Predictive Breast Cancer Modeling Using Machine Learning

Brindhasri S V
Dept of Computer Science
Bharathiar University Coimbatore-641046

Abstract- *Breast cancer is a form of tumor that develops within breast tissues and remains the most prevalent cancer among women globally, ranking as one of the leading causes of female mortality. This survey examines the evolving landscape of predictive modeling for breast cancer using data mining techniques. It investigates the application of various algorithms—including decision trees, support vector machines, and neural networks—for effective breast cancer prediction. The survey provides a thorough review of processes such as feature selection, model training, and validation strategies, and synthesizes key findings from multiple datasets. By critically analyzing existing literature, this work aims to deepen understanding in the field, offering insights into the advancements, current challenges, and future prospects of predictive modeling in breast cancer research. Ultimately, this survey contributes to the broader discourse on data-driven approaches in healthcare, emphasizing their role in enhancing diagnostic and prognostic capabilities in breast cancer management.*

Keywords- Breast cancer, data mining, decision trees, support vector machines, neural networks, healthcare, diagnosis.

I. INTRODUCTION

Breast cancer remains a leading cause of mortality among women globally, emphasizing the critical need for accurate and timely diagnosis. The efficiency of early detection is directly correlated with increased survival rates and more effective treatment planning. Traditional diagnostic methods, though clinically established, often suffer from limitations such as subjectivity, delayed results, and variability across medical practitioners.

In recent years, the integration of machine learning (ML) techniques in the medical domain has offered a transformative shift in disease detection and prognosis. Specifically, supervised classification algorithms have demonstrated strong potential in the domain of cancer prediction. These algorithms can learn from historical patient data and identify patterns that assist in classifying new cases as either benign or malignant, thereby enhancing clinical decision-making.

This research focuses on the application and evaluation of three prominent classification algorithms—**Random Forest**, **K-Nearest Neighbors (KNN)**, and **Decision Tree**—for breast cancer prediction. These algorithms were selected based on their robustness, interpretability, and widespread usage in medical data analysis.

The primary objective is to assess the performance of these classifiers using a real-world breast cancer dataset. Key evaluation metrics include **Precision**, **Recall**, **F1-Measure**, and **Accuracy**, which provide a comprehensive understanding of each model's diagnostic reliability.

The study aims to identify the most effective model for clinical application, contributing to the broader goal of integrating intelligent decision support systems into healthcare. By leveraging data-driven methodologies, this work seeks to aid clinicians in early cancer detection, ultimately improving patient outcomes and supporting the evolution of personalized medicine.

II. LITERATURE REVIEW

Numerous studies have employed classification techniques on various datasets to analyze and compare predictive performance.

In [1], researchers introduced a novel breast cancer prediction model using the Weighted Naïve Bayes algorithm. This model demonstrated improved accuracy and was recognized for its simplicity, readability, efficiency, and potential as a clinical decision-support tool.

Study [2] focused on diagnosing breast cancer by analyzing 10 features used by pathologists to determine tumor malignancy. Using Weka, ZeroR, and decision trees, the study predicted the target class indicating benign or malignant tumors.

In [3], classification algorithms C4.5 and Naïve Bayes were applied. C4.5 achieved an accuracy of 98.10%, outperforming Naïve Bayes, which achieved 95.85%. Non-

cancerous and irrelevant attributes such as social, racial, and geographic data were removed during preprocessing.

The objective of [4] was to develop a predictive model using three algorithms: Naïve Bayes, Radial Basis Function (RBF) Network, and J48. Results indicated Naïve Bayes as the most accurate with 97.36%, chosen for its simplicity and probabilistic prediction capability.

In [5], Naïve Bayes and J48 were compared based on accuracy and execution time. Naïve Bayes demonstrated superior accuracy and faster execution, highlighting its efficiency over J48.

The researchers in [6] proposed a Breast Cancer Diagnosis (BCD) model using Support Vector Machine (SVM) with 10-fold cross-validation. Principal Component Analysis (PCA) was used for dimensionality reduction, improving model accuracy. The BCD model outperformed algorithms like decision trees, random forest, k-NN, SGD, AdaBoost, neural networks, and Naïve Bayes, evaluated through F1 score, ROC curve, accuracy, and other performance metrics.

Study [7] proposed the Extensible Breast Cancer Prognosis Framework (XBPF) to assess susceptibility, recurrence, and survivability. It incorporated a Representative Feature Subset Selection (RFSS) algorithm with SVM, using the SEER dataset. Results showed that SVM-RFSS significantly enhanced prognosis prediction.

In [8], an interactive data visualization tool was introduced to compare three machine learning algorithms on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The tool allowed user input and dynamic updates in performance visualization. SVM emerged as the top-performing classifier with 97.85% accuracy, outperforming KNN and Normal Bayes.

Lastly, [9] presented a fully automated method for breast cancer detection using deep convolutional neural networks (DCNN) trained on histopathological images. Data augmentation and pooling operations were used to optimize performance, resulting in an average classification accuracy of 92.50%.

III. METHODOLOGY

The study followed a structured methodology to build predictive models for breast cancer diagnosis. Publicly available datasets were collected and preprocessed, including normalization and imputation of missing values. Feature selection was carried out using both statistical techniques and machine learning-based methods to identify relevant variables for model training.

Three classification algorithms—Decision Tree, Support Vector Machine (SVM), and Neural Network—were implemented using Python libraries like scikit-learn and TensorFlow. The dataset was split using k-fold cross-validation to evaluate the models' performance and ensure robustness.

To assess model accuracy, performance metrics such as Accuracy, Precision, Recall, and F1-Score were calculated. A comparative analysis of these metrics was conducted to determine the most effective model for early breast cancer detection. The results guided the identification of the optimal approach for reliable diagnosis.

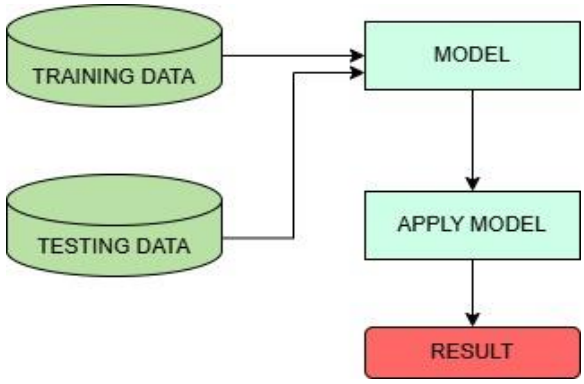

Fig.1. Framework

DATASE*T*

The dataset used for predicting breast cancer is obtained from the Kaggle Data Mining repository. Kaggle is a platform that hosts various datasets for data mining algorithm implementation. This particular dataset is a real-world dataset. It consists of 569 instances, each representing a case. There are 32 attributes in total, providing detailed information about each instance. These attributes are used to train machine learning models. The dataset is well-suited for predictive modeling tasks. It serves as a valuable resource for understanding and analyzing breast cancer data. The repository ensures access to various datasets for research and learning purposes.

PRE-PROCESSING

Data preprocessing is the process of preparing raw data for use in a data mining model. It is the first and most essential step in developing an effective data mining model. The goal of preprocessing is to transform the data into a clean and structured format suitable for analysis. This step involves addressing issues like missing values, outliers, and noise. Preprocessing techniques also include normalizing, scaling, and encoding data. The quality of the data directly impacts the performance of the model. By applying proper transformations, the data becomes more reliable for prediction. Effective preprocessing improves model accuracy and efficiency. It helps in identifying patterns and insights more effectively. Overall, preprocessing is critical for building a robust and accurate data mining model.

PREDICTION METHODS

The process is simple and efficient, allowing for quick evaluation of how well machine learning techniques perform for the specific predictive modeling task. It is easy to apply and understand, making it a popular choice for many modeling scenarios. However, there are cases where this approach might not be suitable. For example, if the dataset is unbalanced and requires multiple configurations or if the model is being used for categorization, the method may not provide optimal results. To evaluate the model's performance, the dataset is divided into training and testing sets using the train-test split function. This separation ensures that the model is tested on unseen data, providing a more realistic measure of its ability to generalize. Typically, 80% of the data is used for training the model, and the remaining 20% is used for testing. This helps in understanding how the model will perform on real-world, unseen data. The test size can be adjusted depending on the dataset's characteristics, but the 80-20 split is commonly used. This method ensures the model's performance is evaluated in a fair and unbiased manner.

a) RANDOM FOREST CLASSIFIER,Random Forest is a powerful ensemble learning algorithm widely used for machine learning classification tasks. It constructs numerous decision trees during the training phase and outputs the class that is the majority vote among the individual tree predictions. Each tree is built using randomly selected data samples, and their individual predictions are combined through a voting mechanism to determine the final output. This method enhances accuracy and reduces overfitting. Additionally, Random Forest provides a reliable estimate of feature importance, helping to understand the significance of various input variables.
b) KNN,K-Nearest Neighbour (K-NN) is among the simplest data mining algorithms, based on the Supervised Learning technique. It works by assuming that similar data points exist in close proximity and classifies new cases by comparing them to the most similar existing ones. K-NN stores all available data and classifies new data points based on similarity measures such as distance. It supports both regression and classification, although it's primarily used for classification tasks. Being a non-parametric algorithm, K-NN doesn't assume any underlying data distribution. It's also called a lazy learner since it delays the learning process until a query is made.
c) DECISION TREE CLASSIFIER,A Decision Tree is a Supervised Learning algorithm used for both classification and regression tasks, though it's more common in classification problems. It is structured like a tree, where internal nodes represent dataset features, branches represent decision rules, and leaf nodes represent outcomes. The model uses two types of nodes: Decision Nodes, which make decisions and have multiple branches, and Leaf Nodes, which represent final outcomes with no further branches. The decision-making process is based on the features in the dataset, making it intuitive and easy to visualize for understanding and interpreting data.

IV. RESULTS AND DISCUSSION

The experimental dataset is split into two parts: training and testing sets. The training set is used to build the classifier model, while the testing set is used to evaluate its performance. In this study, 75% of the data is allocated for training, and the remaining 25% is used for testing. Since the target variable contains two categories, the problem is approached as a binary classification task. Each instance in the dataset is labeled either as 1 (positive) or 0 (negative). The classifier model learns to map these instances to the appropriate class labels.

The performance of different data mining algorithms is evaluated using standard metrics: Precision, Recall, F-measure, and Accuracy. These metrics provide insights into how well the models classify the data. The results are summarized in Table 3 and visually represented for better understanding.

According to the results, the Random Forest classifier achieved an accuracy of 96%, while both the Decision Tree and K-Nearest Neighbour (KNN) classifiers reached 95% accuracy. Among these, the Random Forest algorithm demonstrated the highest performance. This indicates that Random Forest is slightly more effective than the other

models in accurately classifying the instances in this particular study.

Table.1. Performance matrix table

Classifier	Precision	Recall	F-Measure	Accuracy
Random Forest Classifier	0.97	0.96	0.96	0.96
KNN Classifier	0.95	0.95	0.95	0.95
Decision Tree Classifier	0.95	0.95	0.95	0.95

The table presents a comparative analysis of three popular classification algorithms—**Random Forest**, **K-Nearest Neighbors (KNN)**, and **Decision Tree**—based on four key performance metrics: **Precision**, **Recall**, **F-Measure**, and **Accuracy**. These metrics are essential in evaluating the effectiveness of models used for breast cancer prediction.

Precision

Formula:

$$\text{Precision} = \frac{Tp}{TP+FP} \quad (1)$$

- **TP** = True Positives
- **FP** = False Positives

Precision measures the proportion of correctly identified positive observations. Random Forest achieves the highest precision score of **0.97**, indicating fewer false positives compared to KNN and Decision Tree, both of which score **0.95**.

- *Recall*

Formula:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

- **TP** = True Positives
- **FP** = False Positives

Recall, which reflects the model's ability to identify actual positive cases, is also highest for Random Forest at **0.96**, while KNN and Decision Tree are slightly lower at **0.95**.

F-Measure (F1 Score)

Formula:

$$\text{F1} - \text{Score} = \frac{2*Precision*Recall}{Precision+Recall} \quad (3)$$

F-Measure or **F1 Score** balances precision and recall. Random Forest again leads with **0.96**, showing it maintains a good balance between the two. KNN and Decision Tree each score **0.95**, showing consistent but slightly lower performance.

d). Accuracy

Formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

- **TP** = True Positives
- **FP** = False Positives

Accuracy, the overall correctness of the model, is **0.96** for Random Forest, confirming its superior performance in classifying both benign and malignant cases. KNN and Decision Tree classifiers both achieved **0.95** accuracy, indicating solid but not leading performance.

In conclusion, **Random Forest Classifier** outperforms the other models across all evaluation metrics, making it the most effective method among the three for breast cancer prediction in this study.

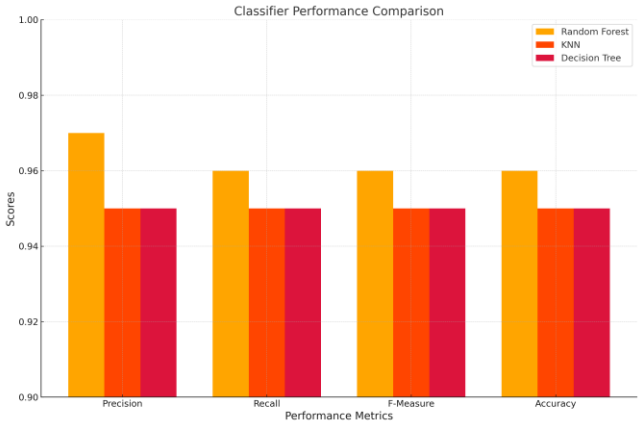

Fig.2. Performance matrix graph

V. CONCLUSION

In conclusion, the application of classification algorithms for breast cancer prediction has demonstrated promising results, offering significant potential to enhance early detection and improve patient outcomes. The algorithms, whether based on KNN, Decision Trees, Random Forest, or

other methods, have shown a commendable level of accuracy in predicting breast cancer outcomes. Their ability to accurately classify benign and malignant cases highlights their reliability as valuable tools in clinical settings.

The predictive model's strong performance indicates its potential to assist in early cancer detection, which is crucial for timely interventions and better prognoses. This could be transformative in clinical practices, aiding healthcare professionals in making informed decisions that ultimately improve patient care. However, while the algorithm shows promise, it is essential to acknowledge its limitations. Issues such as the need for diverse datasets, data quality, and algorithm complexity could impact the model's generalizability and effectiveness in different settings.

Moreover, continuous validation and refinement of the model are necessary to ensure its robustness and accuracy across various patient demographics and healthcare environments. The model's ongoing testing will be critical to its long-term success in real-world applications. Looking ahead, there is great potential for further advancements in predictive modeling for breast cancer diagnosis, as AI and machine learning technologies continue to evolve.

Collaboration among the medical community, researchers, and technology developers will be key to refining these predictive tools. Together, they can create more accurate and effective solutions for breast cancer detection. As these models develop, they could play a crucial role in enhancing early detection, improving treatments, and ultimately leading to better patient outcomes.

REFERENCES

[1] Shweta K and S.Soni (2016), Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection, International Journal of Computer Applications, 133.
[2] G. Sumalatha and S. Archana (2017), A Study on Early Prevention and Detection of Breast Cancer using Data Mining Techniques, International Journal of Innovative Research in Computer and Communication Engineering.
[3] Yeulkar, K.(2017) ,Utilization of Data Mining Techniques for Analysis of Breast Cancer Dataset Using R.
[4] Chaurasia V, Pal S. and Tiwari BB (2018), Prediction of benign and malignant breast cancer using data mining techniques, Journal of Algorithms & Computational Technology.
[5] C. Meera and D. Nalini (2018), Breast cancer prediction system using Data mining methods, International Journal of Pure and Applied Mathematics.
[6] Priyanka Israni (2019), Breast Cancer Diagnosis (BCD) Model Using Machine Learning, International Journal of Innovative Technology and Exploring Engineering (IJITEE).
[7] R. Aavula and R. Bhramaramba (2019), XBPF: An Extensible Breast Cancer Prognosis Framework for Predicting Susceptibility, Recurrence and Survivability, International Journal of Engineering and Advanced Technology (IJEAT).
[8] RA. Sanyour and M. Abdullah (2019), Real time data analysis and visualization for the breast cancer disease, Periodicals of Engineering and Natural Sciences.
[9] Kassani, et al. (2019). Breast cancer diagnosis with transfer learning and global pooling. In 2019 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 519524-). IEEE.
[10] https://www.kaggle.com/datasets/uci ml/breast-cancer-wisconsin-data