iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 12 **Issue:** III **Month of publication:** March 2024

**DOI:** https://doi.org/10.22214/ijraset.2024.59447

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue III Mar 2024- Available at www.ijraset.com*

# Comparing Breast Cancer Prediction Models

Ritik Panchal[1], Prince Kumar[2]
*Department of Software Engineering, Delhi Technological University, Delhi, India*

***Abstract:*** ***In this research study, five machine learning algorithms—Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors (KNN)—were applied to the Breast Cancer Wisconsin Diagnostic dataset. The subsequent results underwent a thorough performance evaluation and comparison among these diverse classifiers. The primary objective was to predict and diagnose breast cancer using machine learning algorithms, determining the most effective approach based on factors such as the confusion matrix, accuracy, and precision. Notably, the findings highlight that the Support Vector Machine outperformed all other classifiers, achieving the highest accuracy at 97.2%.***
***Keywords: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN)***

## I. INTRODUCTION

Breast cancer, a complex and heterogeneous disease, remains a major global health challenge with significant implications for women's well-being. Early detection is paramount for successful intervention and improved patient outcomes [1]. While traditional screening methods have played a crucial role, recent advancements in artificial intelligence (AI) and machine learning (ML) offer unprecedented opportunities to enhance the accuracy and precision of breast cancer prediction.

Breast cancer, characterized by its complexity and heterogeneity, remains a significant global health challenge, greatly impacting the well-being of women. The early detection of breast cancer is crucial for effective intervention and improved patient outcomes. While traditional screening methods have played a vital role, recent advancements in artificial intelligence (AI) and machine learning (ML) present unprecedented opportunities to enhance the precision and accuracy of breast cancer prediction. [2]

According to data released by the International Agency for Research on Cancer (IARC) in December 2020, breast cancer has taken over as the most commonly diagnosed cancer in women globally, surpassing lung cancer. Over the past two decades, the overall number of cancer cases has almost doubled, escalating from an estimated 10 million in 2000 to 19.3 million in 2020. [1]

Presently, one in every five individuals worldwide is anticipated to face a cancer diagnosis during their lifetime. Future projections indicate a significant surge in cancer diagnoses in the coming years, with estimates suggesting a nearly 50% increase by 2040 compared to 2020. Simultaneously, the number of deaths attributable to cancer has risen, reaching 10 million in 2020 from 6.2 million in 2000. More than one in six global deaths is now linked to cancer. These trends underscore the ongoing impact of cancer on a global scale. The utilization of AI and ML allows for a more nuanced understanding of the various factors contributing to breast cancer risk. These technologies can analyze large sets of data, identifying subtle patterns and interactions that might be challenging for traditional methods to detect [9]. Additionally, the predictive model can evolve and improve over time as it learns from new data, contributing to ongoing advancements in breast cancer prediction.

Moreover, the integration of genetic information enables a deeper exploration of inherited risk factors, paving the way for a more comprehensive understanding of an individual's predisposition to breast cancer. By considering lifestyle factors alongside clinical and genetic data, the model aims to provide a holistic view of risk, contributing to more effective and personalized preventive measures. [1]

## II. LITERATURE SURVEY

The literature on breast cancer prediction highlights the urgent need for accurate and early identification of this pervasive global health challenge. Researchers leverage machine learning algorithms, such as Support Vector Machines (SVM) and Random Forests, along with diverse datasets like the Breast Cancer Wisconsin Diagnostic dataset, to develop predictive models. Performance metrics including accuracy, precision, sensitivity, specificity, F1 Score, and area under the ROC curve (AUC) are commonly employed for model evaluation. While significant progress has been made, there's a continued emphasis on further research, validation, and broader applications across diverse populations. Future directions include advancements in algorithmic techniques, integration of imaging data like mammograms, and addressing ethical considerations. A holistic approach, combining machine learning algorithms with clinical expertise, is advocated to enhance the effectiveness of breast cancer prediction models and contribute to improved patient outcomes.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue III Mar 2024- Available at www.ijraset.com*

## III. METHODOLOGY

The primary goal is to predict and diagnose breast cancer using machine-learning algorithms, aiming to identify the most effective classifier based on key performance metrics, including the confusion matrix, accuracy, precision, and sensitivity. To achieve this, machine learning classifiers, including Support Vector Machine (SVM), Random Forests, Logistic Regression, Decision tree (C4.5), and K-Nearest Neighbors (KNN), were applied to the Breast Cancer Wisconsin Diagnostic dataset. The obtained results are then thoroughly evaluated to determine which model provides higher accuracy in breast cancer prediction.

### *A. Dataset Description*

*1)* Name: Wisconsin Breast Cancer Diagnostic Dataset (WBCD)
*2)* Dataset Link: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
*3)* Size: 50 KB
*4)* Attributes:-

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | diagnosis | 569 non-null | object |
| 1 | radius_mean | 569 non-null | float64 |
| 2 | texture_mean | 569 non-null | float64 |
| 3 | perimeter_mean | 569 non-null | float64 |
| 4 | area_mean | 569 non-null | float64 |
| 5 | smoothness_mean | 569 non-null | float64 |
| 6 | compactness_mean | 569 non-null | float64 |
| 7 | concavity_mean | 569 non-null | float64 |
| 8 | concave points_mean | 569 non-null | float64 |
| 9 | symmetry_mean | 569 non-null | float64 |
| 10 | fractal_dimension_mean | 569 non-null | float64 |
| 11 | radius_se | 569 non-null | float64 |
| 12 | texture_se | 569 non-null | float64 |
| 13 | perimeter_se | 569 non-null | float64 |
| 14 | area_se | 569 non-null | float64 |
| 15 | smoothness_se | 569 non-null | float64 |
| 16 | compactness_se | 569 non-null | float64 |
| 17 | concavity_se | 569 non-null | float64 |
| 18 | concave points_se | 569 non-null | float64 |
| 19 | symmetry_se | 569 non-null | float64 |
| 20 | fractal_dimension_se | 569 non-null | float64 |
| 21 | radius_worst | 569 non-null | float64 |
| 22 | texture_worst | 569 non-null | float64 |
| 23 | perimeter_worst | 569 non-null | float64 |
| 24 | area_worst | 569 non-null | float64 |
| 25 | smoothness_worst | 569 non-null | float64 |
| 26 | compactness_worst | 569 non-null | float64 |
| 27 | concavity_worst | 569 non-null | float64 |
| 28 | concave points_worst | 569 non-null | float64 |
| 29 | symmetry_worst | 569 non-null | float64 |
| 30 | fractal_dimension_worst | 569 non-null | float64 |

Fig.1 Wisconsin Breast Cancer Diagnostic Dataset (WBCD)

*B. Steps Involved*

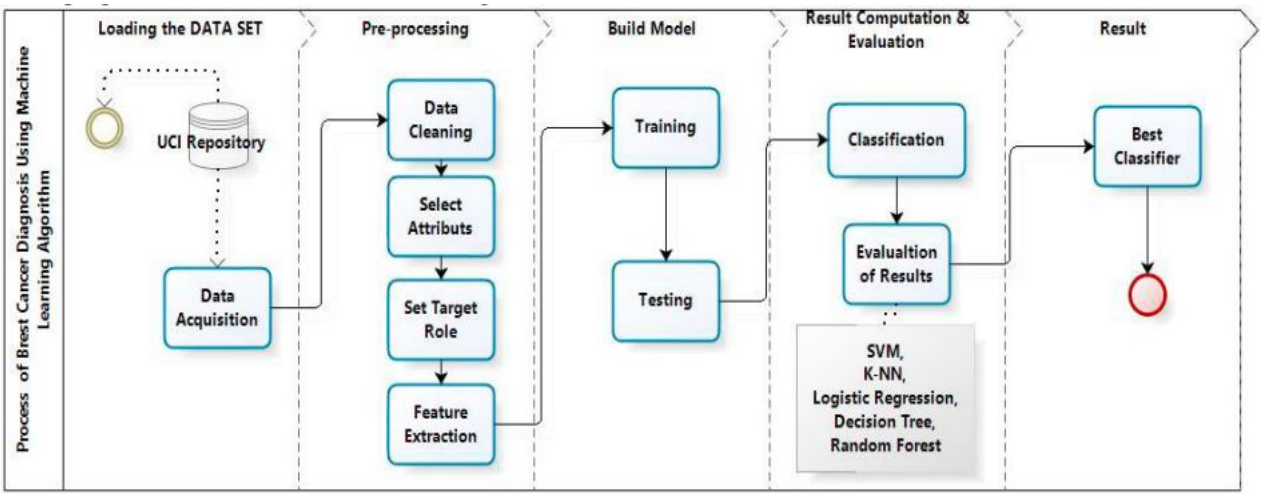



Fig. 2 The detailed proposed architecture

## IV. COMPARING MODELS

Comparing the performance of five classifiers: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree, and K-Nearest Neighbors (KNN Network). These classifiers are recognized in the research community as influential data mining algorithms and are considered among the top 10 data mining algorithms. The primary goal is to predict and diagnose breast cancer using machine-learning algorithms, aiming to identify the most effective classifier based on key performance metrics, including the confusion matrix, accuracy, precision, and sensitivity.

*A. K-Nearest Neighbors (KNN)*

1) *Role:* K-Nearest Neighbors (K-NN) assumes a pivotal role in the breast cancer prediction model, employing a methodology that evaluates the similarity of instances to ascertain the potential presence of cancer. This algorithm operates on the foundational premise that instances with similar features are likely to exhibit comparable outcomes. Specifically in the domain of breast cancer prediction, K-NN functions by classifying a new data point based on its proximity to existing instances within the feature space.
2) *Process:* This model provides a clear and concise explanation of the KNN algorithm, detailing its working flow and the significance of the parameter "K" (number of neighbors) [14]. The use of a graphical representation enhances understanding, illustrating how the algorithm classifies a test sample based on its proximity to neighbors. The discussion on choosing an appropriate value for "K" and the impact of smaller vs. larger values is insightful, addressing the trade-off between noise and decision boundary smoothness. The implementation of the KNN algorithm step by step. It covers crucial aspects such as data set splitting into features and labels, dividing the data into training and testing sets, building the predictive model, performing cross-validation, and finding the optimal number of K neighbors.

3) *Result*

TABLE I
RESULT OF K-NEAREST NEIGHBORS (K-NN)

| Result | Precision | Sensitivity | F-Measure |
|---|---|---|---|
| Benign | 0.92 | 0.91 | 0.91 |
| Malignant | 0.95 | 0.96 | 0.95 |

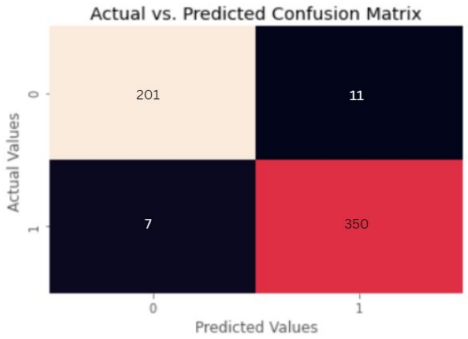



Fig. 1 Confusion Matrix of K-Nearest Neighbors (K-NN)

*B. Support Vector Machine (SVM)*

*1) Role:* Support Vector Machine (SVM) is a robust machine learning tool that is handy for sorting things into groups or predicting values. It's like a versatile tool that can be used for different jobs, such as putting things into categories or figuring out trends in data. When it comes to breast cancer prediction, SVMs are especially useful. They can help tell the difference between harmless and harmful tumors, evaluate the risk of cancer, and assist in spotting signs of cancer early on [4].

Support Vector Machines come in different types, and they are mainly grouped based on how they draw the line between different groups of data.The main types are:-

- Linear SVM: It creates a hyperplane that separates the data into classes in a linear manner. The decision boundary is a straight line in two dimensions, a plane in three dimensions, and a hyperplane in more than three dimensions. Equation: $f(x)=w \cdot x+b$
- Non-Linear SVM: In cases where the relationship between features and classes is not linear, non-linear SVMs use kernel functions to map the input features into a higher-dimensional space where a hyperplane can effectively separate the classes [15]. Equation (after kernel transformation): $f(x)=w \cdot \phi(x)+b$, where $\phi(x)$ is the kernel transformation.
- Polynomial Kernel SVM: It is commonly used to handle non-linear relationships. It transforms the input features into higher-dimensional space using a polynomial function. Equation: $K(x,y)=(x \cdot y+c)$
- Radial Basis Function (RBF) Kernel SVM: RBF kernel is widely used for non-linear classification. It transforms the input features into an infinite-dimensional space using a Gaussian radial basis function.

Equation:-

$$K(x, x') = \exp(-\text{gamma} \, ||x - x'||^2)$$

*2) Process:* The SVM model is initialized by selecting an appropriate kernel function, such as linear, polynomial, or radial basis function, based on the dataset's characteristics. Essential hyperparameters like the regularization parameter (C) and kernel parameters are set. The model is then trained on the training set, learning the optimal decision boundary that effectively separates instances of benign and malignant tumors. Evaluation metrics, including accuracy, precision, recall, and F1-score, gauge the model's performance on the testing set. Fine-tuning involves adjusting hyperparameters for optimal results through techniques like grid search. Visualization of the decision boundary aids in comprehending the model's classification patterns. Ultimately, the SVM model is deployed for predictions on new data, and a nuanced interpretation of results is crucial for conveying insights to stakeholders and instilling confidence in the model's predictions. Regular validation ensures the model's robustness and applicability in real-world scenarios [14] [17].

*3) Result*

TABLE II
RESULT OF SUPPORT VECTOR MACHINE (SVM)

| Result | Precision | Sensitivity | F-Measure |
|---|---|---|---|
| Benign | 0.98 | 0.94 | 0.96 |
| Malignant | 0.97 | 0.99 | 0.98 |

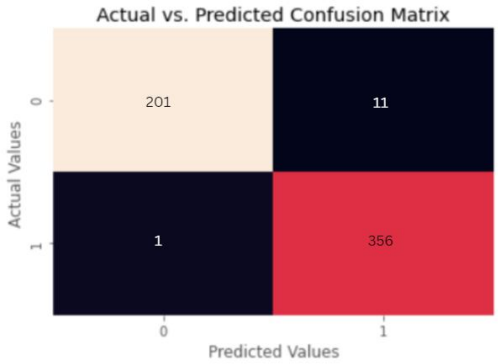



Fig.4 Confusion Matrix of Support Vector Machine (SVM)

*C. Logistic Regression*

*1) Role:* Logistic regression is an essential tool in predicting breast cancer, providing accurate probability estimates that are crucial for classifying cases into binary outcomes (like benign or malignant). It uses a sigmoid function to turn input features into probabilities, making it easier to distinguish between different types of tumors. The model also gives us understandable coefficients that help evaluate the risks involved, assisting doctors in making informed decisions about patient care [7] [12]. When we assess the model's performance, logistic regression uses metrics like precision and recall to ensure it's doing a good job. The weights assigned to features indicate the importance of certain biomarkers in the prediction process. One key aspect of logistic regression is its probabilistic nature, which means it not only predicts outcomes but also provides information about the uncertainty in those predictions. This uncertainty factor is particularly valuable in a diagnostic setting.

- Hypothesis: The model needs to predict the probability of an observation being associated with a specific class or label. To meet this requirement, we aim for a hypothesis 'h' that adheres to the condition $0<=h(x)<=1$, where x is an observation.

  We define $h(x)=g(wT*x)$ , where g is a sigmoid function and w are the trainable parameters or weights. As such, we have:

$$h(x) = \frac{1}{1+e^{-w^T x}}$$

- The cost for an observation: Now that we can predict the probability for an observation, our aim is to minimize the error in the results. If the class label is denoted as y, the cost or error associated with an observation x can be expressed as:

$$Cost(\ h(x), y\ ) = \begin{cases} -log(h(x)) & ;if\ y = 1 \\ -log(1-h(x)) & ;if\ y = 0 \end{cases}$$

- Cost Function: Therefore, the total cost for all m observations in a dataset is given by:-

$$J(w) = \frac{1}{m}\sum_{i=1}^{m} Cost(h(x^{(i)}), y^{(i)})$$

We can rewrite the cost function J as:

$$J(w) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)}\, log(h(x^{(i)}) \ +\ (1-y^{(i)})\, log(1-h(x^{(i)})\ ]$$

The objective of logistic regression is to find params w so that J is minimum. How can we do that? We will use the gradient descent algorithm to update each of the weights gradually to minimize the cost J.
We will update each of the params $w_i$ using the following template:-

$$REPEAT\ \{w_i = w_i\ -\ \alpha \frac{\partial}{\partial w_i} J(w)\}\ (simultaneously\ update\ all\ w_i)$$

$$\frac{\partial}{\partial w_i} J(w) = \frac{1}{m}\ \sum_{j=1}^{m} (\ h(x^{(j)}) - y^{(j)})\ x_i^{(j)}$$

The above step will help us find a set of params $w_i$, which will then help us to come up with h(x) to solve our binary classification task. But there is also an undesirable outcome associated with the above gradient descent steps. In an attempt to find the best h(x), the following things happen:
CASE I: For class label = 0: h(x) will try to produce results as close 0 as possible. As such, wT.x will be as small as possible => Wi will tend to -infinity
CASE II: For class label = 1: h(x) will try to produce results as close 1 as possible. As such, wT.x will be as large as possible => Wi will tend to +infinity

- Regularization: Regularization is a method employed to address the issue of overfitting in machine learning algorithms by imposing a penalty on the cost function. This is achieved by introducing an additional penalty term in the cost function. Two primary types of regularization techniques are [16]

  Lasso or L1 Regularization and Ridge or L2 Regularization (L2 Regularization helps to prevent overfitting) The new cost function:

$$J(w) = \frac{1}{m}\sum_{i=1}^{m} Cost(\, h(x^{(i)}),\, y^{(i)}) \,+\, \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2$$

The regularization term will heavily penalize large wi. The effect will be less on smaller wi's. As such, the growth of w is controlled. The h(x) we obtain with these controlled params w will be more generalizable.
NOTE: λ is a hyper-parameter value. We have to find it using cross-validation.
A larger value λ of will make wi shrink closer to 0, which might lead to underfitting. λ=0 will have no regularization effect. When choosing λ , we have to take proper care of bias vs variance trade-off. [7] [16]

*2) Process:* In breast cancer prediction using logistic regression, several key steps contribute to the model's efficacy. Firstly, the sigmoid function plays a crucial role in calculating the z-value, transforming the input features into probabilities between 0 and 1. This step establishes the foundation for the binary classification task by mapping predictions to a probability scale. Subsequently, forward-backward propagation and parameter updating refine the model through multiple iterations. These processes involve adjusting the weights and biases to minimize the cost function, optimizing the model's ability to accurately classify benign and malignant instances. Following parameter optimization, predictions are made on both the training and test datasets. The accuracy of the model is assessed by comparing these predictions with the actual labels [24]. The logistic regression algorithm's effectiveness in breast cancer prediction is quantified through train and test accuracy metrics, providing insights into its generalization performance. Additionally, the implementation includes the verification of results using scikit-learn's logistic regression module, ensuring consistency and validating the custom logistic regression implementation. Overall, this iterative process, from sigmoid transformation to accuracy evaluation, embodies the systematic approach of logistic regression in predicting breast cancer, combining mathematical rigor with practical model assessment.

*3) Result*

TABLE III
RESULT OF LOGISTIC REGRESSION

| Result | Precision | Sensitivity | F-Measure |
|---|---|---|---|
| Benign | 0.98 | 0.91 | 0.94 |
| Malignant | 0.95 | 0.99 | 0.97 |

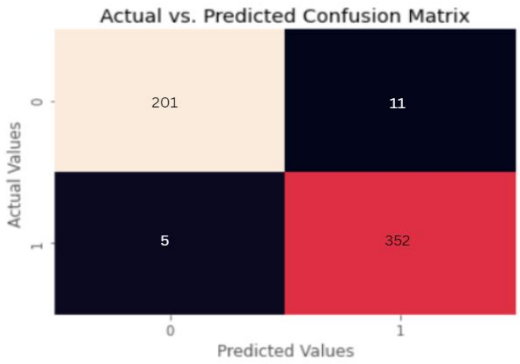



Fig.5 Confusion Matrix of Logistic Regression

*D. Random Forest and Extreme Gradient Boosting*

*1) Role:* Both gradient boosting and random forest are powerful machine learning algorithms, each with its own set of advantages and disadvantages. Random forest is known for its speed, scalability, and ability to provide reliable results even with noisy or limited data. On the other hand, gradient boosting excels at handling complex data and assessing the importance of features, albeit at a slower pace compared to random forest. The choice between these two algorithms depends on the specific problem and dataset being addressed. Data scientists and machine learning practitioners must carefully evaluate the characteristics and applications of both algorithms to determine which one will yield optimal results for a given task.

*2) Process*

- Random Forest (RF): RF consists of numerous decision trees, where the accuracy of the results correlates directly with the number of trees in the forest. Introduced by Breiman in 2001, RF utilizes C4.5 or J48 as its classifier and combines Bagging with random feature selection for decision trees. It operates as a supervised classification algorithm.
- Extreme Gradient Boosting (XGBoost): XGBoost is a decision-tree-based ensemble machine learning algorithm integrated into the gradient boosting framework. While decision trees are generally easy to visualize and interpret, grasping the intricacies of the next generation of tree-based algorithms, such as XGBoost, can pose a challenge.

*3) Result*

TABLE IV
RESULT OF RANDOM FOREST

| Result | Precision | Sensitivity | F-Measure |
|---|---|---|---|
| Benign | 0.96 | 0.94 | 0.95 |
| Malignant | 0.97 | 0.98 | 0.97 |

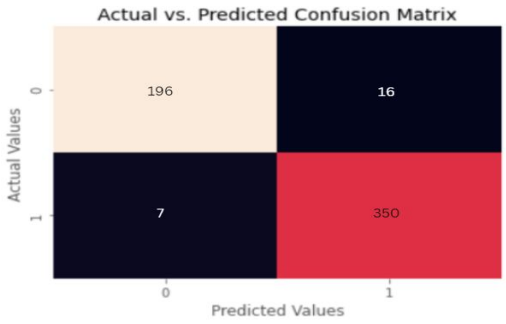



Fig.6 Confusion Matrix of Random Forest

*E. Decision Tree*

*1) Role:* In the initial stages of breast cancer identification, various algorithms such as Support Vector Machine (SVM), K Nearest Neighbor (KNN), MLP, etc., were employed. However, these algorithms did not achieve the desired accuracy in cancer detection. Our approach is to utilize the Decision Tree algorithm for breast cancer detection. Decision Tree is a supervised learning technique, and our goal is to improve the accuracy of breast cancer detection. The primary advantage of the Decision Tree algorithm lies in its ability to identify whether the predicted cancer is of malign or benign type, achieving an impressive accuracy rate of 98.8%. This suggests that the Decision Tree algorithm can be a promising method for enhancing the precision and reliability of breast cancer detection compared to other algorithms used in the early days.

*2) Process:* In the decision tree algorithm, analysis is performed by assigning importance to all attributes, both high and low. The value for the root node is determined by examining the entire trained dataset.

- Step 1: In order to do the process of learning training dataset is selected.
- Step 2: Make a map of each individual attribute to respective classes.
- Step 3: Catch all practicable values for each attribute that correlate with feasible classes.
- Step 4: Compute values of every attributes which belongs to distinctive classes.
- Step 5: Root node are generated to that attribute which has minimum number of values which reside in the unique classes.
- Step 6: Comparably pick another attribute for next extent in decision tree from prevailing attributes based on least number of values which has distinctive classes.
- Step 7: Stop

*3) Result:*

TABLE V
RESULT OF DECISION TREE

| Result | Precision | Sensitivity | F-Measure |
|---|---|---|---|
| Benign | 0.94 | 0.92 | 0.93 |
| Malignant | 0.96 | 0.97 | 0.96 |

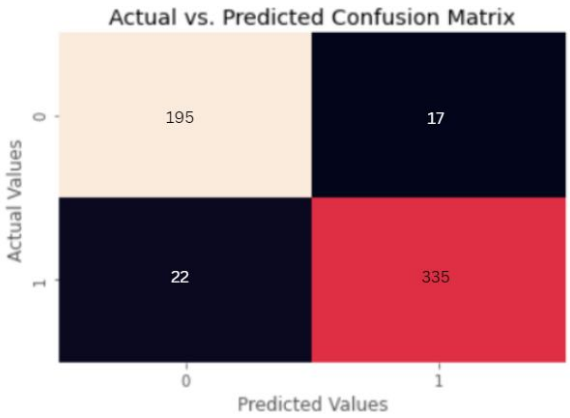



Fig. 7 Confusion Matrix of Decision Tree

## V. RESULTS AND DISCUSSIONS

After applying Machine Learning Algorithms on Breast Cancer Wisconsin Diagnostic dataset. We used Confusion Matrix, Accuracy, Precision, Sensitivity, F1 Score, AUC as performance metrics to evaluate and compare the models and identify the best algortihm for the brest cancer Prediction.

In Table VI and Figure 8, the accuracy percentages for the Wincson Breast Cancer Diagnostic datasets are presented. Upon analyzing the results from both the training set and testing set, it is observed that all the classifiers exhibit different accuracies.

Notably, SVM consistently demonstrates the highest accuracy in the testing set, achieving 97.2%, surpassing the accuracies of the other classifiers.

TABLE VI
ACCURACY PERCENTAGE FOR BREAST CANCER PREDICTION

| Models | Accuracy Training Set (%) | Accuracy Testing Set (%) |
|---|---|---|
| SVM | 98.4 | 97.2 |
| Random Forest | 99.8 | 96.5 |
| Logistic Regression | 95.5 | 95.8 |
| Decision Tree | 98.8 | 95.1 |
| K-NN | 94.6 | 93.7 |

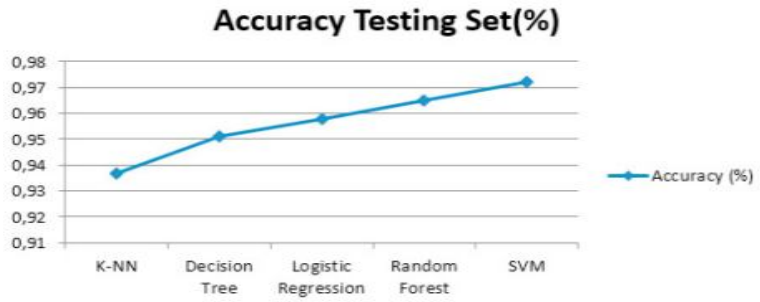



Fig. 8 Comparative graph of different classifiers

TABLE VII
CONFUSION MATRIX

| Models | Malignant | Benign | Class |
|---|---|---|---|
| SVM | 201 | 11 | Malignant |
| | 1 | 356 | Benign |
| Random Forest | 196 | 16 | Malignant |
| | 7 | 350 | Benign |
| Logistic Regression | 201 | 11 | Malignant |
| | 5 | 352 | Benign |
| Decision Tree | 195 | 17 | Malignant |
| | 22 | 335 | Benign |
| K-NN | 201 | 11 | Malignant |
| | 7 | 350 | Benign |

TABLE VIII
CLASSIFIERS PERFORMANCES

| Models | Precision | Sensitivity | F-Measure | Class |
|---|---|---|---|---|
| SVM | 0.98 | 0.94 | 0.96 | Benign |
| | 0.97 | 0.99 | 0.98 | Malignant |
| Random Forest | 0.96 | 0.94 | 0.95 | Benign |
| | 0.97 | 0.98 | 0.97 | Malignant |
| Logistic Regression | 0.98 | 0.91 | 0.94 | Benign |
| | 0.95 | 0.99 | 0.97 | Malignant |
| Decision Tree | 0.94 | 0.92 | 0.93 | Benign |
| | 0.96 | 0.97 | 0.96 | Malignant |
| K-NN | 0.92 | 0.91 | 0.91 | Benign |
| | 0.95 | 0.95 | 0.95 | Malignant |

TABLE IX
THE AREA UNDER ROC CURVE (AUC)

| Models | AUC (%) |
|---|---|
| SVM | 0.966 |
| Random Forest | 0.960 |
| Logistic Regression | 0.947 |
| Decision Tree | 0.945 |
| K-NN | 0.952 |

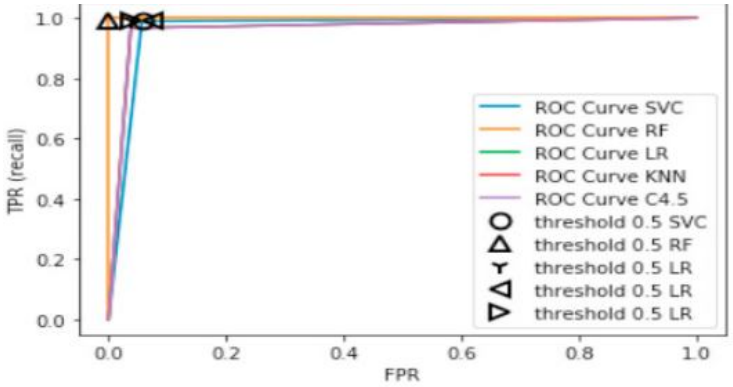



Fig.9 ROC Curves

## VI. CONCLUSION

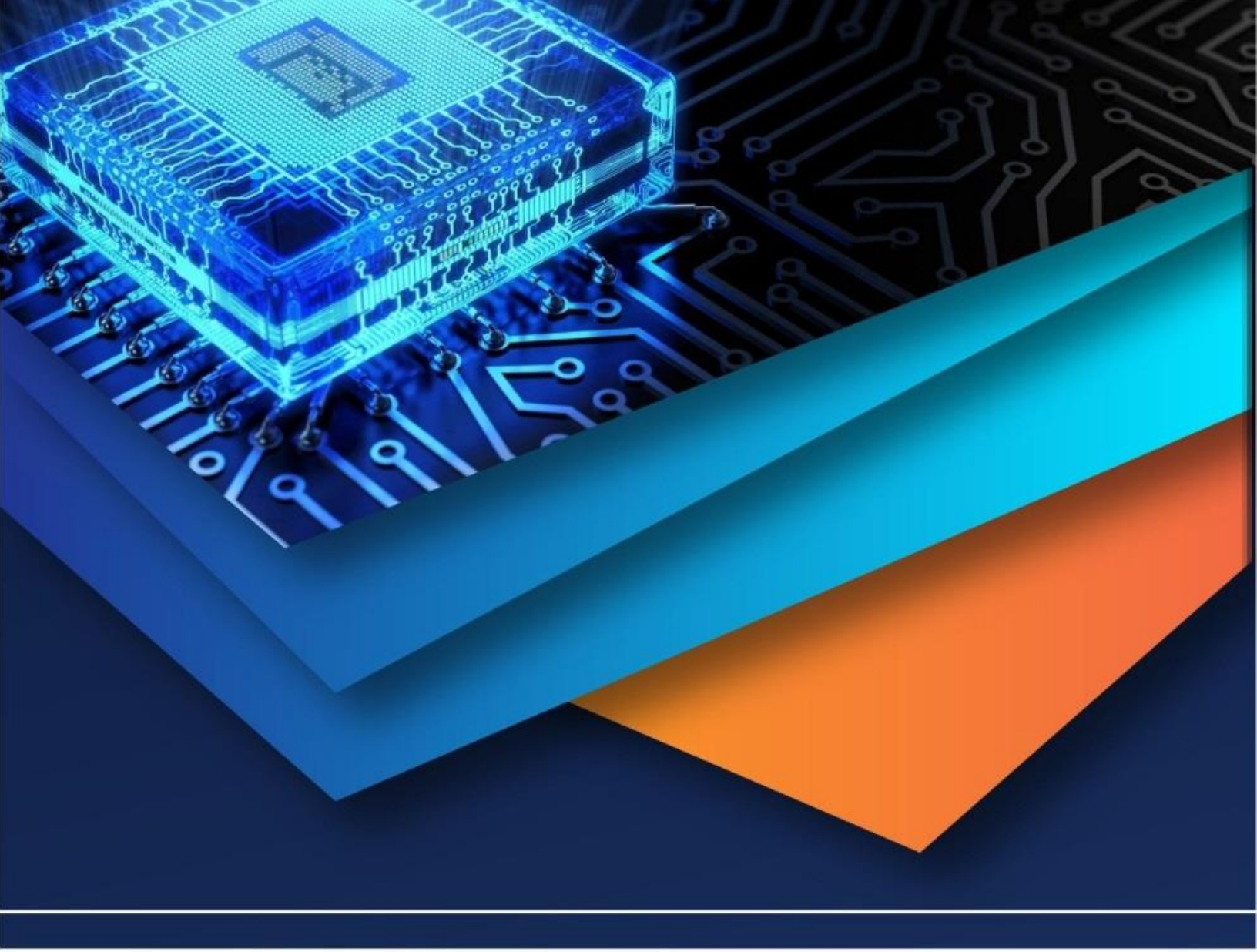
On the Wisconsin Breast Cancer Diagnostic dataset (WBCD) we applied five main algorithms which are: SVM, Random Forests, Logistic Regression, Decision Tree, K-NN, calculate, compare and evaluate different results obtained based on confusion matrix, accuracy, sensitivity, precision, AUC to identify the best machine learning algorithm that are precise, reliable and find the higher accuracy. All algorithms have been programmed in Python using scikit-learn library in Anaconda environment. After an accurate comparison between our models, we found that Support Vector Machine achieved a higher efficiency of 97.2%, Precision of 97.5%, AUC of 96.6% and outperforms all other algorithms. In conclusion, Support Vector Machine has demonstrated its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of accuracy and precision. It should be noted that all the results obtained are related just to the WBCD database, it can be considered as a limitation of our work, it is therefore necessary to reflect for future works to apply these same algorithms and methods on other databases to confirm the results obtained via this database, as well as, in our future works, we plan to apply our and other machine learning algorithms using new parameters on larger data sets with more disease classes to obtain higher accuracy.

## VII. ACKNOWLEDGMENT

We would like to express our profound gratitude to Assistant Professor Ms. Priya Singh of the Department of Software Engineering at Delhi Technological University for her invaluable guidance, insightful feedback, and unwavering support throughout the course of this research project. Her expertise and dedication not only shaped this work but also inspired us to explore our potential. We are truly grateful for her mentorship and for providing us with the opportunity to work under her guidance. This project would not have reached its fruition without her encouragement and constructive criticism

## REFERENCES

[1] R. K. Barwal and N. Raheja, "A Classification System for Breast Cancer Prediction using SVOF-KNN method," 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2022
[2] M. P. Behera, A. Sarangi, D. Mishra and S. K. Sarangi, "Breast Cancer Prediction Using Long Short-Term Memory Algorithm," 2022 5th International Conference on Computational Intelligence and Networks (CINE), Bhubaneswar, India, 2022
[3] Y. Wankhade, S. Toutam, K. Thakre, K. Kalbande and P. Thakre, "Machine Learning Approach for Breast Cancer Prediction: A Review," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023
[4] M. R. Karim, G. Wicaksono, I. G. Costa, S. Decker and O. Beyan, "Prognostically Relevant Subtypes and Survival Prediction for Breast Cancer Based on Multimodal Genomics Data," in IEEE Access, vol. 7
[5] N. Arya and S. Saha, "Multi-Modal Classification for Human Breast Cancer Prognosis Prediction: Proposal of Deep-Learning Based Stacked Ensemble Model," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 2, pp. 1032-1041, 1 March-April 2022
[6] S. Alghunaim and H. H. Al-Baity, "On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context," in IEEE Access, vol. 7, pp. 91535-91546, 2019
[7] X. Wang, W. Yu, Z. Ding, X. Zhai and S. Saha, "Modeling and Analyzing of Breast Tumor Deterioration Process with Petri Nets and Logistic Regression," in Complex System Modeling and Simulation, vol. 2, no. 3, pp. 264-272, September 2022
[8] M. Byra, K. Dobruch-Sobczak, Z. Klimonda, H. Piotrzkowska-Wroblewska and J. Litniewski, "Early Prediction of Response to Neoadjuvant Chemotherapy in Breast Cancer Sonography Using Siamese Convolutional Neural Networks," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 3, pp. 797-805, March 2021
[9] M. R. Karim, G. Wicaksono, I. G. Costa, S. Decker and O. Beyan, "Prognostically Relevant Subtypes and Survival Prediction for Breast Cancer Based on Multimodal Genomics Data," in IEEE Access, vol. 7, pp. 133850-133864, 2019
[10] Z. Huang and D. Chen, "A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm," in IEEE Access, vol. 10, pp. 3284-3293, 2022
[11] E. K. Jadoon, F. G. Khan, S. Shah, A. Khan and M. ElAffendi, "Deep Learning-Based Multi-Modal Ensemble Classification Approach for Human Breast Cancer Prognosis," in IEEE Access, vol. 11
[12] C. McIntosh and T. G. Purdie, "Contextual Atlas Regression Forests: Multiple-Atlas-Based Automated Dose Prediction in Radiation Therapy," in IEEE Transactions on Medical Imaging, vol. 35, no. 4
[13] G. Sruthi, C. L. Ram, M. K. Sai, B. P. Singh, N. Majhotra and N. Sharma, "Cancer Prediction using Machine Learning," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddha Nagar, India, 2022
[14] Y. Wankhade, S. Toutam, K. Thakre, K. Kalbande and P. Thakre, "Machine Learning Approach for Breast Cancer Prediction: A Review," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023
[15] A. Bharat, N. Pooja and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis," 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), Bangalore, India, 2018.
[16] M. Sugimoto, M. Takada and M. Toi, "Comparison of robustness against missing values of alternative decision tree and multiple logistic regression for predicting clinical data in primary breast cancer," 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 2013
[17] M. Akhil and P. V. S. Kumar, "Breast Cancer Prognosis using Machine Learning Applications," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY