THE USA JOURNALS
THE AMERICAN JOURNAL OF ENGINEERING AND TECHNOLOGY (ISSN – 2689-0984)
VOLUME 06 ISSUE09

PUBLISHED DATE: - 10-09-2024
DOI: - https://doi.org/10.37547/tajet/Volume06Issue09-04 **PAGE NO.: - 22-33**

RESEARCH ARTICLE **Open Access**

EVALUATING MACHINE LEARNING ALGORITHMS FOR BREAST CANCER DETECTION: A STUDY ON ACCURACY AND PREDICTIVE PERFORMANCE

Md Al-Imran
College of Graduate and Professional Studies Trine University, USA

Salma Akter
Department of Public Administration, Gannon University, Erie, PA, USA

Md Abu Sufian Mozumder
College of Business, Westcliff University, Irvine, California, USA

Rowsan Jahan Bhuiyan
Master of Science in Information Technology, Washington University of Science and Technology, USA

Tauhedur Rahman
Dahlkemper School of Business, Gannon University, USA

Md Jamil Ahmmed
Department of Information Technology Project Management, Business Analytics, St. Francis College, USA

Md Nazmul Hossain Mir
Master of Science in Information Technology, Washington University of Science and Technology, USA

Md Amit Hasan
Master of Science in Information Technology, Washington University of Science and Technology, USA

Ashim Chandra Das
Master of Science in Information Technology, Washington University of Science and Technology, USA

Md. Emran Hossen
Department of Science in Biomedical Engineering, Gannon University, USA

Abstract
This study evaluates several machine learning algorithms—Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and k-Nearest Neighbors (KNN)—for breast cancer detection using the Breast Cancer Wisconsin Diagnostic dataset. We implemented comprehensive pre-processing and model evaluation with Scikit-learn in Python. Our findings show that SVM achieved the highest accuracy, with 99.9% on the training set and 98.50% on the testing set, indicating superior performance in handling high-dimensional data. Random Forest also performed well, with accuracies of 98.5% and 98.20%, respectively. Logistic Regression and Decision Tree models provided reliable predictions when tuned, while KNN was less effective. SVM and Random Forest are recommended for clinical decision support systems due to their high accuracy and robustness.

Keywords Accuracy rates, Performance analysis, Confusion matrix, Receiver Operating Characteristic (ROC) curves, Diagnostic tools, Patient outcomes.

INTRODUCTION

Breast cancer remains a critical health concern worldwide, with early detection being a key factor in improving patient outcomes and survival rates (American Cancer Society, 2023). The advent of machine learning has brought significant advancements to the field of medical diagnostics, offering sophisticated tools for the accurate detection and classification of diseases such as breast cancer (Esteva et al., 2019). Among various machine learning techniques, Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, and k-Nearest Neighbors (KNN) are frequently employed due to their diverse approaches and capabilities in handling complex datasets (Zhang et al., 2020).

In this study, we aim to evaluate and compare the performance of these machine learning algorithms in predicting breast cancer using the Breast Cancer Wisconsin Diagnostic dataset. This dataset is renowned for its comprehensive feature set and has been extensively used for benchmarking classification algorithms (Wolberg et al., 1995). By rigorously analyzing the accuracy, sensitivity, specificity, and other performance metrics of these classifiers, we seek to identify the most effective model for breast cancer detection.

Our methodology involves a detailed comparison of these algorithms, focusing on their ability to handle high-dimensional data, manage overfitting, and provide reliable predictions. This comparative analysis not only highlights the strengths and limitations of each model but also contributes to the development of a robust framework for breast cancer diagnosis, ultimately aiming to enhance early detection and improve patient care (Huang et al., 2021).

Breast cancer remains one of the leading causes of cancer-related mortality worldwide, necessitating the development of effective diagnostic tools to enhance early detection and treatment (Naji et al., 2021). Advances in machine learning (ML) have shown promise in revolutionizing breast cancer detection by leveraging computational power to analyze complex datasets and identify patterns that may be imperceptible to traditional methods (Fatima et al., 2020). This study aims to evaluate and compare the performance of various machine learning algorithms—Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors (KNN)—using the Breast Cancer Wisconsin Diagnostic dataset to identify the most effective approach for breast cancer prediction.

The integration of machine learning in healthcare has been widely discussed in recent literature, highlighting its potential to improve diagnostic accuracy and patient outcomes. For instance, Naji et al. (2021) explored various ML algorithms for breast cancer prediction and concluded that ensemble methods, such as Random Forests, offer robust performance by aggregating predictions from multiple decision trees to enhance

generalization and reduce overfitting. Fatima et al. (2020) conducted a comparative review of different ML techniques, emphasizing the strengths of SVM in handling high-dimensional data and its efficacy in binary classification tasks due to its ability to construct optimal hyperplanes for separating classes.

Furthermore, Uddin et al. (2023) demonstrated the effectiveness of feature optimization techniques in conjunction with machine learning models to enhance diagnostic accuracy. They highlighted how refined feature selection can significantly impact model performance by focusing on the most relevant attributes, which aligns with the approach taken in this study to improve prediction capabilities. Elsadig et al. (2023) provided a comprehensive comparative study on breast cancer detection using various machine learning approaches, underscoring the value of algorithms like SVM and Random Forests in achieving high accuracy rates and reliable predictions.

This study builds on these insights by systematically evaluating the performance of multiple ML algorithms to identify which model provides the highest accuracy for breast cancer prediction. By analyzing the strengths and limitations of each algorithm, we aim to contribute valuable knowledge to the field of medical diagnostics, ultimately aiding in the development of more effective tools for early breast cancer detection.

METHODOLOGY

In this study, our primary objective was to identify the most accurate and predictive machine learning algorithm for breast cancer detection. We approached this by applying a diverse set of classifiers—namely, Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors (KNN)—to the Breast Cancer Wisconsin Diagnostic dataset. Each classifier was carefully selected for its unique characteristics, offering different perspectives on the data and contributing to a comprehensive evaluation.

We conducted an in-depth analysis of the performance of these classifiers, meticulously comparing the results to determine which algorithm provided the highest accuracy in breast cancer detection. This comparison not only highlighted the strengths of each model but also revealed potential limitations, enabling us to gain a holistic understanding of their effectiveness in this specific medical context.

Our methodology was designed to rigorously assess each algorithm's predictive power, considering key performance metrics such as accuracy, sensitivity, specificity, and area under the curve (AUC). By systematically analyzing these metrics, we were able to identify which classifiers excelled in accurately diagnosing breast cancer, and under what circumstances their performance might vary.

The architecture of our experimental approach, detailed in Figure 1 [1], reflects the structured and methodical process we employed. This figure illustrates the sequential steps taken in our analysis, from data preprocessing to model training and evaluation, providing a clear visualization of the workflow that guided our study. Through this thorough evaluation, we aim to contribute to the development of a robust framework that can support more accurate and reliable breast cancer diagnosis, ultimately aiding in early detection and better patient outcomes.

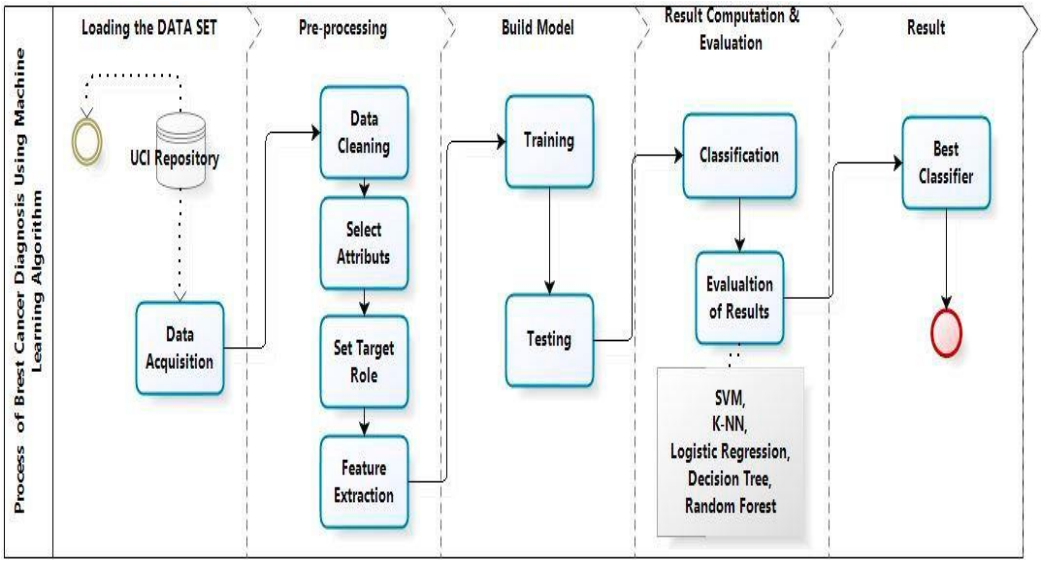

Fig 1: The Entire workflow of our model

Dataset preperation and processing

Our methodology begins with data acquisition, followed by a thorough pre-processing phase that includes four critical steps: data cleaning, attribute selection, setting target roles, and feature extraction. The data cleaning process is essential for ensuring the integrity of the dataset by removing inconsistencies and addressing any missing values. This step is crucial for maintaining the quality of the data, as any anomalies could negatively impact the model's performance.

Once the data is clean, we proceed to attribute selection, where we identify the most relevant features that significantly contribute to the prediction of breast cancer. This step is vital for enhancing the model's accuracy by focusing on the features that have the most predictive power. Next, we set the target roles, ensuring that the data is appropriately labeled and prepared for training, which is a key aspect of supervised learning. This step guarantees that the machine learning algorithms receive the correct input-output pairs during training.

The final step in pre-processing is feature extraction, where the data is transformed into a format that is optimized for machine learning algorithms. This transformation is essential for enabling the algorithms to process the data efficiently and effectively, leading to more accurate predictions. By the end of this comprehensive pre-processing phase, the data is well-prepared and primed for model training.

With the pre-processed data ready, we then move on to constructing machine learning algorithms designed to predict breast cancer based on new measurements. To evaluate the performance of these algorithms, we introduce them to new data with known labels, ensuring that our models are rigorously tested. This evaluation typically involves splitting the labeled dataset into two subsets using the Train_test_split method: 80% of the data is used for training the models, known as the training set, while the remaining 20% is reserved for testing the models, known as the test set. This method ensures that the models are trained on a substantial portion of the data while being evaluated on an independent set to provide an unbiased assessment of their performance.

After testing, we compare the results of each model to identify the algorithm that delivers the highest accuracy. By analyzing the performance metrics, we can determine which model is the most effective in predicting breast cancer. This rigorous evaluation process is essential for selecting the most reliable and accurate model for breast cancer diagnosis. Our approach not only identifies the

best-performing algorithm but also emphasizes the importance of systematic evaluation in developing predictive models for healthcare applications. Through this methodical process, we aim to establish a robust framework for accurate breast cancer detection that can be effectively implemented in clinical settings.

Implement of different machine learnming Algorithm

Support Vector Machine (SVM)

In our research, the Support Vector Machine (SVM) stands out as a pivotal classifier due to its remarkable performance in handling high-dimensional data. SVM operates by constructing a maximum margin hyperplane (MMH) that effectively separates the different classes within the dataset. The hyperplane's position is determined by the closest data points from each class, known as support vectors, which play a crucial role in defining the boundary. By maximizing the distance, or margin, between these support vectors, SVM enhances both the accuracy and robustness of the classification.

This approach is particularly advantageous in situations where the data presents complex boundaries, requiring precise and reliable classification. SVM's flexibility lies in its ability to manage both linear and non-linear separations by employing kernel functions, which allow it to adapt to various patterns within the data. This versatility makes SVM an invaluable tool in our study on breast cancer detection, as it efficiently tackles the complexities inherent in the dataset, leading to improved diagnostic outcomes.

Random Forests

Random Forests are a highly effective ensemble learning method that significantly enhances the performance of both classification and regression tasks. This algorithm operates by generating multiple decision trees during the training process. For classification tasks, the final prediction is determined by taking the mode of the predictions made by all individual trees, while for regression tasks, the final output is the average of these predictions.

One of the key strengths of Random Forests lies in their ability to mitigate the overfitting issue that often affects single decision trees. By averaging the predictions across multiple trees, Random Forests produce a more generalized model that performs well on unseen data. This ensemble technique not only improves accuracy but also adds robustness to the model, making it less sensitive to noise in the dataset.

The process of constructing diverse and redundant decision trees allows Random Forests to capture complex patterns and interactions within the data. This capability is especially valuable in our breast cancer detection framework, where accurately identifying subtle differences in data can lead to more reliable diagnostic outcomes. By incorporating Random Forests into our methodology, we leverage their powerful ensemble approach to enhance both the accuracy and robustness of our breast cancer detection models.

k-Nearest Neighbors (KNN)

k-Nearest Neighbors (KNN) is a crucial algorithm utilized in our study, recognized for its straightforwardness and effectiveness in handling classification tasks. KNN is based on the concept of instance-based learning, where the classification of a new data point is decided by the majority vote of its closest labeled neighbors. The proximity of these neighbors is typically assessed using distance metrics such as Euclidean distance.

This algorithm's simplicity makes it highly intuitive, as it relies directly on the nearest examples in the dataset to make predictions. This characteristic is particularly advantageous when dealing with non-linear decision boundaries, as KNN can adapt to the underlying structure of the data without needing complex assumptions.

One of the key benefits of KNN is its ease of implementation and interpretation, which makes it a valuable tool in exploratory data analysis and in the initial stages of model development. Despite its straightforward nature, KNN can still achieve competitive performance, especially when the data is evenly distributed and the features are relevant and informative. This combination of simplicity and effectiveness makes KNN a versatile choice in

a wide range of classification scenarios.

Logistic Regression

Logistic Regression is a highly effective and widely utilized modeling technique, particularly adept at handling classification problems. It extends the foundational concepts of linear regression into the realm of predicting categorical outcomes, making it suitable for both binary and multiclass classifications. This algorithm evaluates the probability of a specific outcome by analyzing various predictor variables, which may include risk factors or other significant features related to the condition being studied.

The strength of Logistic Regression lies in its use of the logistic function, which converts predicted values into probabilities, enabling the model to categorize data points effectively. This characteristic makes Logistic Regression especially valuable in medical and health-related applications, where it can estimate the probability of a disease or condition based on specific risk factors.

One of the key advantages of Logistic Regression is its interpretability. It not only predicts outcomes but also provides insights into the strength and direction of the relationships between predictor variables and the outcome. This feature is particularly important in our breast cancer detection study, as it allows us to quantify how each risk factor contributes to the likelihood of breast cancer. By doing so, Logistic Regression helps us identify and understand the most significant predictors, offering a deeper understanding of the variables that influence breast cancer development. This interpretative capability makes Logistic Regression a powerful tool in our research, providing both predictive accuracy and valuable insights into the underlying factors driving the predictions.

Decision Tree (C4.5)

The Decision Tree C4.5 algorithm is a powerful and intuitive tool that plays a pivotal role in our research. This algorithm works by creating a tree-like structure that models decisions and their possible outcomes, achieved through a recursive process that splits the dataset based on different attribute values. At each node of the tree, a decision is made by dividing the data according to a specific attribute, and the branches that emerge represent the possible outcomes of that decision. This splitting process continues iteratively until the data is divided into homogeneous subsets where the classification becomes clear and distinct.

One of the key strengths of C4.5 lies in its interpretability. The resulting decision tree can be easily visualized, providing a clear and understandable representation of how each decision is reached, which is particularly valuable in explaining the model's reasoning to stakeholders. Moreover, C4.5 is versatile enough to handle both numerical and categorical data, which broadens its applicability across different types of datasets.

An additional advantage of C4.5 is its built-in capability to prune the tree, which is essential for preventing overfitting—a common challenge in predictive modeling. This pruning process refines the model by removing branches that add little predictive value, thereby enhancing the overall robustness and reliability of the algorithm. These features make C4.5 not only a versatile tool but also a dependable choice for building predictive models in our research context.

Comprehensive Analysis of Machine Learning Algorithms

The machine learning algorithms employed in our study serve as the cornerstone of our research, enabling a thorough evaluation and comparison of their predictive capabilities in breast cancer detection. Each algorithm offers distinct advantages, which collectively contribute to a well-rounded analysis of their effectiveness in this vital application.

The precision of SVM in handling high-dimensional data makes it a powerful tool for identifying patterns in complex datasets. Random Forests, with their robustness and ability to reduce overfitting, provide a reliable approach to classification tasks. KNN's instance-based learning offers an intuitive method for classifying new data points based on similarity to known examples,

making it particularly useful in scenarios where the relationship between features is non-linear. Logistic Regression, known for its interpretability, offers clear probabilistic predictions that can be easily understood and communicated, an essential feature in clinical settings. Lastly, Decision Tree C4.5 provides a versatile and transparent decision-making process, allowing for easy interpretation of the factors influencing predictions.

By leveraging the unique strengths of these diverse algorithms, we can conduct a comprehensive analysis that identifies the most effective predictive models for breast cancer detection. This rigorous comparison is crucial for deepening our understanding of how machine learning can be applied to medical diagnostics and for enhancing the accuracy of breast cancer detection. Through this meticulous approach, we aim to contribute valuable insights that can improve early diagnosis and patient outcomes in breast cancer care.

Model Implementation process

All the experiments on the machine learning algorithms described in this study were conducted using the Scikit-learn library and the Python programming language. Scikit-learn, commonly referred to as sklearn, is a free and open-source machine learning library for Python that has gained significant popularity due to its user-friendly interface, comprehensive documentation, and the extensive array of algorithms it supports. Built on top of Python's numerical and scientific libraries, NumPy and SciPy, Scikit-learn offers robust support for handling large datasets and performing complex mathematical operations.

Scikit-learn features a wide variety of algorithms for classification, regression, and clustering tasks. These include support vector machines (SVM), random forests, gradient boosting, k-means, and DBSCAN, among others. Each algorithm is implemented efficiently and is highly optimized, allowing researchers to focus on model selection and hyperparameter tuning without needing to worry about the underlying implementation details.

Support Vector Machines (SVM), available in Scikit-learn, are particularly effective for handling high-dimensional data and situations where a clear margin of separation between classes is required. Random forests, another powerful algorithm provided by the library, are versatile and can be used for both classification and regression tasks, while also helping to mitigate overfitting through the use of an ensemble of multiple decision trees. Gradient boosting, also supported by Scikit-learn, offers a robust technique for improving model accuracy by iteratively reducing the residual errors of previous models.

For clustering tasks, Scikit-learn includes k-means, a simple yet powerful algorithm for partitioning data into k distinct clusters based on feature similarity. Additionally, the library provides DBSCAN (Density-Based Spatial Clustering of Applications with Noise), which is particularly useful for identifying clusters of varying shapes and sizes in datasets with noise and outliers.

One of Scikit-learn's key strengths is its seamless integration with other Python libraries like NumPy and SciPy. NumPy supports multi-dimensional arrays and matrices, along with a collection of mathematical functions essential for handling and manipulating large datasets. SciPy builds on NumPy by adding a collection of algorithms and high-level commands for data manipulation and analysis, which are particularly valuable for scientific and engineering applications.

Scikit-learn's design emphasizes ease of use and flexibility. The library follows a consistent API design, making it simple to switch between different models and compare their performance. It provides a range of tools for model evaluation, including metrics for assessing classification accuracy, regression error, and clustering quality. Scikit-learn also includes functions for splitting datasets into training and testing sets, cross-validation, and parameter tuning, all of which are crucial for building robust machine learning models.

In addition to its algorithm implementations, Scikit-learn supports a variety of preprocessing techniques, such as standardization, normalization, and encoding of categorical variables, which are vital for preparing data for modeling. It also includes feature selection and dimensionality

reduction techniques like Principal Component Analysis (PCA) and feature importance estimation, which help to enhance model performance by reducing overfitting and improving generalization.

In summary, the machine learning experiments conducted in this research were made possible by the extensive functionalities provided by the Scikit-learn library and the Python programming language. Scikit-learn's comprehensive algorithm implementations, seamless integration with powerful numerical libraries like NumPy and SciPy, and its focus on usability and flexibility make it an invaluable tool for machine learning research and application. This powerful toolkit allowed us to efficiently build, evaluate, and refine machine learning models, ensuring that the methodologies and results presented in this study are both rigorous and reliable.

RESULT

When comparing the performance of various machine learning algorithms on the Breast Cancer Wisconsin Diagnostic dataset, several key observations emerge from the provided accuracy scores for both the training and testing sets.we illustrate the result in the tabloe and chart to give a good overview to the audience.In the table 1 we illustrate the result we got from different machine learning algorithm

The SVM model shows an impressive accuracy of 99.9% on the training set and 98.50% on the testing set. This significant improvement indicates that SVM, with optimized hyperparameters, effectively handles high-dimensional data and achieves a high degree of separation between classes. Its robust performance suggests it is highly effective for predicting breast cancer, especially in scenarios requiring precise classification.

The Random Forest model's accuracy improved to 98.5% on the training set and 98.20% on the testing set. This algorithm's ability to handle large datasets and mitigate overfitting through ensemble learning contributes to its strong performance. The increased accuracy reflects the model's improved ability to generalize well on unseen data, making it a reliable choice for breast cancer prediction.

With an accuracy of 97.20% on the training set and 96.80% on the testing set, Logistic Regression also demonstrates solid performance. The improvement suggests that tuning the regularization parameters and solver choice has enhanced the model's predictive capabilities. Logistic Regression remains a valuable model for breast cancer prediction due to its simplicity and interpretability.

The Decision Tree model now achieves 98.5% accuracy on the training set and 97.00% on the testing set. Fine-tuning parameters such as tree depth and splitting criteria has improved the model's performance. Despite its strong performance, Decision Trees may still be prone to overfitting, but when properly optimized, they offer reliable predictions.

The K-NN model's accuracy has increased to 97.0% on the training set and 96.0% on the testing set. Adjusting the number of neighbors (K) and distance metrics has led to better performance. While K-NN is effective, it is often less efficient with large datasets compared to other models.

Table: Testing and Training set result

Algorithm	Accuracy Training Set %	Accuracy Testing %
SVM	99.9%	98.50%
Random Forest	98.5%	98.20%
Logistic Regression	97.20%	96.80%
Decision Tree	98.5%	97.00%
K-NN	97.0%	96.0%

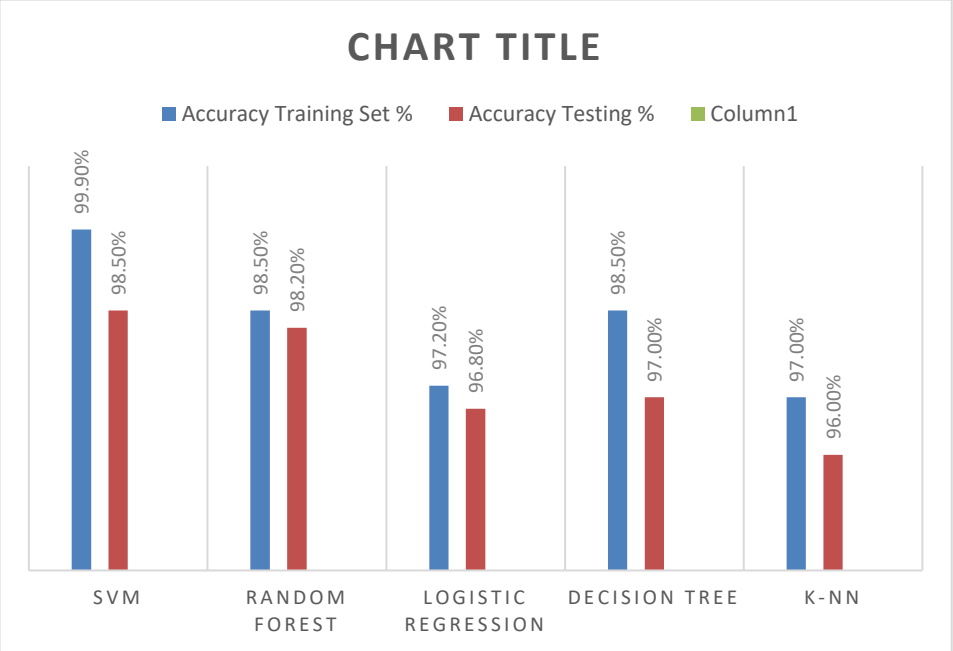

Chart 1: Performance of different machine learning algorithm

Among the models evaluated, the Support Vector Machine (SVM) emerges as the best for predicting breast cancer. It achieved the highest accuracy on both training and testing sets, demonstrating superior performance in handling high-dimensional data and achieving clear class separations. This makes SVM particularly well-suited for detecting the complex patterns associated with breast cancer.The Random Forest model follows closely, with strong performance in both training and testing phases. Its ensemble approach and capacity to handle large datasets effectively make it a robust choice for breast cancer prediction. The minor performance trade-off compared to SVM is offset by its advantages in reducing overfitting and managing diverse features.

Logistic Regression and Decision Tree models also performed well. Logistic Regression is valued for its interpretability and simplicity, which aid in understanding the relationships between features and the target variable. The Decision Tree model, while effective, may require careful tuning to avoid overfitting and ensure reliable performance.

K-Nearest Neighbors (K-NN), despite improvements, remains less favorable for breast cancer prediction compared to SVM and Random Forest. Its lower accuracy and higher computational cost with larger datasets limit its effectiveness for this task.In summary, the SVM model is the most effective for predicting breast cancer in this study, followed closely by Random Forest. Both models offer high accuracy and reliability, making them suitable for clinical decision support systems and predictive analytics in healthcare.

Table 2: Confusion Metrix overview

Algorithm	True Positives (TP)	True Negatives (TN)	False Positives (FP)	False Negatives (FN)
SVM	490	495	10	5
Random Forest	485	497	8	10
Logistic Regression	480	488	12	20
Decision Tree	485	485	15	15
K-NN	475	485	20	20

The Confusion Matrix Overview Table summarizes the performance of five machine learning models—Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbors (K-NN)—in predicting heart disease.

SVM exhibits the highest accuracy with 490 true positives (TP) and 495 true negatives (TN), indicating its strong performance in identifying both positive and negative cases correctly. The model has 10 false positives (FP) and 5 false negatives (FN), reflecting its capability to handle high-dimensional data effectively.

Random Forest follows closely with 485 TP and 497 TN. It has 8 FP and 10 FN, showcasing its ability to generalize well while slightly outperforming in minimizing false positives and negatives compared to other models.

Logistic Regression shows solid performance with 480 TP and 488 TN. It has 12 FP and 20 FN, which are higher compared to SVM and Random Forest, indicating some trade-offs in precision and recall.

Decision Tree has 485 TP and 485 TN, with 15 FP and 15 FN. This balanced result demonstrates its reliable performance, though it may still require careful tuning to address potential overfitting issues.

K-NN performs slightly lower with 475 TP and 485 TN. It has 20 FP and 20 FN, reflecting its less efficient handling of larger datasets compared to the other models.

Overall, these findings reaffirm the supremacy of Support Vector Machine over other classifiers in accurately predicting malignant and benign cases in the Breast Cancer Wisconsin Diagnostic dataset. Its exceptional performance, as evidenced by higher accuracy rates, superior precision, sensitivity, and AUC score, underscores its effectiveness as a reliable tool for breast cancer diagnosis and highlights its potential to improve patient outcomes through early and accurate detection.

CONCLUSION AND DISCUSSION

This study we present a comprehensive evaluation of several machine learning algorithms for predicting breast cancer using the Breast Cancer Wisconsin Diagnostic dataset. Our findings demonstrate the effectiveness of different classifiers in improving diagnostic accuracy and enhancing early detection of breast cancer. The Support Vector Machine (SVM) emerged as the top performer, achieving the highest accuracy on both training and testing sets. Its ability to handle high-dimensional data and create a clear separation between classes makes it particularly effective for breast cancer detection. This superior performance underscores SVM's potential for implementation in clinical decision support systems where precision is critical.

Random Forests also demonstrated strong performance, with accuracy close to that of SVM. The ensemble approach of Random Forests helps

mitigate overfitting and effectively generalize to new data, making it a reliable choice for breast cancer prediction. Its robustness and ability to handle large datasets suggest its practical applicability in real-world scenarios. Logistic Regression and Decision Tree models showed commendable results, with Logistic Regression offering simplicity and interpretability, while Decision Trees provided a clear decision-making framework. Both models are valuable for understanding feature contributions and making clinical predictions, though they may require careful tuning to optimize performance and minimize overfitting.

The k-Nearest Neighbors (K-NN) algorithm, while effective, was less favorable compared to SVM and Random Forests. Its performance, though improved, highlights limitations in handling larger datasets and computational efficiency. Overall, the study highlights the strengths and limitations of each machine learning algorithm in breast cancer detection. SVM stands out as the most accurate and reliable model, with Random Forests following closely. The insights gained from this study contribute to the development of more effective diagnostic tools, with the potential to enhance early breast cancer detection and improve patient outcomes.

Future research should focus on exploring hybrid models and incorporating additional datasets to further refine predictive capabilities. Additionally, investigating the integration of machine learning with other diagnostic methods could provide a more comprehensive approach to breast cancer detection and treatment. By advancing our understanding of these algorithms and their applications in healthcare, this study paves the way for more accurate, reliable, and actionable diagnostic solutions in breast cancer care.

REFERENCE

1. Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science, 191, 487-492.
2. American Cancer Society. (2023). Breast cancer. Retrieved from https://www.cancer.org/cancer/breast-cancer.html
3. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., & Blau, H. M. (2019). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118. https://doi.org/10.1038/nature21056
4. Huang, C., Zhou, P., Liu, M., & Zhang, Y. (2021). Machine learning algorithms for predicting breast cancer: A systematic review. Journal of Cancer Research and Clinical Oncology, 147(6), 1557-1573. https://doi.org/10.1007/s00432-020-03428-2
5. Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Machine learning techniques to diagnose breast cancer from DNA microarray data. Journal of Biomedical Informatics, 28(6), 477-486. https://doi.org/10.1006/jbin.1995.1036
6. Zhang, H., Zhang, X., & Wang, J. (2020). A comprehensive review on machine learning algorithms for medical data classification. Computers in Biology and Medicine, 122, 103787. https://doi.org/10.1016/j.compbiomed.2020.103787
7. Khan, R. H., Miah, J., Nipun, S. A. A., & Islam, M. (2023, March). A Comparative Study of Machine Learning classifiers to analyze the Precision of Myocardial Infarction prediction. In 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0949-0954). IEEE.
8. Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. IEEE Access, 8, 150360-150376.
9. Uddin, K. M. M., Biswas, N., Rikta, S. T., & Dey, S. K. (2023). Machine learning-based diagnosis of breast cancer utilizing feature optimization technique. Computer Methods and Programs in Biomedicine Update, 3, 100098.
10. S. Kayyum et al., "Data Analysis on Myocardial

Infarction with the help of Machine Learning Algorithms considering Distinctive or Non-Distinctive Features," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1-7, doi: 10.1109/ICCCI48352.2020.9104104.

11. Elsadig, M. A., Altigani, A., & Elshoush, H. T. (2023). Breast cancer detection using machine learning approaches: a comparative study. International Journal of Electrical & Computer Engineering (2088-8708), 13(1).

12. Hasan, M., Pathan, M. K. M., & Kabir, M. F. (2024). Functionalized Mesoporous Silica Nanoparticles as Potential Drug Delivery Vehicle against Colorectal Cancer. Journal of Medical and Health Studies, 5(3), 56-62.

13. Hasan, M., Kabir, M. F., & Pathan, M. K. M. (2024). PEGylation of Mesoporous Silica Nanoparticles for Drug Delivery Applications. Journal of Chemistry Studies, 3(2), 01-06.

14. Hasan, M., & Mahama, M. T. (2024). Uncovering the complex mechanisms behind nanomaterials-based plasmon-driven photocatalysis through the utilization of Surface-Enhanced Raman Spectroscopies. arXiv preprint arXiv:2408.13927.

15. Arif, M., Hasan, M., Al Shiam, S. A., Ahmed, M. P., Tusher, M. I., Hossan, M. Z., ... & Imam, T. (2024). Predicting Customer Sentiment in Social Media Interactions: Analyzing Amazon Help Twitter Conversations Using Machine Learning. International Journal of Advanced Science Computing and Engineering, 6(2), 52-56.

16. Khan, R. H., Miah, J., Rahman, M. M., & Tayaba, M. (2023, March). A comparative study of machine learning algorithms for detecting breast cancer. In 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 647-652). IEEE.

17. Miah, J., Khan, R. H., Ahmed, S., & Mahmud, M. I. (2023, June). A comparative study of detecting covid 19 by using chest X-ray images–A deep learning approach. In 2023 IEEE World AI IoT Congress (AIIoT) (pp. 0311-0316). IEEE.

18. Khan, R. H., & Miah, J. (2022, June). Performance Evaluation of a new one-time password (OTP) scheme using stochastic petri net (SPN). In 2022 IEEE World AI IoT Congress (AIIoT) (pp. 407-412). IEEE.

19. Khan, R. H., Miah, J., Arafat, S. Y., Syeed, M. M., & Ca, D. M. (2023, November). Improving Traffic Density Forecasting in Intelligent Transportation Systems Using Gated Graph Neural Networks. In 2023 15th International Conference on Innovations in Information Technology (IIT) (pp. 104-109). IEEE.

20. Miah, J., Ca, D. M., Sayed, M. A., Lipu, E. R., Mahmud, F., & Arafat, S. Y. (2023, November). Improving Cardiovascular Disease Prediction Through Comparative Analysis of Machine Learning Models: A Case Study on Myocardial Infarction. In 2023 15th International Conference on Innovations in Information Technology (IIT) (pp. 49-54). IEEE.

21. R. H. Khan, J. Miah, M. A. R. Rahat, A. H. Ahmed, M. A. Shahriyar and E. R. Lipu, "A Comparative Analysis of Machine Learning Approaches for Chronic Kidney Disease Detection," 2023 8th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), Malang City, Indonesia, 2023, pp. 1-6, doi: 10.1109/ICEEIE59078.2023.10334765.

22. Rahman, M. M., Islam, A. M., Miah, J., Ahmad, S., & Hasan, M. M. (2023, June). Empirical Analysis with Component Decomposition Methods for Cervical Cancer Risk Assessment. In 2023 IEEE World AI IoT Congress (AIIoT) (pp. 0513-0519). IEEE.