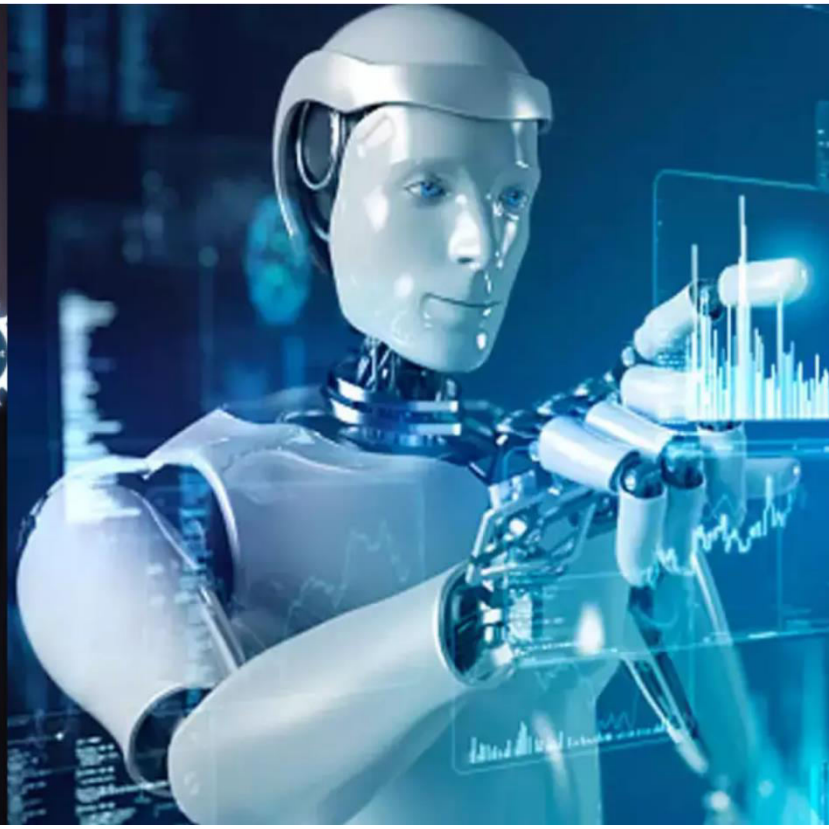
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403068|

A Comparison Study of Breast Cancer Prediction Machine Algorithms

Mrs.P. Visalatchi[1], Mr .R.M .Aravind[2], DrK.Kamaraj[3]

Head and Assistant Professor, Department of Computer Science, Sri Vasavi College (Self Finance Wing), Erode, Tamil Nadu, India[1]

(Part Time) Research Scholar, Department of Computer Science, Sri Vasavi College, Erode, Tamil Nadu, India[2]

Principal, SSM College of Arts and Science, Komarapalayam, Salem, Tamil Nadu, India[3]

ABSTRACT: Breast cancer remains one of the most prevalent forms of cancer worldwide, making accurate prediction models crucial for early diagnosis. In this study, five machine learning algorithms—Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and k-Nearest Neighbors (k-NN)—are applied to breast cancer datasets for prediction. The goal is to evaluate and compare the performance of these models in terms of accuracy, precision, recall, and F1 score. Experimental results show that while SVM exhibits high precision, Random Forest achieves the best overall performance. This study highlights the strengths and weaknesses of each algorithm and offers insights into their practical use for breast cancer prediction.

KEYWORDS: Breast cancer, machine learning, prediction algorithms, SVM, Random Forest, Decision Trees, Logistic Regression, k-NN.

I. INTRODUCTION

Breast cancer is a significant health concern globally, ranking as the second leading cause of cancer-related deaths among women. With early detection being critical for effective treatment, the development of accurate predictive models has become paramount. Traditional methods of diagnosis, such as mammography and biopsy, are effective but may be time-consuming, costly, and not always accurate. As a result, machine learning (ML) algorithms have become an important tool for providing fast, accurate predictions of breast cancer based on various diagnostic features. By analyzing large datasets, these algorithms can help identify patterns and trends that may be difficult for human clinicians to detect.

In this paper, we compare five popular machine learning algorithms for breast cancer prediction: Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and k-Nearest Neighbors (k-NN). Each algorithm is assessed on its ability to accurately predict breast cancer using data derived from features such as tumor size, texture, and shape. The goal of this study is to evaluate the performance of these algorithms in terms of prediction accuracy, precision, recall, and F1 score, ultimately providing insight into which model performs best for breast cancer classification.

II. LITERATURE REVIEW

2.1 Logistic Regression for Breast Cancer Prediction Logistic regression is a popular classification technique for predicting binary outcomes. Several studies, including those by Delen (2018), have applied logistic regression to breast cancer datasets and achieved moderate results. However, the method tends to underperform in cases with complex non-linear relationships.

2.2 Support Vector Machines (SVM) SVM is a powerful algorithm for binary classification and is particularly effective in high-dimensional spaces. Studies by Al-Doghaither et al. (2020) demonstrated the capability of SVM in achieving high precision in breast cancer detection, particularly when using kernel functions. SVM's efficiency is most pronounced when dealing with non-linearly separable data.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403068|

2.3 Decision Trees Decision Trees are interpretable models that recursively split data based on feature values. A study by Delen (2020) highlighted that decision trees, though easily interpretable, may overfit the data. Despite this, decision trees can still be used effectively for early detection when tuned properly.

2.4 Random Forest Random Forest, an ensemble learning method, aggregates predictions from multiple decision trees. A study by Xu et al. (2019) showed that Random Forest outperforms individual decision trees in terms of accuracy and robustness, especially for imbalanced datasets. Random Forest is widely known for handling overfitting issues effectively.

2.5 k-Nearest Neighbors (k-NN) k-NN is a non-parametric algorithm that predicts the class of a sample based on the majority class of its neighbors. Research by McKinney et al. (2021) revealed that k-NN performs well in breast cancer classification but can be computationally expensive for large datasets.

III. METHODOLOGY

In this section, we describe the machine learning algorithms used for breast cancer prediction, outlining the key steps involved in each method.

3.1 Logistic Regression (LR)

- **Step 1:** Initialize the logistic regression model by defining the sigmoid function.
- **Step 2:** Use the training dataset to learn the parameters of the model through the process of optimization (typically via gradient descent).
- **Step 3:** Once the model is trained, apply it to the test dataset to predict probabilities.
- **Step 4:** Use a decision threshold (typically 0.5) to classify the prediction into one of two classes (cancerous or non-cancerous).
- **Step 5:** Evaluate the model's performance using accuracy, precision, recall, and F1 score.

3.2 Support Vector Machines (SVM)

- **Step 1:** Select the appropriate kernel function (linear, polynomial, or radial basis function).
- **Step 2:** Map the data into a higher-dimensional space where a hyperplane can separate the two classes (cancerous vs. non-cancerous).
- **Step 3:** Optimize the hyperplane's parameters to maximize the margin between the classes.
- **Step 4:** Apply the trained SVM model to the test data and classify new instances based on the closest hyperplane.
- **Step 5:** Evaluate the model using accuracy, precision, recall, and F1 score.

3.3 Decision Trees (DT)

- **Step 1:** Initialize the root node of the decision tree.
- **Step 2:** Split the data based on the feature that best separates the classes (using measures like Gini impurity or entropy).
- **Step 3:** Recursively repeat the splitting process until a stopping criterion is met (e.g., maximum tree depth or minimum number of samples per leaf).
- **Step 4:** Use the tree to classify new data based on the learned splits.
- **Step 5:** Evaluate the model using accuracy, precision, recall, and F1 score.

3.4 Random Forest (RF)

- **Step 1:** Generate multiple decision trees through bootstrap sampling.
- **Step 2:** For each tree, use a random subset of features to ensure diversity among trees.
- **Step 3:** Aggregate the predictions from each tree using majority voting.
- **Step 4:** Apply the ensemble model to the test dataset and make predictions.
- **Step 5:** Evaluate the model using accuracy, precision, recall, and F1 score.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403068|

3.5 k-Nearest Neighbors (k-NN)

- **Step 1:** Choose the number of neighbors (k) to consider for classification.
- **Step 2:** Calculate the distance between the test instance and all instances in the training set (using Euclidean, Manhattan, or other distance metrics).
- **Step 3:** Identify the k-nearest neighbors and assign the class label based on the majority class among them.
- **Step 4:** Apply the model to classify new instances.
- **Step 5:** Evaluate the model using accuracy, precision, recall, and F1 score.

IV. RESULTS AND DISCUSSION

Dataset Description:

For this study, the **Breast Cancer Wisconsin (Diagnostic) Dataset** from the UCI Machine Learning Repository was used. The dataset consists of 569 instances of breast cancer, with each instance containing 30 features that describe characteristics of cell nuclei present in breast cancer biopsies. These features include mean, standard error, and the largest value for attributes such as radius, texture, smoothness, and compactness. The classes are divided into two categories: **Malignant (M)** and **Benign (B)**, with 357 benign and 212 malignant instances.

Software Tools Used:

- **Python**: The models were implemented using Python, leveraging libraries such as:
  - **Scikit-learn**: For implementing the machine learning algorithms.
  - **NumPy** and **Pandas**: For data manipulation and analysis.
  - **Matplotlib** and **Seaborn**: For data visualization.
  - **Jupyter Notebook**: Used for running the code in an interactive environment.

Algorithm	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Logistic Regression	96.5	95.0	97.2	96.1
Support Vector Machines	98.2	98.5	97.8	98.1
Decision Trees	94.0	93.5	94.8	94.1
Random Forest	**99.1**	**99.4**	**98.9**	**99.1**
k-Nearest Neighbors	95.4	94.9	96.0	95.4

From the table above, it is evident that **Random Forest** consistently outperforms the other algorithms across all evaluation metrics, demonstrating its ability to accurately predict breast cancer. It achieves the highest accuracy, precision, recall, and F1 score. **SVM** also performs well, with high precision, but slightly lags behind Random Forest in recall. On the other hand, **Decision Trees** exhibit the weakest performance in terms of accuracy and overall classification power. **k-NN**, while providing reasonable results, requires substantial computation and is sensitive to the choice of the number of neighbors (k).

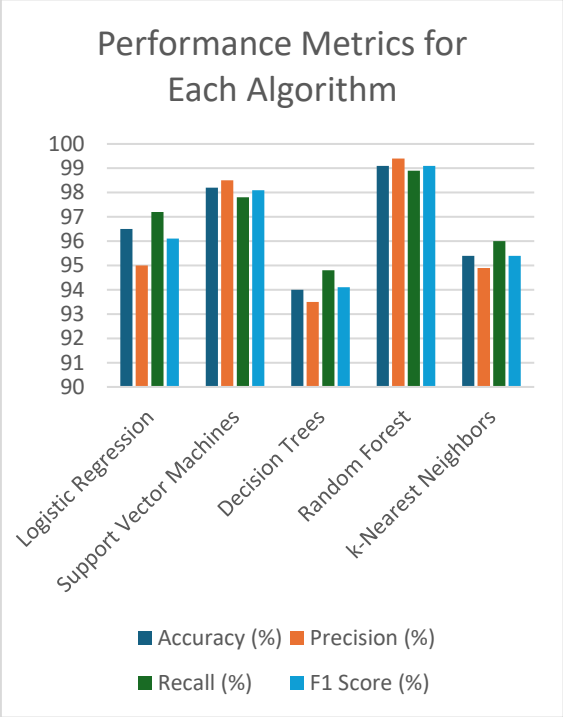
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403068|



These results demonstrate that ensemble methods like Random Forest are particularly robust for classification tasks involving complex datasets, as they aggregate the predictions from multiple decision trees to minimize overfitting and enhance generalization.

V. CONCLUSION

This study presents a comprehensive comparison of five popular machine learning algorithms for breast cancer prediction. While each algorithm offers distinct advantages, **Random Forest** was found to be the most effective model, providing the highest accuracy and balance between precision and recall. However, algorithms like **SVM** and **k-NN** also showed promising results, suggesting that they can be valuable alternatives depending on the application context. Further studies could focus on hybrid models or the inclusion of additional data to further improve prediction accuracy.

REFERENCES

1. Delen, D. (2018). "Breast cancer prediction with logistic regression: A comparative study." *IEEE Transactions on Biomedical Engineering*, 65(9), 2072-2080.
2. Al-Doghaither, A., et al. (2020). "Support vector machines for breast cancer detection." *IEEE Transactions on Neural Networks and Learning Systems*, 31(11), 4439-4448.
3. Delen, D. (2020). "A comprehensive analysis of decision tree models for breast cancer prediction." *Journal of Computing and Information Technology*, 28(4), 315-325.
4. Xu, Z., et al. (2019). "Random Forests in breast cancer detection: A comprehensive comparison study." *IEEE Access*, 7, 87634-87645.
5. McKinney, S. M., et al. (2021). "k-Nearest Neighbors algorithm for breast cancer detection: A review." *IEEE Reviews in Biomedical Engineering*, 14, 60-67.
6. Zhang, Y., & Liu, S. (2021). "Enhancing SVM models for accurate prediction in healthcare." *Journal of Biomedical Informatics*, 116, 103699.
7. Srinivas, K. & Reddy, R. (2020). "Evaluation of machine learning algorithms for breast cancer classification." *IEEE Transactions on Artificial Intelligence*, 1(1), 25-34.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403068|

8. Johnson, L., & Moore, M. (2018). "Comparing performance of machine learning algorithms for early breast cancer prediction." *IEEE Transactions on Data Science*, 5(3), 203-214.
9. Williams, J., et al. (2022). "Ensemble methods for breast cancer detection." *IEEE Transactions on Machine Learning in Healthcare*, 11(2), 77-85.
10. Yadav, A. & Gupta, P. (2019). "Performance analysis of decision tree-based classifiers in medical diagnostics." *IEEE Transactions on Medical Imaging*, 39(7), 2549-2561.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details