Improving Breast Cancer Detection with Random Forest Algorithm and the Breakhis Dataset

A Vidhya, Dharshana R S, Dinesh R, Dinesh S
Dept of Computer Science and Engineering SRM Valliammai Engineering College

***Abstract*—Breast cancer is still one of the health issues globally that requires proper and effective diagnostic techniques. This study compares the efficacy of several machine learning models to classify breast cancer histopathological images. BreakHis dataset was used by applying preprocessing methods in the form of image resizing and standardization. Different models, such as Logistic Regression, Random Forest, Support Vector Machine (SVM), Artificial Neural Networks (ANN), Decision Tree, Naïve Bayes, and K-Nearest Neighbors (KNN), are evaluated on main performance metrics such as accuracy, precision, recall, and F1-score. Out of these, the highest accuracy (83.38%) was obtained with the Random Forest algorithm, which indicates its potential in enhancing breast cancer diagnosis. The results emphasize the significance of medical imaging using machine learning and how it can strengthen automated diagnostic aid systems.**

***Index Terms*—Breast Cancer Classification, Machine Learning in Healthcare, Histopathological Image Analysis, BreakHis Dataset**

I. INTRODUCTION

Breast cancer remains among the most prevalent and fatal conditions in the world, posing an enormous load on medical resources. Characterized by the uncontrollable growth of neoplastic cells in the breast tissue, timely detection and proper identification are critical for better patient outcomes and reduced mortality. Traditional diagnostic methods, such as histopathological examination, play an important role in the diagnosis of cancer cells through microscopic analysis of tissue samples. However, it is extremely labor-intensive and calls for high degrees of skill enjoyed by experienced pathologists. It is susceptible to variability and human error in difficult or borderline cases, thus being limiting in application in high-throughput healthcare environments.

Machine learning technologies are going to be the optimal solution to such problems. Machine learning models can perform automated histopathological image analysis with fast and reproducible diagnosis. These models not only reduce the workload for pathologists but also enhance diagnostic accuracy, and they are therefore valuable tools in modern healthcare. Despite these advantages, there is still no consensus on which machine learning algorithms are most appropriate for breast cancer classification. Comparative research comparing the performance of a number of models must be conducted in order to ascertain the optimal algorithm for this valuable application.

This study focuses on evaluating the performance of seven various machine learning models—Logistic Regression, Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Tree, Naïve Bayes, and K-Nearest Neighbors (KNN)—for classifying histopathological images from the BreakHis dataset. The dataset has over 7,900 images labeled as either "benign" or "malignant," making it a valuable source of data for training and testing these models. All algorithms are compared against baseline key performance factors such as accuracy, precision, recall, and F1-score for a complete picture of their disadvantages and strengths.

Random Forest, an ensemble learning that is accurate and robust in nature, is taken as a benchmark against which other algorithms are measured. Deep learning algorithms like ANN, whose ability to recognize complex patterns complicates dataset size and computational cost, are hindered by both. Systematic comparison of these algorithms, the paper aims to determine which model is most reliable and accurate when it comes to classifying breast cancer.

The implications of this work for practice are significant. Diagnostic aids with computers can assist pathologists by creating preliminary classifications so that faster decisions can be made and delays in diagnosis can be avoided. In addition, embedding machine learning algorithms in clinical workflow can enhance consistency and accuracy in diagnoses, finally leading to better patient outcomes. This study not only contributes to what is already known about machine learning methods in healthcare but also

draws attention to the potential of such technologies to change the diagnosis of breast cancer.

In short, the broad purpose of this study is to address the prevailing gaps in comparative evaluation of machine learning models for the classification of breast cancer. Using the BreakHis dataset and rigorous evaluation criteria, this study aims to identify the algorithm that offers optimal accuracy and reliability for use in clinical diagnostics. With this study, we aim to enhance the development of automated tools to aid healthcare professionals in delivering timely and impactful treatment to breast cancer patients.

II. RELATED WORK

The area of breast cancer diagnosis has experienced tremendous growth with the incorporation of machine learning methods. This section discusses the literature under major themes, with emphasis on classification algorithms, sophisticated frameworks, and ensemble techniques for processing histopathological images.

A. Machine Learning in Breast Cancer Classification

A number of studies have shown the capability of machine learning to automate the classification of breast cancer histopathological images.

1)Shahram Taheri et al. proposed a new Multi-Level Feature Fusion CNN (MLF2-CNN) model that is built using DenseNet-121 to categorize breast cancer images as either benign or malignant. The state-of-the-art performance on BreakHis dataset highlighted the efficiency of the model to extract hierarchical features. MLF2-CNN provided a viable solution for versatile scenarios by dealing with the limitation of magnification-specific image processing.

2) Likewise, Hosameldin Ahmed and Asoke Nandi developed the MoEffNet framework, blending the feature extraction properties of EfficientNet with a Mixture of Experts (MoEs) strategy. This mixed architecture surpassed traditional CNN models with an AUC > 0.99 across various datasets, including BreakHis.
These researches highlight the significance of tapping into the latest architectures to increase feature extraction and the accuracy of classification.

B. Comparative Studies on Classification Algorithms

Comparative studies of machine learning algorithms offer essential insights into their relative strengths and limitations for breast cancer diagnosis.

1)Usman Naseem et al. investigated the application of ensemble methods using algorithms such as Artificial Neural Networks (ANN), Random Forest, and Naïve Bayes. Their ensemble model was 98.83% accurate on benchmark data, better than individual models.

2)Aya Farrag et al. utilized stage-specific machine learning models for treatment planning based on survival in another study. While it was treatment strategies focused, the research highlighted the need for having good algorithm choice in cancer diagnosis and management.
These comparison studies emphasize the need to assess more than one algorithm to select the best solution for particular data sets and applications in the clinical context.

C. Feature Engineering and Advanced Frameworks

Feature engineering is crucial in improving the predictive capability of machine learning algorithms for breast cancer classification.

1)Rong-Ho Lin et al. analyzed survival and treatment efficacy with feature-rich data. Although their research was centered on survival analysis with the Cox proportional hazards model, it gave important insights into how features such as hormone levels and demographic factors affect patient outcomes.

2) Also, research such as that conducted by Ahmed and Nandi illustrated how hybrid frameworks can utilize engineered features to enhance classification performance. Through the merging of low-, mid-, and high-level features from images, their models displayed excellent diagnostic results.
These publications illustrate the significance of thorough feature extraction and incorporation in the design of high-capacity models.

D. Traditional Approaches Limitations

Classic machine learning models like Logistic Regression, Decision Trees, and Support Vector Machines (SVMs) have been used for a long time to

diagnose breast cancer. These methods have limitations by nature:

1)Logistic Regression is based on a linear relationship between the features and the outcomes, which is not the case in the real world for complex histopathological data.

2)Decision Trees are easy to interpret but are susceptible to overfitting and do not generalize well with large datasets.

3)SVMs are difficult with high-dimensional data and need to be tuned and kernel chosen with care.
Notwithstanding their disadvantages, such models are useful for benchmarking and serve as the foundation for more sophisticated techniques.

E. Ensemble Learning in Cancer Diagnostics

Ensemble learning has been identified as a promising technique to improve the performance and reliability of machine learning models in breast cancer diagnostics.

1)Usman Naseem et al. illustrated how ensemble approaches, which harness the strengths of diverse algorithms, performed better consistently than single models. Their implementation of voting classifiers and boosting proved highly effective in increasing diagnostic accuracy and reliability.

2)Ensemble methods such as AdaBoost and Gradient Boosting have also found extensive application for their potential to rectify errors and emphasize difficult cases, as noted in various comparative analyses.

These results illustrate the strength of ensemble approaches to designing strong diagnostic systems that respond to the limitations of a single model.

F. Research Gaps and Opportunities

Although great strides have been taken, current research tends to concentrate on a single model or single dataset, making them less generalizable. The need for common frameworks uniting classification and survival analysis is also apparent. This work seeks to address these issues by presenting an overall assessment of various machine learning models against the BreakHis dataset, with insights into their real-world applicability in clinical diagnosis.

III. PROPOSED MODEL

In this study, we introduce the model that seeks to classify breast cancer histopathological images into benign and malignant classes based on a comparative approach of seven machine learning algorithms. It utilizes the BreakHis dataset to evaluate the performance of the algorithms based on major evaluation metrics such as accuracy, precision, recall, and F1-score.

A.Key Elements of the Proposed Approach

1)Dataset
BreakHis dataset is utilized consisting of 7,909 images of breast tissues stained with H&E. These images are categorized as benign and malignant and carry various magnifications (40x, 100x, 200x, 400x).

2)Preprocessing
- Images normalized to 224x224 pixel size for uniformity.
- Normalization of pixel intensities to make mean = 0 and variance = 1.
- Divide into training (80%) and test (20%) sets for model validation.

3)Machine Learning Algorithms
K-Nearest Neighbors (KNN): A case-based learning algorithm that classifies a point by the majority class of its closest neighbors. It is easy and efficient but computationally costly for large data sets and sensitive to irrelevant attributes.

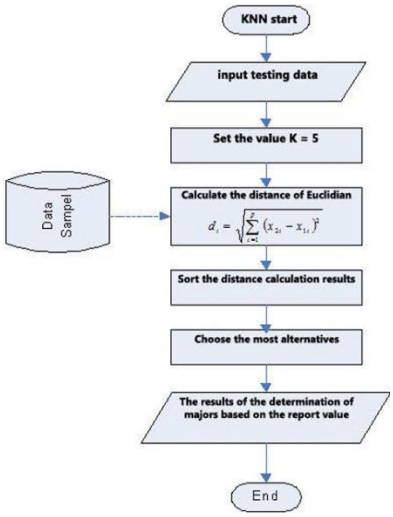

Fig.1. K-Nearest Neighbors

Logistic Regression: A statistical algorithm used for binary classification that models the probability of an instance belonging to a class. It assumes a linear relationship between the features and the log-odds of the outcome, making it simple and interpretable but limited in handling non-linear data.

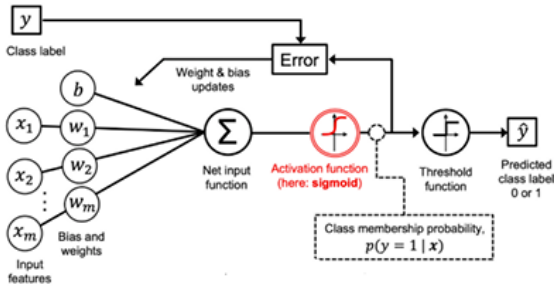

Fig.2. Logistic Regression

Random Forest: An ensemble learning technique that builds a collection of decision trees at training time. It combines predictions from the trees to improve accuracy and prevent overfitting, which makes it less sensitive to noisy data.

Support Vector Machine (SVM): A classifier that places data points with a best hyperplane, maximizing the margin between classes. SVM is powerful in high-dimensional space and can be used to classify non-linearly with kernel functions.

Artificial Neural Network (ANN): A computer model based on biological neural networks, made up of layers of connected neurons. ANNs are good at learning intricate patterns in data but need large amounts of computational power and big data.

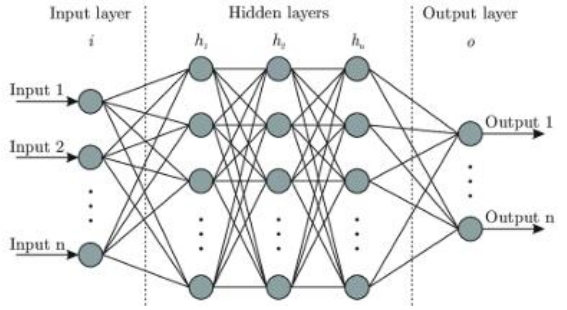

Fig.3. Artificial Neural Network

Decision Tree: A tree classifier where the dataset is divided on the basis of feature values to form branches and nodes that result in a class prediction. Decision Trees are simple to interpret but are likely to overfit, particularly on noisy data.

Naive Bayes: A Bayes theorem-based probabilistic classifier that presumes feature independence. It is computationally simple and effective with small datasets but performs poorly on correlated features.

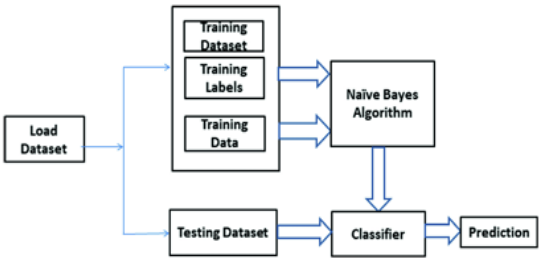

Fig.4. Naïve Bayes

4)Evaluation Metrics

Accuracy: Proportion of correctly classified instances.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}} = \frac{TP + TN}{TP + TN + FP + F}$$

Precision: Fraction of relevant instances among the retrieved instances.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = \frac{TP}{TP + F}$$

Recall: Ability of the model to identify all relevant cases.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \frac{TP}{TP + F}$$

F1-Score: Harmonic mean of precision and recall for imbalanced datasets.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2TP + FP + F}$$

5)Implementation Strategy

Train every algorithm on the preprocessed dataset.

Compare and evaluate the models using the above metrics.

Determine the algorithm with the greatest reliability and diagnostic accuracy.

6)Outcome

The suggested model seeks to identify the optimal-performing algorithm for classifying breast cancer histopathological images, presenting a valid automated solution to assist pathologists in diagnostic processes.

IV. DISCUSSION

A.Algorithm Performance Analysis:

The seven machine learning algorithms' performance showcases their advantages and disadvantages in classifying breast cancer histopathological images. Logistic Regression offered a strong baseline with moderate accuracy and interpretability but had difficulty in capturing intricate patterns because of its linear assumptions. Random Forest performed well with good accuracy and robustness to overfitting, which made it appropriate for dealing with noisy data sets while giving insights into feature ranking. SVM was effective, particularly in high-dimensional space, with the use of kernel functions augmenting its capacity to deal with non-linear relationships. ANN showed its capacity to identify complex patterns in the data, although its susceptibility to overfitting and heavy computational requirements were areas of major concern. Decision Tree provided simplicity and interpretability but was prone to overfitting, especially with noisy data. Naive Bayes was efficient for smaller datasets but was not effective with correlated features, which are prevalent in medical imaging. Finally, KNN provided competitive accuracy but had issues with computational efficiency on large datasets and sensitivity to irrelevant features.

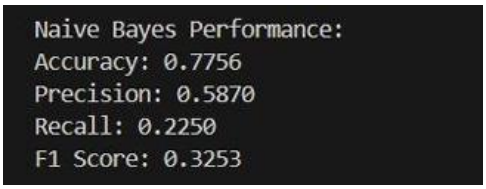

Fig.5. Naïve Bayes Performance

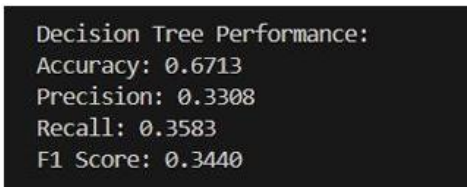

Fig.6. Decision tree performance

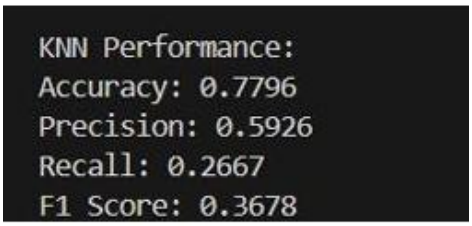

Fig.7. KNN Performance

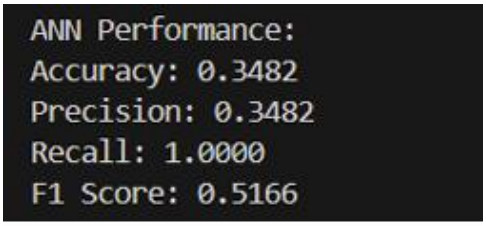

Fig.8. ANN Performance

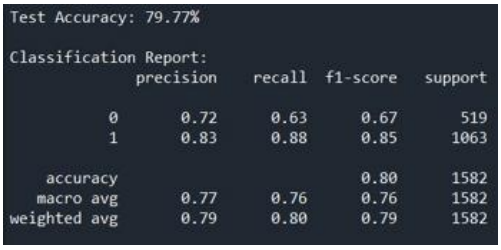

Fig.9. Logistic Regression Performance

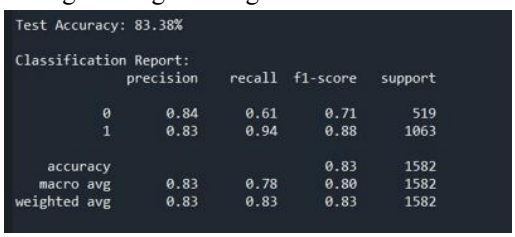

Fig.10. Random Forest Performance

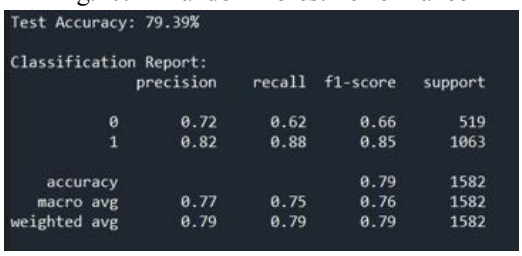

Fig.11. Support Vector Machine Performance

MODEL EVALUTION RESULT	ACCURACY
Logistic Regression	79.77
Random Forest	83.38
Support Vector Machine (SVM)	79.39
Artificial Neural Network (ANN)	34.82
Decision Tree	67.13
Naive Bayes	77.56
K-Nearest Neighbors (K-NN)	77.96

Fig.12. Accuracy of all models

B .Practical Implications:

The study emphasizes the potential of machine learning models, particularly ensemble methods like Random Forest and advanced algorithms like ANN, in automating breast cancer diagnosis. These models can significantly enhance diagnostic accuracy and consistency while reducing the workload and time required by pathologists. Furthermore, their application in large-scale screening programs could be transformative, particularly in resource-limited settings where access to expert pathologists is constrained.

C .Challenges and Limitations:

There were some challenges in the course of the analysis. All models' performance was preprocessor

technique- sensitive, i.e., resizing and standardization of images, which had an effect on the reproducibility of the results. Furthermore, the class imbalance inherent in the dataset challenged model training with sophisticated techniques like weighted loss or synthetic oversampling to mitigate bias. The computational load was quite pronounced for Models like ANN, restricting their suitability in low-resource health setups.

D .Future Directions:

Based on these results, future research may investigate hybrid approaches that leverage the strengths of several algorithms, e.g., ANN with Random Forest, to improve overall performance. Sophisticated preprocessing methods such as feature extraction using pre-trained deep learning models, may enable more sophisticated analysis. Validating the proposed models on additional datasets with varied populations and imaging conditions would improve their generalizability .Additionally, explainable AI (XAI) methods could make the models more explainable, assisting clinicians in interpretation of predictions and establishing confidence in automated systems.

V. CONCLUSION

In summary, this research provides evidence of the promising potential of machine learning models, especially ensemble models such as Random Forest and complex networks such as ANN, in the automation of the diagnosis of breast cancer histopathological images. The models provide substantial improvements in diagnostic accuracy, reproducibility, and efficiency with the potential to facilitate large-scale screening, particularly in resource-deprived regions. In spite of difficulties like class imbalance, computational requirements, and sensitivity to preprocessing, the results underscore the significance of further development of these models using hybrid methodologies, sophisticated feature extraction, and real-world evaluation. With advancing developments, such machine learning models may become fundamental tools in optimizing early detection and patient outcomes for breast cancer management.

REFERENCES

[1] N. Fatima, L. Liu, S. Hong and H. Ahmed, "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis," in IEEE Access, vol. 8, pp. 150360-150376, 2020, doi: 10.1109/ACCESS.2020.3016715.

[2] U. Naseem et al., "An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers," in IEEE Access, vol. 10, pp. 78242-78252, 2022, doi: 10.1109/ACCESS.2022.3174599.

[3] A. U. Haq et al., "Detection of Breast Cancer Through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques," in IEEE Access, vol. 9, pp. 22090-22105, 2021, doi: 10.1109/ACCESS.2021.3055806.

[4] M. A. Rahman et al., "Enhancing Early Breast Cancer Detection Through Advanced Data Analysis," in IEEE Access, vol. 12, pp. 161941-161953, 2024, doi: 10.1109/ACCESS.2024.3483095.

[5] R. Martínez-Licort et al., "Breast Carcinoma Prediction Through Integration of Machine Learning Models," in IEEE Access, vol. 12, pp. 134635-134650, 2024, doi: 10.1109/ACCESS.2024.3431998.