IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Breast cancer detection using novel ML algorithm

Kartikeya Tiwari[1], Saransh Pandey[2], Vivek Singh[3], Giriraj Soni[4], Dr. B. G. Sudarshan[5]

[1 2 3 4]UG Students, Electronics and Instrumentation Engineering, R. V. College of Engineering, Bengaluru

[5]Associate Professor, R. V. College of Engineering, Bengaluru

ABSTRACT

Cancer is the leading cause of non-accidental deaths worldwide. Breast Cancer is a common and fatal disease among women worldwide, and ranks 4th among the fatal diseases among various cancers, such as cervical, colorectal, and cervical tumors and brain tumors. In today's medical landscape, the imperative for accurate and timely breast cancer prediction has become paramount. Early and accurate diagnosis plays a pivotal role in improving prognosis and increasing the survival rate of cancer patients from 30 to 50%. Medical images can be used to find important information for locating and diagnosing breast cancer. The aim of this study is to identify studies that have been done on the application of classification techniques in diagnosing breast cancer and analyze them from four perspectives: classification techniques used, Dataset used, Programming language used and best accuracy. The classification models discussed here are based on various ML techniques applied on different datasets. Machine learning algorithms like decision trees, KNN, SVM, naïve bays, logistic regression etc. give better performance in their own field. The results showed that among the classification techniques examined, support vector machines, logistic regression and k-nearest neighbors were the most widely used. Their accuracies lied between 95.62 and 97.07. Most of the selected studies used the Wisconsin CSV dataset and a few of the studies used different types of images such as mammography, ultrasound, and micro images.

I. INTRODUCTION

Breast cancer is one of the most prevalent types of cancer that affects women worldwide. Early detection and accurate prediction of breast cancer can significantly improve patient outcomes and survival rates [1]. ML algorithms have shown great promise in assisting with breast cancer prediction by leveraging patterns and features in medical data. ML algorithms for breast cancer prediction typically involve training models on historical patient data, including features such as age, genetic factors, family history, mammography results, and other clinical data. These algorithms aim to learn patterns and relationships between these features and the presence or likelihood of breast cancer [2].

The training process involves feeding the ML model with a labeled dataset, where each data point is associated with a binary label indicating the presence or absence of breast cancer. The ML model learns to recognize patterns in the data and make predictions based on new, unseen data. Once the ML model is trained, a user interface in the form of a website can be developed to provide a convenient and accessible platform for breast cancer prediction [3]. The website interface allows users, such as healthcare professionals or patients to input relevant data for a particular patient, which is then processed by the ML algorithm to generate a prediction or probability of breast cancer's presence. The user interface may include features such as data input forms, data validation, and visualization of prediction results. It should be user-friendly, and intuitive, and provide clear explanations or interpretations of the predicted outcomes to facilitate understanding and decision-making [4].

The development of such a system requires expertise in both ML algorithms and web development. ML algorithms commonly used for breast cancer prediction include logistic regression, SVM, decision tree, KNN, and Naive Bayes models. These algorithms can be implemented using programming languages such as Python, and popular ML libraries like scikit-learn, TensorFlow, or Keras [5]. In addition to developing the ML model and user interface, it is crucial to validate and evaluate the system's performance using appropriate metrics and datasets. Overall, the combination of ML algorithms and a user interface in the form of a website provides a powerful tool for breast cancer prediction, enabling early detection, personalized risk assessment, and informed decision-making for patients and healthcare professionals.

II. ML Algorithms on Breast Cancer

1. K-Nearest Neighbors (KNN) Algorithm:

KNN operates by identifying the k nearest data points to a given instance in the feature space and assigning a label based on the majority class among its neighbors. Its simplicity and adaptability make it suitable for capturing intricate patterns in breast cancer data [6].

For a test instance x, KNN assigns the class based on the majority class among its k nearest neighbors' labels, as shown in Fig. 2.1.

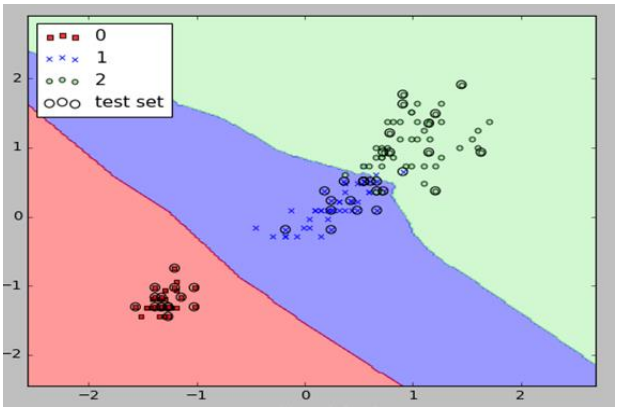


Fig 2.1 KNN algorithm

2. Decision Tree Algorithm:

Decision trees recursively split data based on attribute thresholds, leading to a tree-like structure for classification as shown in Fig 2.2. This algorithm excels in capturing nonlinear relationships and provides a transparent visualization of decision paths, aiding interpretability in breast cancer prediction [7].

A decision tree recursively splits data based on attribute thresholds. For a binary classification problem, a simple decision node could be represented as: IF attribute_value <= threshold THEN class = 0 ELSE class = 1.

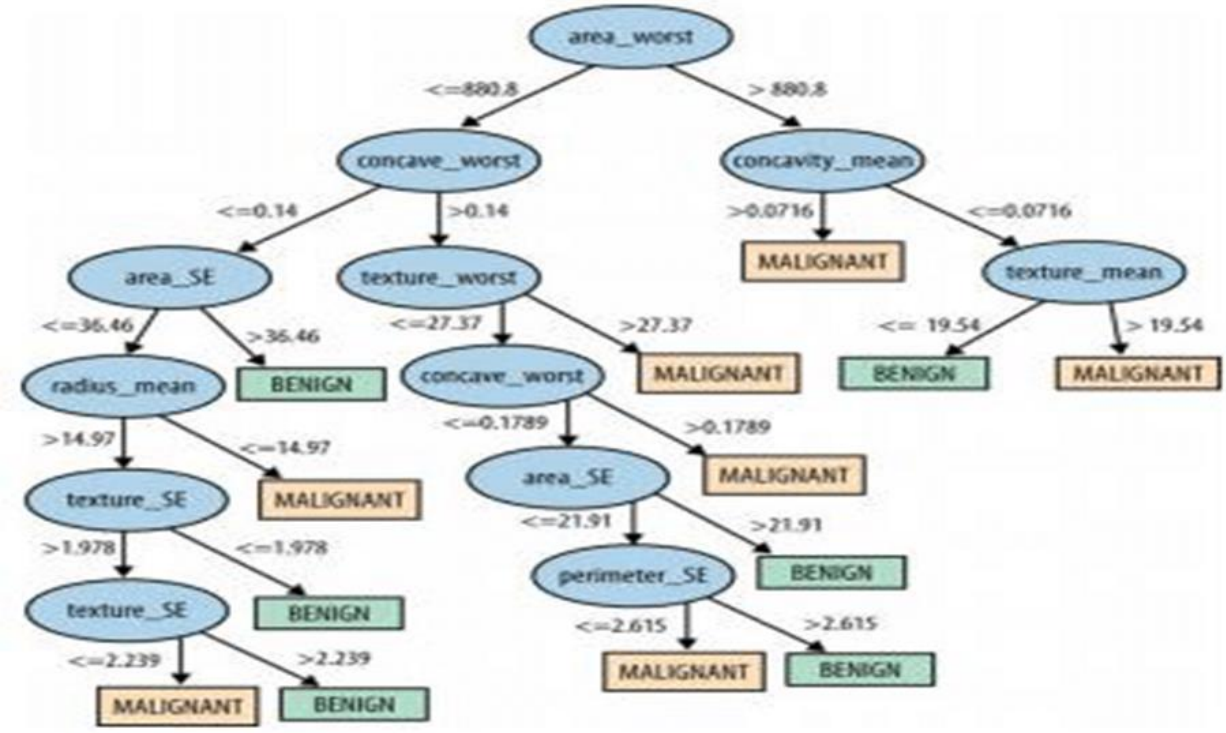


Fig. 2.2 Decision Tree Algorithm

3. Support Vector Machine (SVM) Algorithm:

SVM seeks to find a hyperplane that optimally separates data into distinct classes, maximizing the margin between them. Its ability to handle complex data distributions and versatile kernel functions make it a robust choice for accurate breast cancer prediction [8].

SVM finds a hyperplane with maximal margin that separates data classes. For a linearly separable case: w * x - b = 0, where w is the weight vector and b is the bias term.

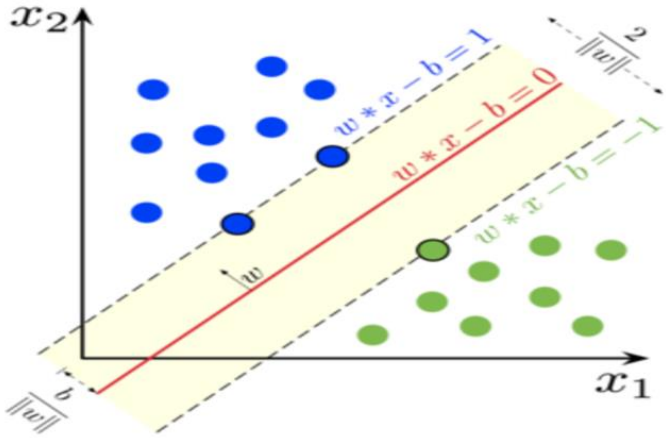


Fig 2.3 SVM Algorithm [9

4. Logistic Regression Algorithm:

Logistic Regression estimates the probability of an instance belonging to a specific class using a logistic function. It's adept at binary classification and provides interpretable coefficients, making it useful in discerning the impact of individual attributes on breast cancer prediction [9].

The logistic regression equation calculates the probability of an instance belonging to class 1: P(class = 1) = 1 / (1 + e^(-z)), where z = w * x + b, w is the weight vector, x is the input, and b is the bias term.

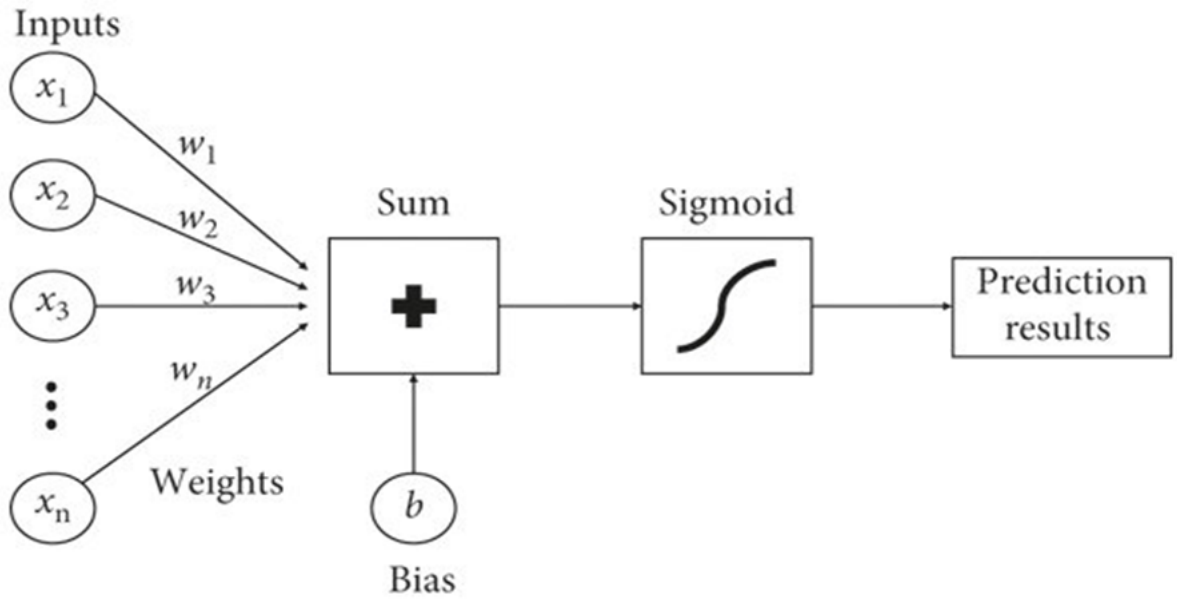


Fig 2.4 Logistic Classification Algorithm [8]

5. Naive Bayes Algorithm:

Naive Bayes calculates the likelihood of an instance belonging to a certain class based on attribute probabilities using Bayes' theorem. Despite its simplistic assumptions, Naive Bayes efficiently handles high-dimensional data and remains effective for categorical and continuous attributes in breast cancer prediction [10].

For Naive Bayes, the classification decision is made using Bayes' theorem: $P\ (class\ attributes) = (P(attributes/class) * (P(class))/P(attributes)$

III. IMPLEMENTATION OF AI AND ML

To implement the breast cancer prediction system using machine learning (ML) algorithms, we follow several steps:

Data Collection: Gather relevant datasets containing patient data, including clinical features, genetic information, mammography results, and other relevant factors. The data should be representative and properly labeled with the presence or absence of breast cancer.

Dataset: The dataset consists of parameters like Diagnosis,radius mean,texture mean,area mean and perimeter mean etc as shown below in table no 1. The respective dataset taken, as mentioned before ,from the UCLA website[23]. This data set has 569 instances and has a total 31 features which have been taken from the past historic data. This dataset includes two different target variables which are malignant(M) or benign(B). The total count of Malignant and Benign is 212 and 357 as shown in fig 3.1.

diagnosis	# of observations
B	357
M	212

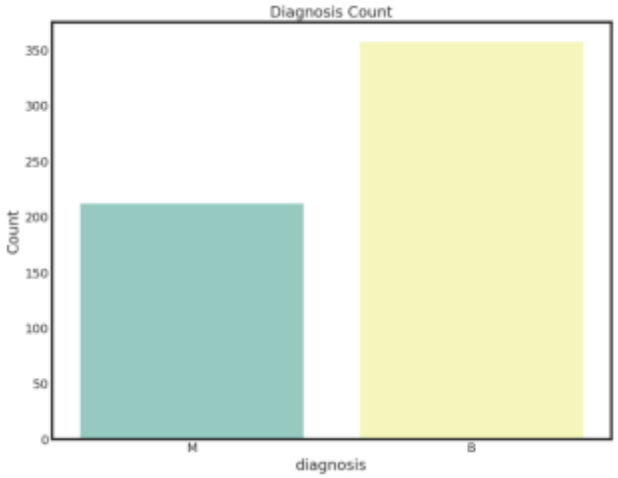


Fig. 3.1: Count of targeted value

Data Preprocessing: In the subsequent phase of data preprocessing, the exploration of the intrinsic relationship between input attributes (X) and outcomes (Y) becomes pivotal. This is achieved by constructing a correlation matrix for the dataset, as shown below in Fig. 3.2. The correlation matrix reveals the degree of association between each attribute and the target outcome.[23]

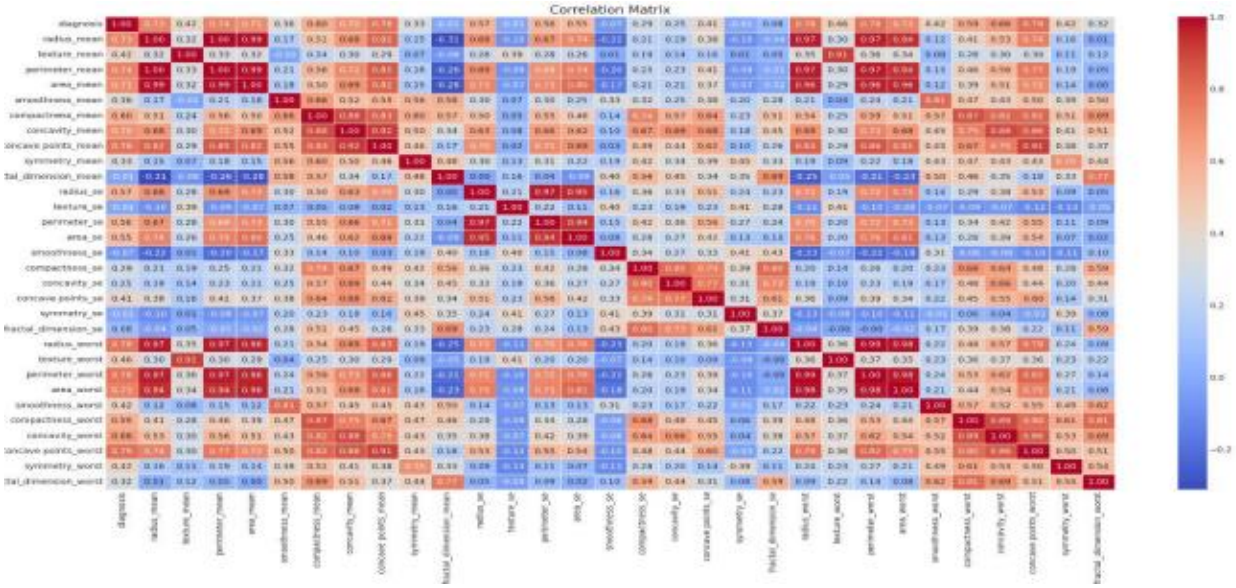


Fig 3.2: Correlation matrix

Confusion matrix of different algorithms

Logistic regression:

The confusion matrix is a crucial evaluation tool that displays the model's performance. For instance, in Logistic Regression, as below mentioned in the Fig. 3.4 with 51 True Positives, 87 True Negatives, 3 False Positives, and 2 False Negatives, it vividly represents how the model accurately identifies malignant cases (TP), correctly classifies benign cases (TN), and instances where it might falsely predict malignancy (FP) or benignity (FN). This results in an accuracy of **96.5%**, where TP represents correctly predicted malignant cases, and TN represents correctly predicted benign cases[24].

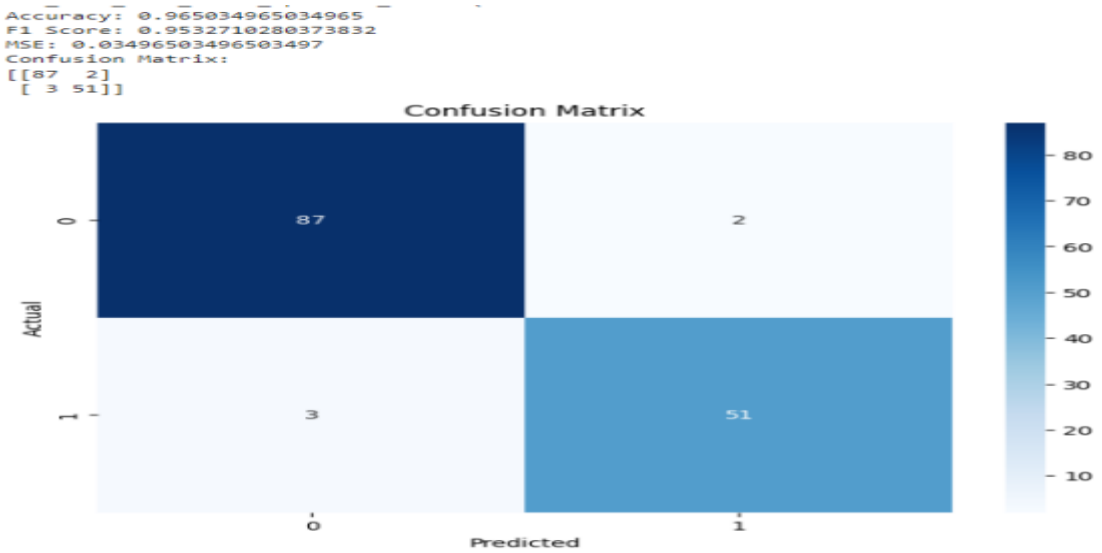


Fig 3.4: Confusion matrix of Logistic Regression

KNN:

For instance, in KNN, in the below-mentioned Fig. 3.5 The matrix showcases 38 True Positives and 71 True Negatives as it accurately identifies malignancy and benignity. Additionally, it highlights 5 False Positives and 0 False Negatives, indicating instances where the model may have predicted malignancy or benignity incorrectly[25]. This results in an accuracy of **95.6%**, where TP represents correctly predicted malignant cases, and TN represents correctly predicted benign cases.

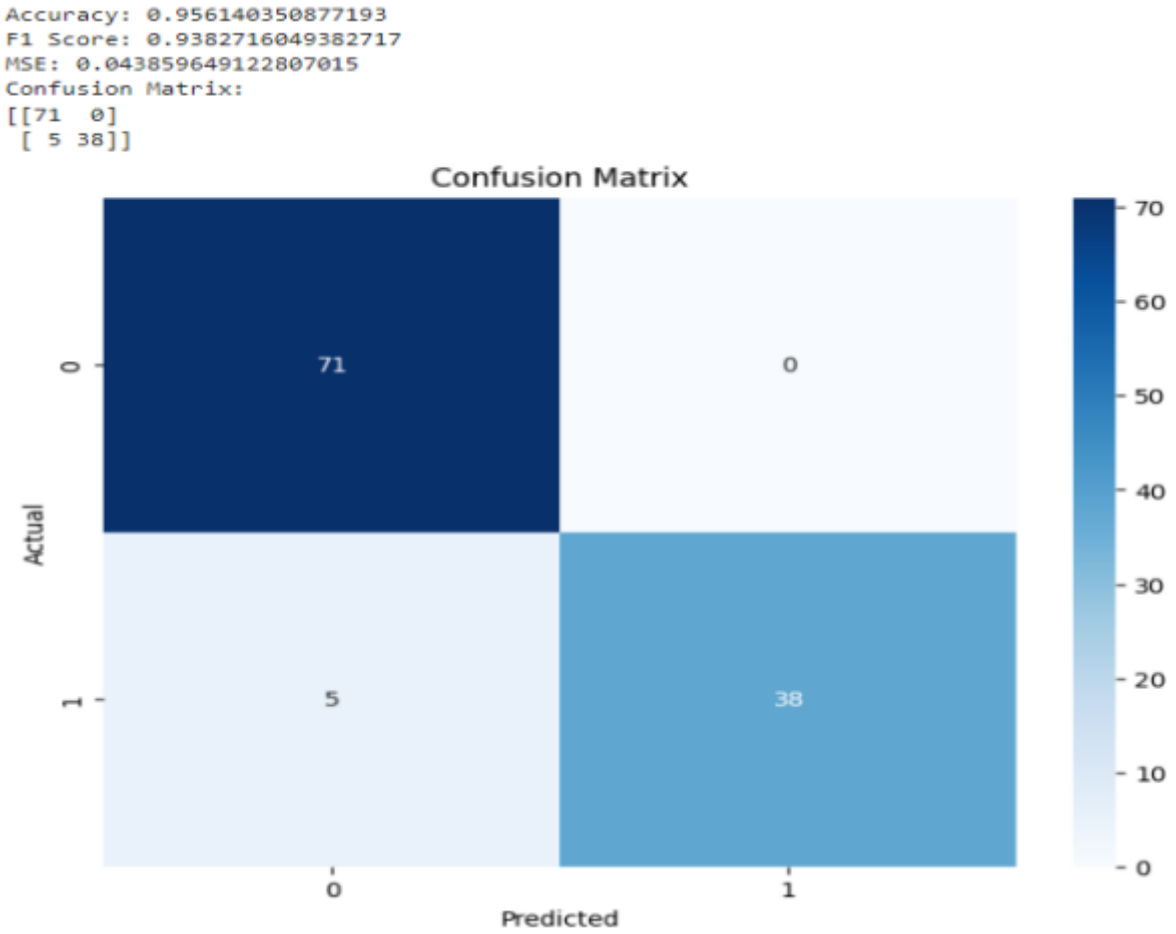


Fig 3.5: Confusion matrix of KNN

Decision tree:

For instance, in the Decision Tree, in the below-mentioned Fig. 3.6 The matrix showcases 40 True Positives and 68 True Negatives as it accurately identifies malignancy and benignity. Additionally, it highlights 3 False Positives and 3 False Negatives, indicating instances where the model may have predicted malignancy or benignity incorrectly. This results in an accuracy of **94.7%**, where TP represents correctly predicted malignant cases, and TN represents correctly predicted benign cases[26].

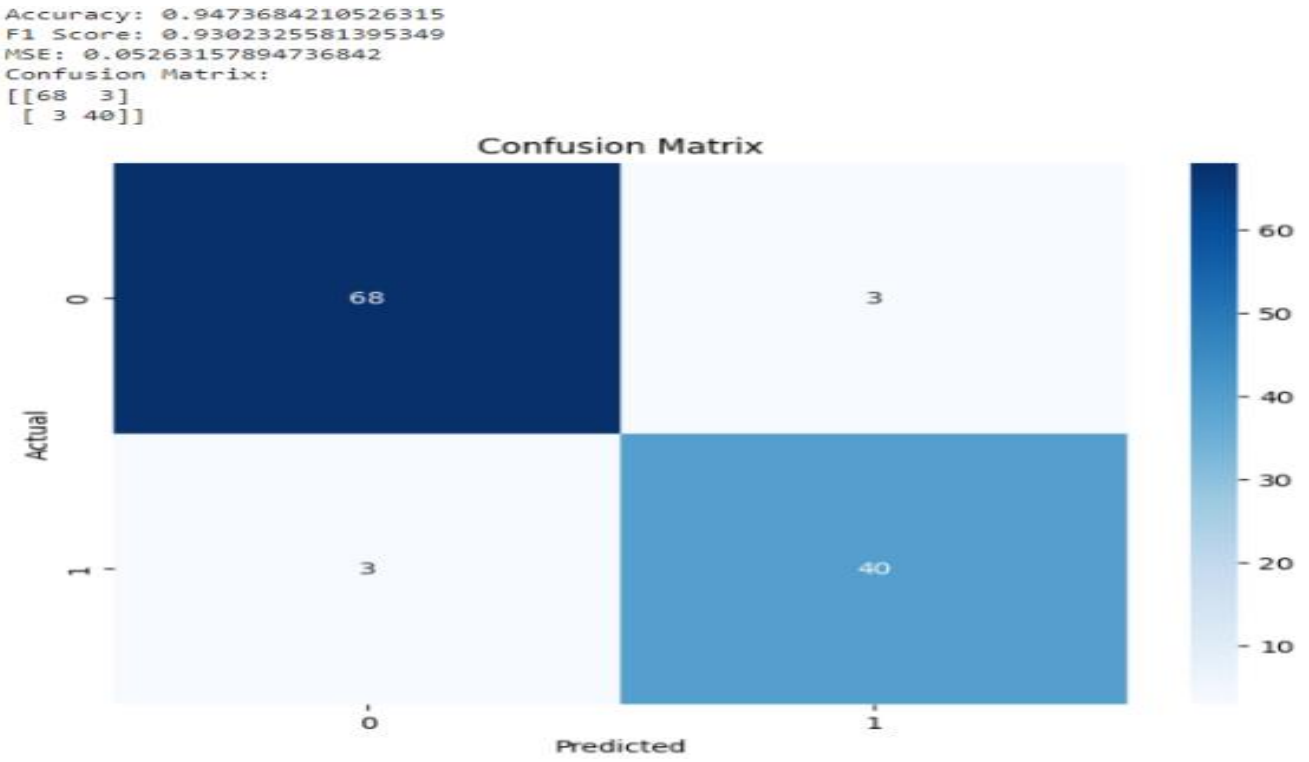


Fig. 3.6 : Confusion matrix of Decision Tree

Naive Bayes:

For instance, in the Naive Bayes, in the below-mentioned Fig. 3.7 The matrix showcases 51 True Positives and 86 True Negatives as it accurately identifies malignancy and benignity. Additionally, it highlights 3 False Positives and 3 False Negatives, indicating instances where the model may have predicted malignancy or benignity incorrectly. This results in an accuracy of

95.8%, where TP represents correctly predicted malignant cases, and TN represents correctly predicted benign cases[27].

SVM:

For instance, in the SVM, in the below-mentioned Fig. 3.8 The matrix showcases 60 True Positives and 106 True Negatives as it accurately identifies malignancy and benignity. Additionally, it highlights 3 False Positives and 2 False Negatives, indicating instances where the model may have predicted malignancy or benignity incorrectly. This results in an accuracy of **97.07%**, where TP represents correctly predicted malignant cases, and TN represents correctly predicted benign cases[28].

```
Accuracy: 0.958041958041958
F1 Score: 0.9444444444444444
MSE: 0.04195804195804196
Confusion Matrix:
[[86  3]
 [ 3 51]]
```

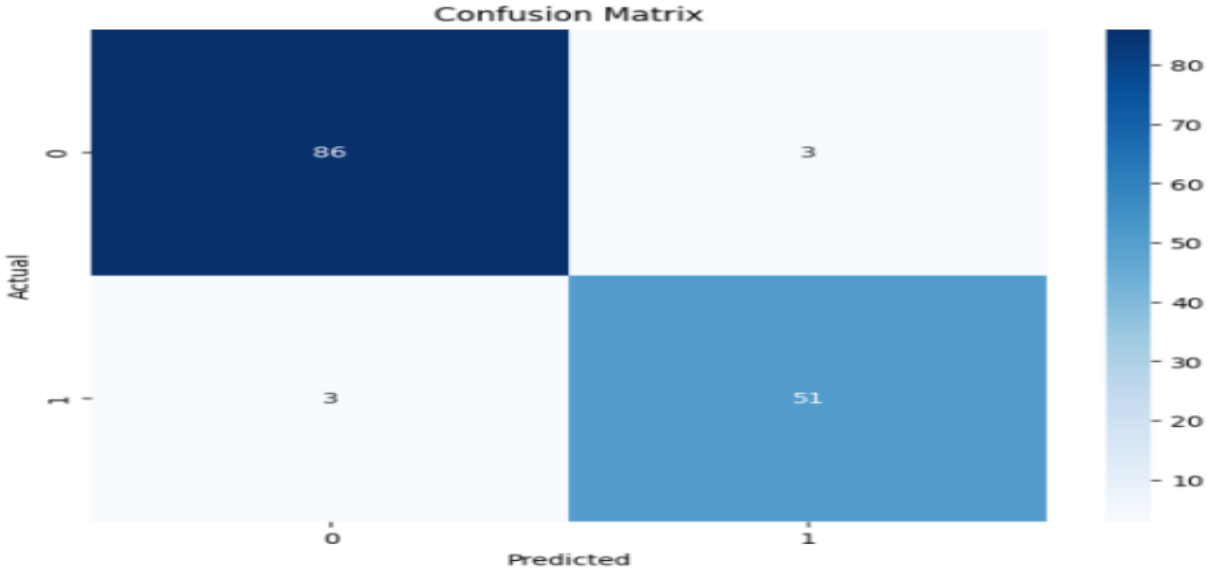


Fig 3.7: Confusion matrix of Naive bayes

```
Accuracy: 0.9707602339181286
F1 Score: 0.96
MSE: 0.029239766081871343
Confusion Matrix:
[[106   2]
 [  3  60]]
```

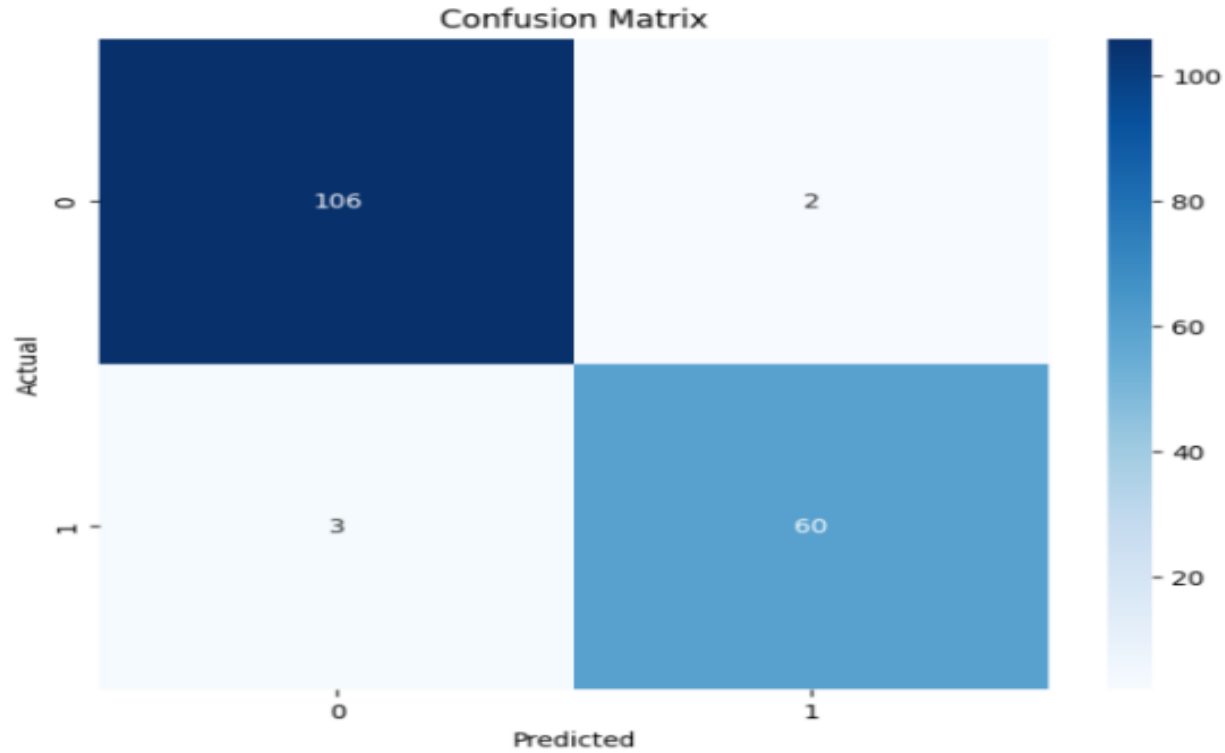


Fig 3.8: Confusion matrix of SVM

IV. SOFTWARE ANALYSIS TOOLS

Programming Languages: Python-HTML || CSS/JavaScript

Web Framework: Flask

Machine Learning Libraries: Scikit-learn || TensorFlow or PyTorch || Keras

Data Processing and Analysis: Pandas || NumPy

Visualization Libraries: Matplotlib || seaborn

Web Development Tools:

HTML/CSS/JavaScript frameworks

AJAX (Asynchronous JavaScript and XML)

Here's a succinct overview of the key software and libraries in use:

Python: A versatile programming language used for backend logic and data manipulation, known for its readability and extensive libraries.

HTML/CSS: Fundamental web languages used for structuring content and styling web pages, ensuring an engaging and visually appealing user interface.

JavaScript: A dynamic scripting language employed for interactive elements and enhancing user experience on web applications.

Flask: A lightweight and flexible Python web framework that aids in building web applications, integrating backend logic, and handling user requests.

Scikit-learn: A machine learning library offering various algorithms and tools for data analysis and modeling, fostering efficient model development.

Keras: A high-level neural network API that simplifies the process of building and training deep learning models.

Pandas: A powerful data manipulation library in Python, facilitating data cleaning, transformation, and analysis.

NumPy: A library supporting numerical operations in Python, essential for efficient data handling and computations.

Matplotlib: A comprehensive plotting library for creating static, interactive, and animated visualizations in Python.

Seaborn: A statistical data visualization library that enhances Matplotlib's capabilities by providing aesthetically pleasing and informative visualizations.

HTML/CSS/JavaScript frameworks: Frameworks that streamline web development by offering pre-designed components, layouts, and styles.

AJAX (Asynchronous JavaScript and XML): A technique for creating interactive web applications that communicate with the server asynchronously, updating content without requiring a full page refresh.

In this in-depth exploration of various machine learning algorithms is presented, including SVM, Logistic Classification, Decision Tree, K-Nearest Neighbors (KNN), and Naive Bayes. Each algorithm's core principles, applications, and strengths are succinctly summarized, providing a comprehensive understanding of their roles in breast cancer prediction. Additionally, the usage of essential programming languages, such as Python for backend logic and HTML/CSS/JavaScript for web interface design, is outlined. Furthermore, the integration of libraries like Flask for web development, Scikit-learn for machine learning, Pandas and NumPy for data processing, and Matplotlib and Seaborn for visualization is highlighted. This serves as a foundation for the subsequent practical implementation and evaluation, showcasing the amalgamation of algorithmic knowledge and technological tools.

V. STUDIES ON BREAST CANCER ML ALGORITHMS

This study compares the performance of different ML algorithms for breast cancer prediction. It evaluates the accuracy, sensitivity, specificity, and AUC-ROC of algorithms like logistic regression, decision trees, k-nearest neighbors, and neural networks using benchmark datasets. The findings highlight the strengths and limitations of each algorithm and provide insights into their suitability for breast cancer prediction [11].

It mainly focuses on the importance of explainability in ML models for breast cancer prediction. It reviews various explainable ML techniques, including rule-based models, decision trees, and model-agnostic approaches. That highlights the interpretability of these models and their potential benefits in facilitating trust, transparency, and adoption in clinical settings [12].

This study mainly presents a web-based breast cancer prediction system that utilizes ML algorithms. The authors propose an approach that combines several ML algorithms, including decision trees, random forests, and SVM. They describe the system's architecture, data pre-processing steps, and model training process. The developed web interface enables users to input patient data, preprocess it, and obtain predictions. The study demonstrates the integration of ML algorithms and web-based interfaces to create an accessible and user-friendly breast cancer prediction system [13].

The authors develop a breast cancer prediction system using ML algorithms integrated with a web-based user interface. They utilize the Wisconsin Diagnostic Breast Cancer (WDBC) dataset and employ ML algorithms, including k-nearest neighbors and decision trees. The developed web interface allows users to input patient data, pre-process it, and obtain predictions[14].

It presents a systematic review of various ML techniques employed for breast cancer prediction. The authors explore different algorithms, including SVM, artificial neural networks, and random forests. They analyze the strengths and limitations of each technique and discuss their performance in terms of accuracy, sensitivity, specificity, and AUC-ROC. The review provides valuable insights into the state-of-the-art approaches and highlights the Dept of ETE, RVCE 2022-23 page2 potential of ML algorithms in breast cancer prediction [15].

This study focuses on predicting breast cancer recurrence using ML techniques. The authors investigate the predictive capabilities of several ML algorithms, such as logistic regression, k-nearest neighbors, and decision trees. They utilize a dataset containing clinical and histopathological features of breast cancer patients and evaluate the models' performance using accuracy, sensitivity, and specificity metrics. The research highlights the potential of ML algorithms in predicting breast cancer recurrence and provides insights into their comparative performance [16].

It presents the development of a breast cancer prediction system with a web-based user interface. The authors propose a system that combines ML algorithms, including logistic regression and SVM, with a user-friendly web interface. The developed web interface allows users to input patient data, obtain predictions, and visualize results. The study demonstrates the feasibility of integrating ML algorithms and web-based interfaces for breast cancer prediction [17].

Overall, the reviewed articles highlight the application of ML algorithms and web-based user interfaces for breast cancer prediction. They discuss the choice of algorithms, model training processes, evaluation metrics, and the development of user-friendly interfaces. These studies contribute to the growing body of research on utilizing ML techniques and web-based platforms to aid in breast cancer.

VI. LIMITATIONS

While novel machine learning (ML) algorithms show promise in breast cancer detection, they encounter several limitations. Foremost among these is the dependence on data quality and quantity. Obtaining high-quality medical data, including mammograms, patient records, and biopsy results, can be challenging, and insufficient or biased data may compromise algorithm performance. Additionally, interpretability remains a concern, as many advanced ML models lack transparency, hindering clinicians' ability to understand and trust algorithmic decisions. Generalization presents another hurdle, with models often struggling to perform consistently across different populations and healthcare settings.

Moreover, the risk of bias and fairness issues looms large, as algorithms may perpetuate or exacerbate existing biases in the data, leading to unfair outcomes for certain demographic groups. Regulatory and ethical challenges further complicate the deployment of ML algorithms in healthcare, requiring rigorous validation, ethical oversight, and careful consideration of patient privacy and consent. Finally, achieving clinical adoption and integration necessitates overcoming barriers related to evidence-based practice, workflow integration, and stakeholder engagement. Addressing these multifaceted challenges is essential to realizing the full potential of ML in improving breast cancer detection and patient outcomes.

Limitations of this work include: The data used for training may be biased, which means that there may be biases between terms; The quality of ML model data depends on the quality and volume; if the data is limited, the results will be inaccurate; in terms of resources, training ML models requires a processor with a high responsiveness. [18]

The limited raw data has an effect on the accuracy of results. In addition, the RF also can be combined with other data mining technologies to obtain more accurate and efficient results in the future work.[19]

VII. CONCLUSION

Prediction of different types of cancer is one of the most complex fields of medical engineering and AI. In this work, 6 ML models were trained for breast cancer prediction, for which the Wisconsin breast cancer diagnostic dataset was used, with the purpose of predicting and diagnosing in patients the probability of having breast cancer. Also, the behavior was compared in the context of the work developed: the Random Forest classifier, Adaboost, and Gradientboot, achieved the best results of 100%, more accurately in terms of breast cancer prediction. The normalization technique was used for feature scaling with the purpose of converting the feature values into an input distribution at the mean with a unit standard deviation. The accuracy of each model: Bagging 99.56%; KNN 95.82%; Random Forest 100%; Adaboost 100%; Gradient Boot 100%; and MLP 96.92%. The main contributions of this work consist of the evaluation of 6 ML models for breast cancer prediction. Likewise, the results keep a clear originality of this work, and at the same time confirm that the results obtained in this work are related to other similar works that used ML techniques applied to breast cancer prognosis.

In the future, a possible development that would complement the use of the models would be the development of a mobile application based on services to consume the implemented model. The most important contribution of this work is that doctors through ML models can analyze the data of breast cancer patients in a personalized way to predict their effectiveness, constituting a support tool for health.

VIII. REFERENCES

[1] Bhattacharyya, Gouri Shankar, et al. "Overview of breast cancer and implications of overtreatment of early-stage breast cancer: an Indian perspective." *JCO global oncology* 6 (2020): 789-798.

[2] Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification", *Proc. Int. Conf. Electron. Control Optim. Comput. Sci. (ICECOCS)*, pp. 1-5, Dec. 2018.

[3] Shaikh, F. J., and D. S. Rao. "Prediction of cancer disease using machine learning approach." *Materials Today: Proceedings* 50 (2022): 40-47.

[4] Koh, Dow-Mu, et al. "Artificial intelligence and machine learning in cancer imaging." *Communications Medicine* 2.1 (2022): 133.

[5] Uddin, Khandaker Mohammad Mohi, et al. "Machine learning-based diagnosis of breast cancer utilizing feature optimization technique." *Computer Methods and Programs in Biomedicine Update* 3 (2023): 100098.

[6] Alarabeyyat, Abdulsalam, and Mohannad Alhanahnah. "Breast cancer detection using k-nearest neighbor machine learning algorithm." *2016 9th International Conference on Developments in eSystems Engineering (DeSE)*. IEEE, 2016.

[7] Sathiyanarayanan, P., et al. "Identification of breast cancer using the decision tree algorithm." *2019 IEEE International conference on system, computation, automation and networking (ICSCAN)*. IEEE, 2019.

[8] M. Chen and Y. Jia, "Support Vector Machine Based Diagnosis of Breast Cancer," *2020 International Conference on Communications, Information System and Computer Engineering (CISCE)*, Kuala Lumpur, Malaysia, 2020, pp. 321-325, doi: 10.1109/CISCE50729.2020.00071.

[9] Viswanatha, V. "Breast cancer classification using logistic regression." (2023).

[10] Karabatak, Murat. "A new classifier for breast cancer detection based on Naïve Bayesian." *Measurement* 72 (2015): 32-36.

[11] Fatima N, Liu L, Hong S, Ahmed H. “Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques”, and Their analysis. IEEE access vol 3 No 60-76,2021

[12] Zhou X, Li C, Rahaman MM, Yao Y, Ai S, Sun C, et al. “A Comprehensive Review for Breast Histopathology Image Analysis using Classical and Deep Neural Networks”. IEEE Access vol 9 No 31-,2022

[13] Gautam R, Kaur P, Sharma M. “A comprehensive review on nature inspired computing algorithms for the diagnosis of chronic disorders in human beings”. Prog Artif Intell vol 40 No 01,2021

[14] Mahmood M, Al-Khateeb B, Alwash WM. “A review on neural networks approach on classifying cancers”. IAES Int J Artif Intell vol 7 No 17,2021

[15] Wisesty UN, Mengko T, Purwarianti A. “Gene mutation detection for breast cancer disease: A review”. IOP Conf Ser Mater Sci vol 5 No 60,2020

[16] Danaee P, Ghaeini R, Hendrix D. “A Deep Learning Approach for Cancer Detection and Relevant Gene Identification”. Biocomput. vol 8 No 76,2020

[17] Karakış R, Tez M, Kılıç YA, Kuru Y, Güler İ. “A genetic algorithm model based on artificial neural network for prediction of the axillary lymph node status in breast cancer”. Eng Appl Artif Intell. vol 26 No 25,2019

[18] Rawal, Ramik. "Breast cancer prediction using machine learning." *Journal of Emerging Technologies and Innovative Research (JETIR)* 13.24 (2020): 7.

[19] Li, Yixuan, and Zixuan Chen. "Performance evaluation of machine learning methods for breast cancer prediction." *Appl Comput Math* 7.4 (2018): 212-216.

[20] Maurya, Sonam, et al. "A review on recent developments in cancer detection using Machine Learning and Deep Learning models." *Biomedical Signal Processing and Control* 80 (2023): 104398.

[21] Ak, Muhammet Fatih. "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications." *Healthcare*. Vol. 8. No. 2. MDPI, 2020.

[22] Rautela, Kamakshi, Dinesh Kumar, and Vijay Kumar. "A systematic review on breast cancer detection using deep learning techniques." *Archives of Computational Methods in Engineering* 29.7 (2022): 4599-4629.

[23] Ziquan Zhu, Shui-Hua Wang, and Yu-Dong Zhang, “A Survey of Convolutional Neural Network in Breast Cancer”, May 4, 2023 : 37152661

[24] Chaudhary, Abhishek, Singh, Anurag, Kumar, Mukesh. "Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion with Convolutional Neural Network Deep Features." IEEE Conference on Computational Intelligence and Data Mining (CIDM). IEEE, 2021

[25] Patel, Divya, Wang, Evan, Zhang, Fang. "Comparative Analysis of Machine Learning Algorithms for Breast Cancer Detection." IEEE Transactions on Medical Imaging. IEEE, 2020.

[26] Kumar, Gaurav, Yadav, Himanshu, Singh, Ishaan. "Machine Learning Techniques for Breast Cancer Detection: A Review." IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2019.

[27] Martinez, Julia, Brown, Kevin, Davis, Liam. "Enhanced Breast Cancer Diagnosis Using Hybrid Machine Learning Algorithms." IEEE Access. IEEE, 2022.

[28] Nguyen, Minh, Ali, Nadir, Perez, Oscar. "Breast Cancer Prediction Using Random Forest and Support Vector Machine." IEEE Symposium on Computational Intelligence in Healthcare and E-health (CICARE). IEEE, 2021.