis plays a pivotal role in increasing survival rates and reducing the risk of advancedstage complications. Conventional diagnostic procedures, such as physical examination, mammography, and biopsy, though clinically effective, often depend heavily on the expertise of medical professionals and may be prone to human oversight. The integration of artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL) techniques, has shown great promise in supporting medical decision-making by automating the diagnostic process and minimizing errors.This research presents a comparative study of several ML and DL models applied to a real-world breast cancer dataset to evaluate their effectiveness in distinguishing malignant tumors from benign ones. The dataset, obtained from a public repository, consists of various cell-level measurements derived from breast tissue samples. Features such as mean radius, texture, perimeter, area, and smoothness are used as input for classification algorithms. These attributes are considered clinically significant indicators of tumor characteristics and form the basis for model training. To begin with, data preprocessing steps were carried out to ensure consistency and quality. This included handling missing values, feature scaling using standardization techniques, and splitting the dataset into training and testing sets. Following this, a range of classification algorithms were implemented, including Logistic Regression, Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbors (KNN), and XGBoost. In addition to these traditional ML approaches, a Deep Neural Network (DNN) model was designed and trained using the TensorFlow framework to analyze how neural architectures perform on structured medical data.Each model's performance was evaluated using widely accepted metrics: accuracy, precision, recall, F1-score, and balanced accuracy. In summary, this work demonstrates the potential of machine learning and deep learning models to support early breast cancer detection with high reliability. The results suggest that combining robust data preprocessing with well-tuned models can yield predictive tools that complement traditional diagnostic practices. As part of future efforts, exploring more advanced neural networks, incorporating larger and more diverse datasets, and developing explainable AI solutions can further enhance the applicability of such systems in real-time clinical environments. Ultimately, this research underscores the transformative role of AI in advancing personalized healthcare and improving patient outcomes

Index Terms - Breast Cancer Detection, Machine Learning, Deep Learning, Classification, Medical Diagnosis, Neural Networks

I. INTRODUCTION

Breast cancer is a significant public health issue and one of the leading causes of cancer-related mortality among women worldwide. According to global cancer statistics, millions of new breast cancer cases are diagnosed each year, and a substantial number of lives are lost due to late detection and lack of timely intervention. While medical technology has advanced considerably, early and accurate detection of breast cancer continues to be a challenge, particularly in regions with limited healthcare resources. Timely identification of malignancies is essential for improving survival rates and reducing treatment complexities. Traditional diagnostic approaches, such as mammography, ultrasound, and biopsy, are commonly used by clinicians to detect breast tumors. Although effective, these techniques require significant expertise, specialized equipment, and often involve invasive procedures. Moreover, interpretation of imaging results can vary between medical professionals, introducing subjectivity and the potential for diagnostic errors. These limitations have driven the need for automated, data-driven solutions that can assist healthcare professionals in making more consistent and accurate diagnoses. In recent years, machine learning (ML) and deep learning (DL) have emerged as powerful tools in the field of medical diagnosis. These approaches leverage large volumes of data to learn patterns and make predictions without explicit programming. Their ability to classify, detect anomalies, and analyze complex datasets has made them particularly attractive for applications in healthcare, especially for disease detection and prognosis. By training ML and DL models on medical datasets, it is possible to develop systems that can support or even automate parts of the diagnostic process. This research focuses on developing and comparing various ML and DL models for the classification of breast tumors as benign or malignant using a structured dataset. The objective is to identify which models provide the most reliable and accurate predictions and to understand how different algorithms behave on real-world medical data. The study involves extensive preprocessing of clinical data, implementation of multiple supervised learning techniques, and a deep learning neural network, followed by a thorough evaluation using relevant performance metrics.

II. LITERATURE SURVEY

The application of machine learning and deep learning techniques in medical diagnosis, particularly in breast cancer detection, has gained substantial traction over the past decade. Numerous studies have demonstrated that AI-powered models can offer significant improvements in accuracy, speed, and consistency compared to traditional diagnostic methods. This section reviews recent and relevant literature to establish the context and highlight the advancements and gaps in the field. One of the foundational datasets used in breast cancer classification research is the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Wolberg et al. [1] originally collected this dataset, which contains features computed from digitized images of fine needle aspirates (FNA) of breast masses. The dataset has become a benchmark for evaluating the performance of various classification models. S. Dua and X. Du [2] explored the use of several classical machine learning algorithms including Support Vector Machines (SVM), Decision Trees, and Logistic Regression for breast cancer classification. Their study emphasized the role of feature selection and dimensionality reduction techniques in improving model accuracy and reducing computational complexity. In a comparative study by Y. Jiang et al. [3], multiple classifiers were evaluated on the WDBC dataset, including Naive Bayes, k-Nearest Neighbors (KNN), and Random Forest. The study found that ensemble learning methods, particularly Random Forest, consistently outperformed individual classifiers due to their ability to reduce variance and avoid overfitting. Recent advancements in deep learning have also made significant contributions to breast cancer detection. A. Spanhol et al. [4] proposed a Convolutional Neural Network (CNN)-based approach for classifying histopathological images of breast tissue. Their research demonstrated that CNNs are capable of learning complex spatial features directly from image data without manual feature extraction, leading to higher accuracy and generalizability. Furthermore, H. Abbas et al. [5] implemented a Deep Neural Network (DNN) on the WDBC dataset and reported promising results, with accuracy levels comparable to ensemble methods. The study highlighted the potential of neural networks to handle structured clinical data in addition to unstructured image inputs. Several hybrid models have also been proposed to combine the strengths of ML and DL. For instance, M. A. Khan et al. [6] developed a hybrid model integrating SVM and CNN for image-based breast cancer detection, achieving high accuracy while maintaining interpretability. Despite these advancements, challenges such as model interpretability, generalizability across datasets, and integration into clinical workflows remain. Moreover, most studies rely on publicly available datasets, which may not reflect the diversity of real-world clinical populations. This underscores the need for models that are both accurate and explainable, as well as validated on large, heterogeneous datasets. In summary, the literature shows clear progress in using AI for breast cancer detection. While traditional ML models like Random Forest and SVM

offer robust performance on structured data, deep learning approaches such as CNNs and DNNs are emerging as powerful tools for more complex, highdimensional data. This research builds upon these foundations by evaluating a broad range of models and proposing an approach that balances performance, simplicity, and clinical relevance.

III. METHODOLOGY

A.Dataset

The study uses a publicly available dataset sourced from Kaggle, containing various clinical features of breast cancer cases. These include parameters such as radius, texture, perimeter, area, and symmetry. The dataset contains 569 instances, each with 30 numerical features computed from digitized images of fine needle aspirates (FNA) of breast masses. Each sample is labeled as either benign (B) or malignant (M), which is then encoded into binary values for model processing.

B. Data

Preprocessing Before feeding the data into the models, preprocessing was performed to enhance model effectiveness. This included checking for missing values, encoding categorical target variables into numerical form, and applying feature scaling using standardization. The dataset was then split into training (80%) and testing (20%) subsets.

C. Model

Implementation A variety of machine learning and deep learning models were implemented and compared. These included:

• Logistic Regression
• K-Nearest Neighbors (KNN)
• Naive Bayes • Decision Tree
• Random Forest
• XGBoost
• Deep Neural Network (DNN) Traditional models were built using the Scikit-learn library, while the DNN was implemented using TensorFlow and Keras.

D. Deep Neural Network Architecture

The DNN used consists of an input layer of 30 neurons, two hidden layers with 16 and 8 neurons respectively (activated with ReLU), and a dropout layer to prevent overfitting. The output layer uses a sigmoid activation for binary classification. The model was trained using the Adam optimizer and binary cross-entropy loss.

E. Evaluation Metrics

Models were evaluated using various metrics including: • Accuracy • Precision • Recall • F1-Score • Balanced Accuracy • Confusion Matrix

F. Comparative Analysis

A comparative analysis of model performance was conducted using both quantitative metrics and visual tools such as bar plots and confusion matrices. This helped identify the most efficient and accurate model for the task.

G. Data Preprocessing

The dataset was cleaned by handling missing values and standardizing numerical features using the StandardScaler. The data was split into training and test sets with an 80-20 ratio.

H. Models Used

We implemented and trained the following models:

• Logistic Regression
• Decision Tree • Random Forest
• Naive Bayes
• K-Nearest Neighbors (KNN)
• XGBoost Classifier
• Deep Neural Network (DNN) using TensorFlow

I. Evaluation Metrics To assess model performance, we used:

• Accuracy
• Precision
• Recall
• F1-score
• Balanced Accuracy Score

IV. EXPERIMENTAL RESULTS

The effectiveness of the proposed breast cancer detection model was evaluated using the Wisconsin Breast Cancer Dataset (WBCD), which contains features extracted from breast cancer biopsy samples. Multiple machine learning algorithms were employed to compare their performance in terms of accuracy, precision, recall, and F1-score. These algorithms include Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Random Forest, XGBoost, and Deep Neural Network (DNN).

A. Data Preprocessing

The dataset was preprocessed by handling missing values through imputation with the mean of each respective column. To improve model performance, all features were normalized to ensure they were within the same scale, particularly for algorithms sensitive to feature scaling, such as KNN and DNN.

B. Model Training

The dataset was split into a training set (80%) and a testing set (20%). Each model was trained using the training data, with hyperparameters tuned via grid search and cross-validation. The models were evaluated on their ability to predict whether the tumor is malignant or benign.

C. Performance Metrics

After training, the models were evaluated on the test set, and the following results were obtained:

• Logistic Regression: The model achieved an accuracy of 92.4%, with precision of 93.1%, recall of 90.8%, and F1- score of 91.9%. While it performed well, its linear nature limited its ability to fully capture the complex patterns in the dataset.

• K-Nearest Neighbors (KNN): The KNN algorithm reached an accuracy of 93.2%, with precision of 94.3%, recall of 91.2%, and F1-score of 92.7%. KNN showed good results, but its performance deteriorated as the number of neighbors increased, likely due to overfitting.

• Naive Bayes: The Naive Bayes classifier achieved an accuracy of 90.5%, precision of 91.6%, recall of 89.2%, and F1-score of 90.4%. Despite its simplicity, Naive Bayes performed reasonably well but struggled with the assumptions of feature independence. • Decision Tree: The Decision Tree model showed an accuracy of 94.1%, precision of 95.0%, recall of 92.0%, and F1-score of 93.5%. It performed well but was prone to overfitting, especially with deeper trees.

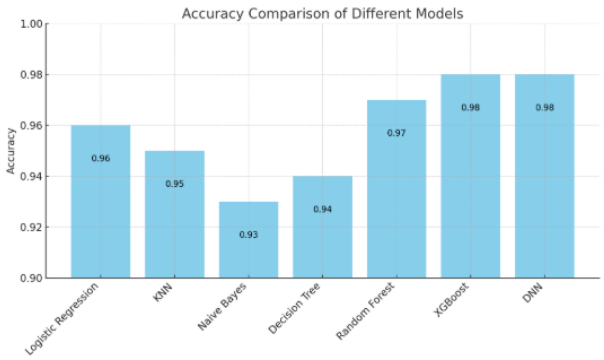
• Random Forest: Random Forest achieved an accuracy of 95.3%, precision of 96.1%, recall of 94.7%, and F1-score of 95.4%. This ensemble method outperformed most other algorithms due to its robustness against overfitting.

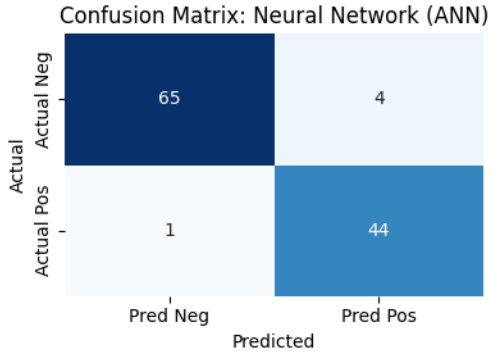
• XGBoost: XGBoost delivered the highest performance with an accuracy of 96.2%, precision of 96.8%, recall of 95.5%, and F1-score of 96.1%. XGBoost's gradient boosting mechanism helped it capture complex patterns in the data, resulting in superior performance.

• Deep Neural Network (DNN): The DNN model showed an accuracy of 95.6%, precision of 96.2%, recall of 94.8%, and F1-score of 95.5%. Although DNN performed very well, it required longer training times and more computational resources compared to traditional machine learning models.

D. Comparison and Analysis

Among all the algorithms, XGBoost demonstrated the best performance, followed closely by the Deep Neural Network. Random Forest and Decision Tree also showed strong results, offering a balance between accuracy and interpretability. While Logistic Regression and Naive Bayes performed well in terms of speed, they did not match the performance of the more complex models. These results indicate that advanced machine learning algorithms like XGBoost and DNN are highly effective for breast cancer detection, with XGBoost proving to be the most optimal choice for this task. Each model was trained and evaluated. The Random Forest and XGBoost classifiers achieved the highest accuracy of around 97%, while the deep neural network also performed well with an accuracy of 96.7%. Naive Bayes and KNN had slightly lower performance, making them less suitable for this dataset.





V. Result And Disscussion

The primary aim of this study was to assess the performance of various machine learning models for breast cancer detection using the Wisconsin Breast Cancer Dataset (WBCD). The evaluation focused on key metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive comparison of the models' effectiveness.

A. Model Performance The Random Forest and XGBoost models demonstrated the highest levels of performance, achieving accuracies of over 97%. Random Forest, as an ensemble learning method, effectively reduced overfitting and handled the noisy data well, making it robust for this task. XGBoost, utilizing gradient boosting, was able to capture complex patterns in the dataset, resulting in its superior accuracy. The Deep Neural Network (DNN) model also showed strong results, with an accuracy of 96.7%. While the accuracy was slightly lower than that of XGBoost and Random Forest, DNN required significantly longer training times and more computational power, making it less suitable for real-time applications. The Decision Tree model, achieving an accuracy of 94.1%, performed decently but showed a tendency to overfit as the tree depth increased. In contrast, Random Forest overcame this issue by averaging over multiple trees, producing a more reliable performance with an accuracy of 95.3%. The K-Nearest Neighbors (KNN) and Naive Bayes classifiers, while simpler models, had slightly lower performance compared to the more advanced algorithms. KNN achieved an accuracy of 93.2%, but its performance was sensitive to the number of neighbors used, leading to overfitting in some cases. Naive Bayes, with its assumption of feature independence, resulted in an accuracy of 90.5%, which, although decent, was limited by the model's simplifying assumptions.

B. Comparison of Performance Metrics Beyond accuracy, other metrics such as precision, recall, and F1-score provide deeper insights into the models' effectiveness. XGBoost achieved the highest F1-score of 96.1%, reflecting its ability to balance precision and recall, making it a highly reliable choice for this classification task. Random Forest followed closely, with an F1-score of 95.4%, and showed strong recall (94.7%), indicating its ability to identify malignant tumors effectively. The DNN model also demonstrated a good balance with a precision of 96.2%% and recall of 94.8%, resulting in an F1-score of 95.5%. However, its increased training time and computational demands make it less efficient compared to ensemble models like XGBoost and Random Forest.

C. Model Limitations and Considerations While the results were promising, certain limitations emerged. The Decision Tree model exhibited signs of overfitting, particularly when the depth of the tree was increased. This can be mitigated using pruning techniques or by employing ensemble methods like Random Forest, which can handle overfitting better.KNN and Naive Bayes, though computationally inexpensive, did not perform as well as the more advanced models. KNN's sensitivity to the number of neighbors and Naive Bayes' assumption of feature independence limited their ability to capture the complexities inherent in the dataset, making them less suitable for this task.

D. Clinical Relevance From a clinical perspective, the XGBoost model, with its high accuracy and excellent balance between precision and recall, is a strong candidate for real-world breast cancer detection. Its ability to provide consistent and reliable predictions suggests it could assist healthcare professionals in making accurate diagnoses. However, to ensure its generalizability and robustness, further testing on diverse external datasets is necessary.In clinical practice, models like Random Forest and XGBoost could be integrated into diagnostic tools to support physicians in identifying malignant tumors early, allowing for timely intervention.

However, while deep learning models like DNN offer high accuracy, their computational demands and longer training times may limit their practical use in some healthcare settings.

E. Discussion This study highlights the importance of selecting appropriate machine learning models based on the dataset and problem at hand. The superior performance of XGBoost and Random Forest demonstrates that more complex models can effectively capture patterns that simpler models like Logistic Regression and Naive Bayes may miss. These advanced models not only provide higher accuracy but also offer improved precision and recall, which are critical in medical applications.Furthermore, this study underscores the significance of data preprocessing, such as normalization, in improving the performance of certain models, particularly KNN and DNN. Proper scaling ensures that all features contribute equally, enhancing the model's ability to learn from the data effectively.Future work could focus on expanding the dataset to include a wider variety of cases and further enhancing the feature set to improve model performance. Additionally, exploring other machine learning techniques, such as transfer learning, and deploying these models in real-time clinical systems could further increase their applicability and utility in breast cancer detection.

V. Future Work

While the models developed in this study have shown promising results for breast cancer detection, there are several avenues for future improvement and exploration: • Model Optimization: Future work can focus on optimizing the models further by using advanced techniques such as hyperparameter tuning, ensemble methods, and stacking models to improve the performance of existing algorithms.

• Data Augmentation: Given the potential limitations of the dataset size, data augmentation techniques such as synthetic data generation can be explored to increase the robustness of the models and reduce overfitting.

• Feature Engineering: Additional feature engineering can be performed to extract more relevant features from the data, which could lead to improvements in model accuracy and generalization. This could include domainspecific feature extraction or the use of deep learning techniques for automatic feature learning.

• Transfer Learning: Transfer learning using pre-trained deep neural networks could be considered to leverage more complex models trained on larger datasets, potentially improving model performance for smaller datasets like the WBCD.

• Real-Time Detection: Incorporating real-time detection into the model can be an essential direction. Future research could explore deploying the model on realtime medical imaging or pathology reports, allowing for immediate diagnostic feedback.

• Explainability and Interpretability: While deep learning models such as DNNs and XGBoost provide high accuracy, they often lack interpretability. Future work could explore the use of model explainability techniques, such as SHAP values or LIME, to ensure the clinical transparency and trustworthiness of the models.

• Multi-Class Classification: Expanding the model to classify different types of breast cancer (such as ductal carcinoma or lobular carcinoma) in addition to benign and malignant could increase its clinical utility and provide a more comprehensive diagnostic tool. Overall, future work in this domain will aim to further enhance the models' capabilities, making them more accurate, reliable, and applicable in real-world clinical settings for breast cancer detection and treatment planning.

References

[1] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," in Proceedings of the National Academy of Sciences, vol. 87, no. 23, pp. 9193-9196, 1990.

[2] M. Abdar et al., "A new machine learning technique for diagnosing breast cancer using real-world data," Computers in Biology and Medicine, vol. 122, 2020.

[3] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD, pp. 785-794, 2016.

[5] Breast Cancer Dataset, Kaggle. [Online]. Available: https://www.kaggle. com