e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

COMPARATIVE ANALYSIS OF MACHINE LEARNING CLASSIFIERS FOR DIAGNOSTIC PRECISION IN BREAST CANCER DETECTION

Riya Satavlekar[*1]**, Sarvesh Chakradeo**[*2]

[*1]Student, MIT School of Bioengineering Sciences and Research, MIT ADT University, Loni Kalbhor, Maharashtra, India.

[*2]Student, Computer Engineering, Pune Institute of Computer Technology, Pune, Maharashtra, India.

DOI : https://www.doi.org/10.56726/IRJMETS45818

ABSTRACT

As reported by the World Health Organization, breast cancer ranks as the most diagnosed life-threatening malignancy among women globally, with an estimated annual incidence of approximately 2.1 million cases.[3] Machine learning techniques have rapidly evolved and become indispensable tools in the field of medical diagnostics. This study aims to explore and evaluate the performance of five widely employed classifiers, including Pocket Perceptron, Support Vector Machine (SVM), Naive Bayes, Random Forest, and K-Nearest Neighbors (KNN). Multiple performance metrics such as accuracy, precision, F1 score, and ROC (AUC), were analyzed, providing a holistic view of each classifier's capabilities. The Wisconsin original breast cancer data set was used as a training set, and all the aforementioned techniques were implemented in a python environment. The results obtained in this study give insights into the strengths and weaknesses of these state of art ML techniques for breast cancer detection. These techniques continue to evolve, and they hold the promise of further enhancing our ability to identify and combat the disease effectively, ultimately improving patient care.

Keywords: Support Vector Machine, K-Nearest Neighbors, Machine Learning, Classifiers, Breast Cancer, Diagnosis.

I. INTRODUCTION

Machine learning techniques have emerged as a ray of hope in the battle against breast cancer, introducing a new era of promise.[4] During the in-depth examination of breast cancer diagnosis, a spectrum of machine learning techniques was used to analyze their unique attributes and intricacies. This section comprises the introduction to the multiple techniques used further in the comparative analysis, where the underlying mechanisms of each method were dissected and their potential in the field of medical diagnostics was assessed. By delving into the inner workings of Pocket Perceptron, Support Vector Machine (SVM), Naive Bayes, Random Forest, and K-Nearest Neighbors (KNN), the objective of study is to unravel the technical intricacies of these classifiers and their suitability for breast cancer detection, ultimately contributing to advancements in patient care and outcomes.

II. METHODOLOGY

Data Collection

The study is anchored in the utilization of the Original Wisconsin Breast Cancer Data set, sourced from the UCI Machine Learning Repository, an accessible online repository. Dr. William H. Wolberg, from the University of Wisconsin Hospitals, systematically collected this dataset over a span of three years. The dataset comprises a total of 669 instances categorized as either malignant or benign, with 458 instances attributed to the benign class and 241 to the malignant class.

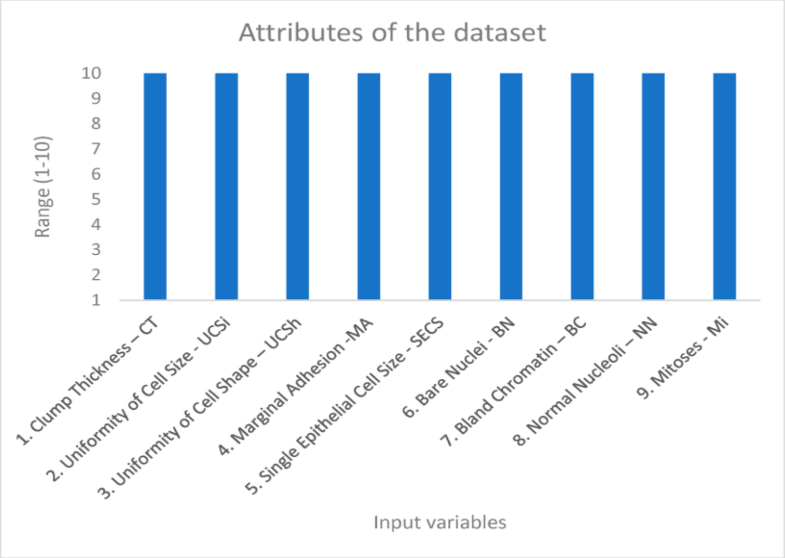


Figure 1: Attributes of Dataset (Ref[22])

Data Preprocessing

This involved steps like Data Cleaning, Data Transformation and Data Splitting. In data cleaning, any missing or inconsistent values in the dataset were handled appropriately. In data transformation, data normalization was done and the categorical data like 'malignant' and 'benign' were converted to binary values using one hot encoding. [2][4][6][8]

Data Splitting

The dataset was split into training and testing sets to evaluate the models effectively. The data set was split into 80-20 for training and testing data.

Model Training

Each classifier was trained on the training dataset using their respective algorithms and hyperparameters. Cross-validation techniques, such as k-fold cross-validation, were applied to optimize model performance and reduce overfitting.

The machine learning classifiers implemented and evaluated in this study were:

a. Pocket Perceptron

b. Support Vector Machine (SVM)

c. Naive Bayes

d. Random Forest

e. K-Nearest Neighbors (KNN)

Model Evaluation

The performance metrics used to assess the performance of each classifier were Accuracy, Precision, F1 Score, Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC).

Model Comparison

The metrics obtained for each classifier were analyzed and compared to provide insights into their respective strengths and weaknesses in diagnosing breast cancer. T

III. MODELING AND ANALYSIS

Pocket Perceptron:

Pocket Perceptron algorithm's linear decision boundary serves as a pivotal tool to delineate benign and malignant tumors based on distinct features or attributes of the tumors.[19] This algorithm progressively refines its decision boundary through an iterative process aimed at minimizing classification errors, rendering it particularly well-suited for discerning various classes of breast tumors. Notably, features such as tumor size

(S), shape (Sh), and texture (T) are harnessed as essential input parameters (x) in the algorithm. The decision boundary can be represented mathematically as:

$$y = w_1 * S + w_2 * Sh + w_3 * T + b$$

Where:

y is the output, representing the classification result.

w_1, w_2, and w_3 are the weights assigned to tumor size, shape, and texture, respectively.

b is the bias term

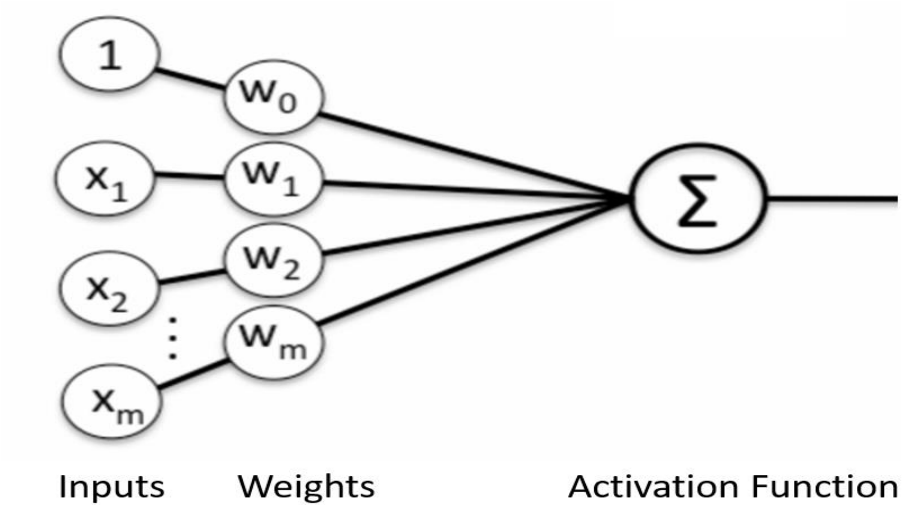


Figure 2: Pocket Perceptron (Ref[20])

SVM:

SVM, as a widely applied supervised machine learning classification technique in cancer diagnosis and prognosis, stands out for its ability to create an optimal separation between benign and malignant tumors within a feature space defined by critical characteristics like tumor size, shape, and medical imaging features. Through the selection of pivotal samples known as support vectors, SVM constructs a linear function that maximizes the separation between these classes. This involves mapping input vectors into a higher-dimensional space to identify the most suitable hyperplane for class division. The primary objective is to maximize the margin, i.e., the distance between the decision hyperplane and the nearest data points. This unique approach, which heavily relies on support vectors closest to the decision boundary, ensures robust classification, as their removal has a more significant impact on the boundary than other distant data points. SVM's efficacy in cancer diagnosis hinges on its unparalleled capacity to optimize the separation of tumor classes, making it a key asset in the field of medical diagnostics.[2][4]

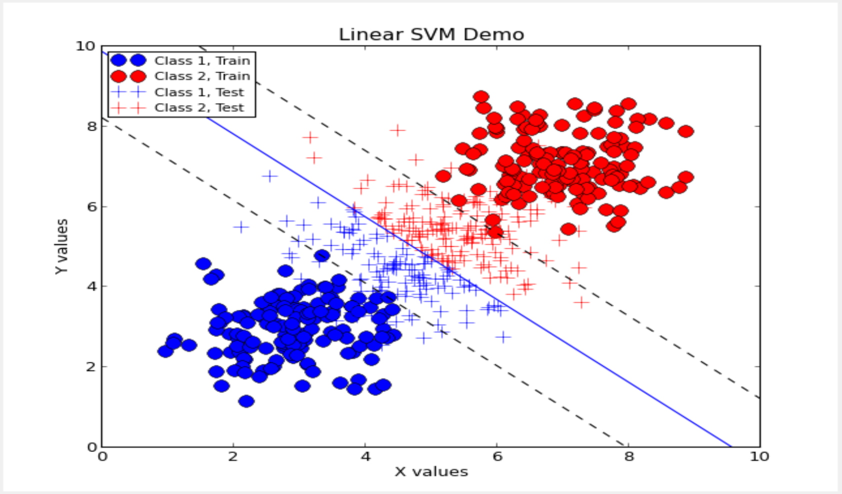


Figure 3: SVM (Ref [21])

Naive Bayes:

Naive Bayes, a probabilistic machine learning algorithm, finds application in breast cancer diagnosis by estimating the likelihood of a tumor's class (benign or malignant) based on certain observed features. This method relies on Bayes' theorem, which calculates the posterior probability of a tumor class given its characteristics. For instance, it computes the probability of a tumor exhibiting specific traits, such as microcalcifications or irregular borders, given its class. The probabilistic approach of Naive Bayes can be expressed through the formula:

P(Class | Features) = (P(Features | Class) * P(Class)) / P(Features)

Where:

P(Class | Features) is the posterior probability of the tumor class given its features.

P(Features | Class) represents the likelihood of observing the features given the class.

P(Class) is the prior probability of the tumor class.

P(Features) denotes the overall probability of observing the given features.

Random Forest:

Random Forest technique offers an effective solution for the intricate task of breast cancer diagnosis. It stands out by considering a diverse range of features, including mammographic findings, clinical data, and genetic markers.[2] What sets Random Forest apart is its ability to create an ensemble of decision trees, much like a panel of jurors in a courtroom. This collective approach results in a well-balanced and resilient model that is less susceptible to noise and fluctuations in the dataset, which is particularly crucial when working with imbalanced cancer data, where malignant cases may represent only a fraction of the overall data.

The Random Forest method operates through a recursive and iterative process. At each step, it randomly selects a subset of the dataset with replacement and another random subset of predictor variables without replacement. These selections are used to construct multiple decision trees. By creating this ensemble of trees, Random Forest combines their individual strengths, offering a more comprehensive and robust evaluation of tumor characteristics. The final classification of breast tumors is determined by a majority vote across these decision trees.

FinalPrediction(x) = Mode(T1(x), T2(x), ..., Tn(x))

Where:

FinalPrediction(x) represents the final classification or prediction for the input data point x.

T1(x), T2(x), ..., Tn(x) are the individual predictions of each decision tree for the input data point x.

Mode calculates the most frequently occurring prediction among the decision trees.

K Nearest Neighbor:

K-Nearest Neighbors (KNN) is known for its proximity-based classification, and offers valuable insights into breast cancer diagnosis by evaluating the resemblance of tumors to their neighboring counterparts. This method employs essential features such as tumor size (S), shape (Sh), and additional characteristics to calculate distances between tumors within the feature space. The distance metric, often based on Euclidean distance, quantifies the dissimilarity or proximity between tumors. Tumors situated in close proximity to one another in this feature space are more likely to be categorized within the same class, furnishing a significant aid in the accurate diagnosis of breast cancer.[18]

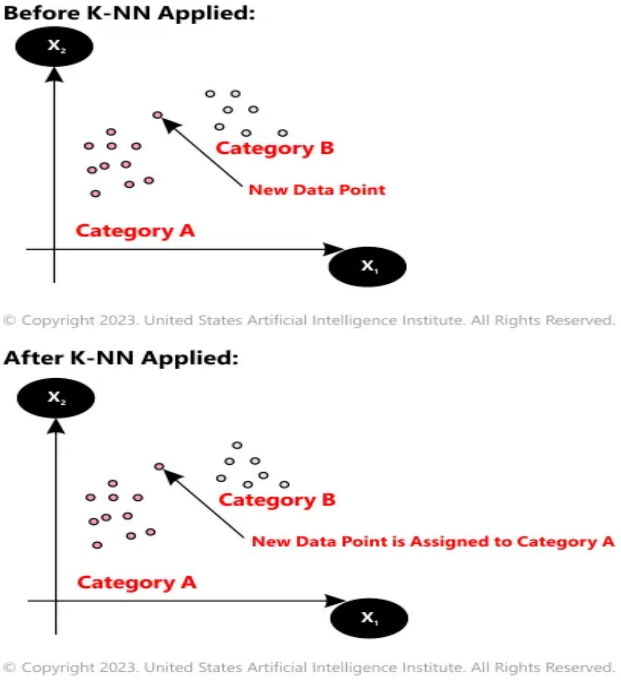
Mathematically, the KNN algorithm computes the distance between a query point (x) and other data points (xi) to identify the k-nearest neighbors, which are crucial for classification. The formula for calculating the distance, typically the Euclidean distance, can be expressed as:

$\text{Distance}(x, xi) = \sqrt{((S - Si)^2 + (Sh - Shi)^2 + ...)}$

Where:

Distance(x, xi) signifies the distance between the query point (x) and a specific data point (xi).

S, Sh, and other variables represent the respective features of the query point (x) and data point (xi).



© Copyright 2023. United States Artificial Intelligence Institute. All Rights Reserved.

© Copyright 2023. United States Artificial Intelligence Institute. All Rights Reserved.

Figure 4: KNN(Ref [23])

IV. RESULTS AND DISCUSSION

The results and discussion may be combined into a common section or obtainable separately. They may also be broken into subsets with short, revealing captions. An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it. This section should be typed in character size 10pt Times New Roman.

Before performing the comparative study of various classifier models, the Breast Cancer Wisconsin dataset was preprocessed to ensure the quality and consistency of the data. The dataset was split into training and testing sets, and standardization was applied to the features to ensure that all features were on the same scale. To assess the performance of the different classifier models, several evaluation metrics were used, including accuracy, precision, F1-score, and the Receiver Operating Characteristic (ROC) curve. These metrics provide a comprehensive view of the models' capabilities in distinguishing between malignant and benign tumors.

The results obtained for each algorithm in terms of accuracy, precision, F1 value, and ROC (AUC) are summarized below:

Table 1. Comparison of performance metrics for each algorithm

Model	Pocket perceptron	SVM	Naive Bayes	Random Forest	KNN
Accuracy	0.956	0.97	0.92	0.92	0.938
Precision	0.956	1.0	0.94	0.95	1.0
F1 Value	0.897	0.96	0.94	0.91	0.930
ROC (AUC)	0.97	0.96	0.89	0.88	0.95

Accuracy and Precision:

Classifier accuracy is a metric that measures classifier's efficacy in accurately predicting the instances into their respective categories. It's calculated as the number of correct predictions divided by the total number of instances in the dataset. While accuracy provides a general sense of how well the classifier is performing, it may not be the most suitable metric for comparing different classifiers. Precision primarily focuses on the classifier's accuracy in classifying positive instances, making it a crucial metric for evaluating the classifier's ability to avoid false positives.[2][7][10]

Accuracy = {(TP + TN) / (TP + TN + FP + FN) }× 100%

Precision = {(TP) / (TP + FP) }× 100%

where,

TP represents True Positives, the cases correctly classified as positive.

TN represents True Negatives, the cases correctly classified as negative.

FP represents False Positives, the cases incorrectly classified as positive.

FN represents False Negatives, the cases incorrectly classified as negative.

Support Vector Machine (SVM) achieved the highest accuracy (0.97) and precision (1.0) among all the algorithms. This indicates that SVM performed exceptionally well in correctly classifying instances and minimizing false positives.

F1 Value:

SVM also achieved the highest F1 value (0.96). This indicates a good balance between precision and recall. It implies that SVM can provide both high precision and recall.

ROC (AUC):

Pocket Perceptron achieved the highest ROC AUC score (0.97), which points out its strong ability to distinguish between positive and negative classes. SVM and KNN also demonstrated good ROC AUC scores (0.96 and 0.95, respectively), indicating high discriminative power.[12][2]

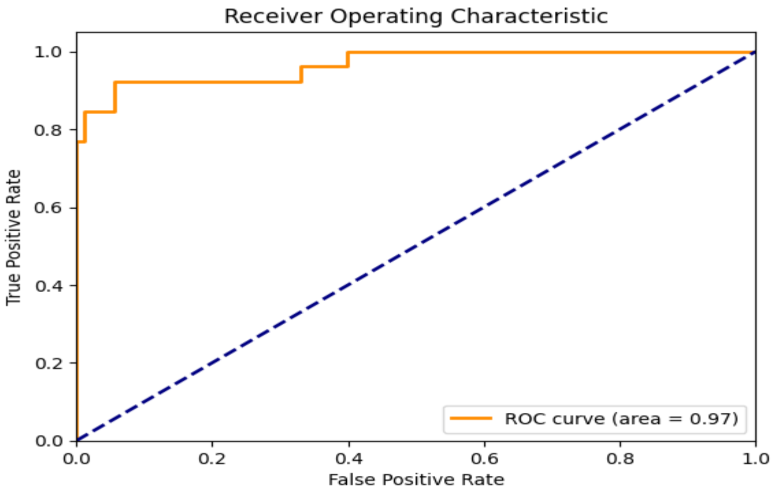


Figure 5: ROC curve and AUC for Pocket Perceptron

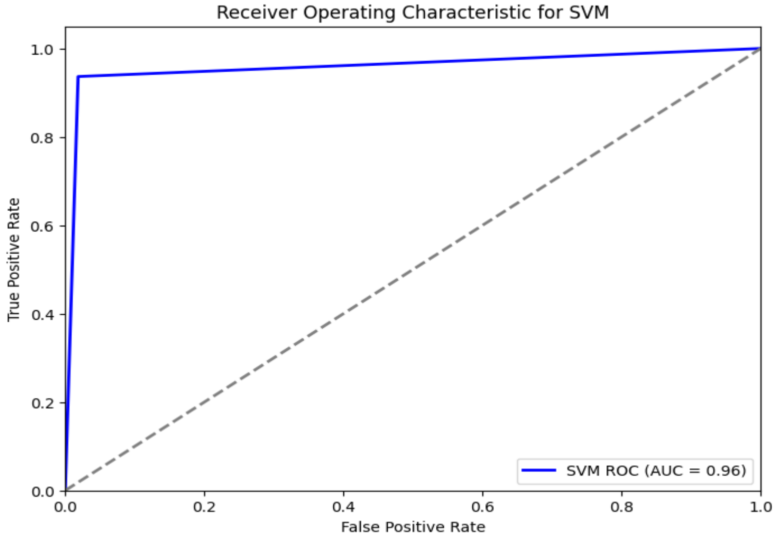


Figure 6: ROC curve and AUC for SVM

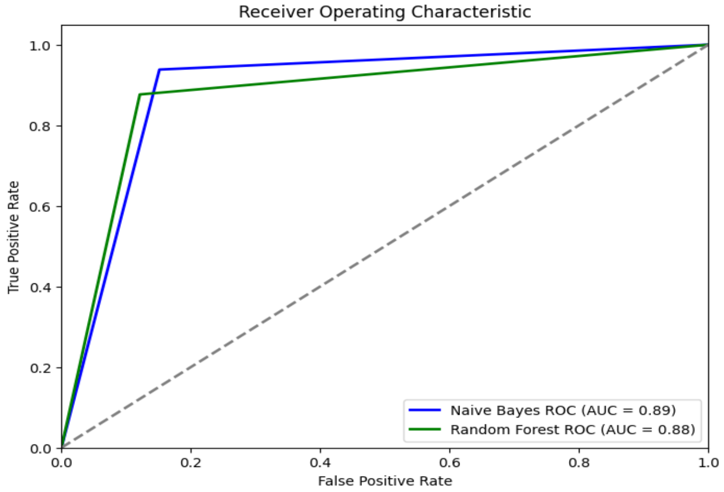


Figure 7: ROC curve and AUC for Naive Bayes and Random Forest

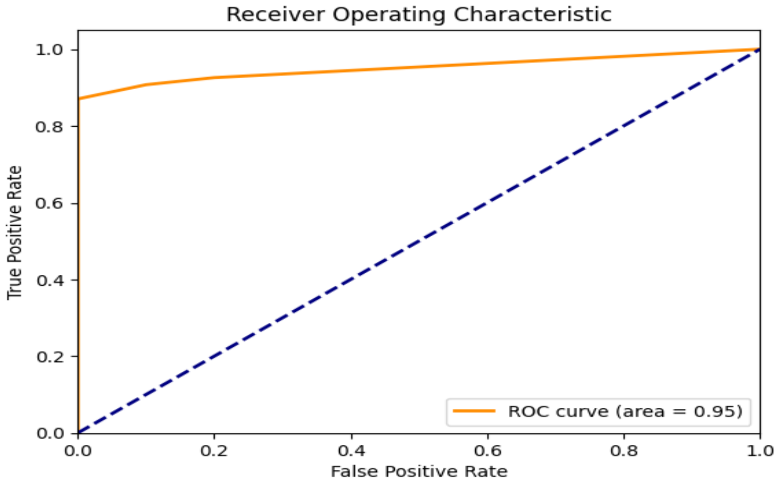


Figure 8: ROC curve and AUC for KNN

V. CONCLUSION

Machine learning techniques have garnered extensive adoption in diagnosis and treatment in the medical field. This study has shed light on five of the most prevalent machine learning classifiers employed in the context of breast cancer detection and diagnosis, namely Pocket Perceptron, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Random Forests (RF) and Naive Bayes. SVM appears to be the top performer in this study, achieving highest accuracy (0.97),highest F1 value (0.96) and precision (1.0). Naive Bayes and Random Forest achieved moderate results in accuracy and precision. While they performed reasonably well, they were outperformed by SVM, Pocket Perceptron, and KNN in these metrics.

VI. REFERENCES

[1] Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

[2] Shubair, Raed. (2016). Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis. 10.1109/ICEDSA.2016.7818560.

[3] "Breast cancer," WHO. Online [Available]: https://www.who.int/news/item/03-02-2021-breast-cancer-now-most-common-form-of-cancer-who-taking-action (accessed Feb. 18, 2022).

[4] Chaurasiya, Satish & Rajak, Ranjit. (2023). Comparative Analysis of Machine Learning Algorithms in Breast Cancer Classification. Wireless Personal Communications. 131. 1-10.
10.1007/s11277-023-10438-9.

[5] Perets, Tatap. (2023). Machine Learning-based Breast Cancer Detection: A Case Study.

10.13140/RG.2.2.32167.73124.

[6] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, Procedia Computer Science, Volume 83, 2016, Pages 1064-1069, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.04.224

[7] Hossin, Md & Shamrat, F. & Bhuiyan, Md Rifat & Hira, Rabea & Khan, Tamim & Molla, Shourav. (2023). Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset. Bulletin of Electrical Engineering and Informatics. 12. 2446-2456.

10.11591/beei.v12i4.4448.

[8] Alsudani, Mustafa & Fakhruldeen, Hassan Falah & Yousif, Israa. (2023). Comparative evaluation of data mining algorithms in breast cancer. Indonesian Journal of Electrical Engineering and Computer Science. 31. 777-784. 10.11591/ijeecs.v31.i2.pp777-784.

[9] S, Bhuvaneswari & S, Karthikeyan. (2023). A Comprehensive Research of Breast Cancer Detection Using Machine Learning, Clustering and Optimization Techniques. 1-6.

10.1109/ICDSNS58469.2023.10245164.

[10] Strelcenia, Emilija & Prakoonwit, Simant. (2023). Effective Feature Engineering and Classification of Breast Cancer Diagnosis: A Comparative Study. BioMedInformatics. 3.

10.3390/biomedinformatics3030042.

[11] Madeswaran, Tamil & Kavuru, Aruna & Theagarajan, Padma & Hadhrami, Nasser & Foori, Maya & Rambabu, Ohm. (2023). An Intelligent System for Predicting the Breast Cancer Threat Using Health Data Registry and Awareness: A Review. European Journal of Engineering and Technology Research. 8. 17-22. 10.24018/ejeng.2023.8.3.3012.

[12] Sunday Samuel, Olofintuyi. (2023). BREAST CANCER DETECTION WITH MACHINE LEARNING APPROACH. FUDMA JOURNAL OF SCIENCES. 7. 216-222. 10.33003/fjs-2023-0702-1392.

[13] Sandhu, Jasjeet Kaur & Kaur, Amandeep & Kaushal, Chetna. (2022). Analysis of Breast Cancer in Early Stage by UsingMachine LearningAlgorithms: A Review. 1-7. 10.1109/CCET56606.2022.10080757.

[14] Mohammed, Siham & Darrab, Sadeq & Noaman, Salah & Saake, Gunter. (2020). Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. 10.1007/978-981-15-7205-0_10.

[15] Vashist, Apurva & Sagar, Anil & Goyal, Anjali. (2023). Breast Cancer Detection by Using Decision Tree. 10.1007/978-3-031-45121-8_7.

[16] Tilwankar, Sarthak & Kirar, Bhupendra. (2021). Breast Cancer Detection using Principal Component Analysis and Machine Learning Models. 80-84. 10.1109/ICACFCT53978.2021.9837342.

[17] Ghiasi, Mahdi & Zendehboudi, Sohrab. (2020). Application of Decision Tree-Based Ensemble Learning in the Classification of Breast Cancer. Computers in Biology and Medicine. 128. 104089.

10.1016/j.compbiomed.2020.104089.

[18] MurtiRawat, Ram & Panchal, Shivam & Singh, Vivek & Panchal, Yash. (2020). Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning. 534-540.

10.1109/ICESC48915.2020.9155783.

[19] Chiu, Huan-Jung & Li, Tzuu-Hseng & Kuo, Ping-Huan. (2020). Breast Cancer–Detection System Using PCA, Multilayer Perceptron, Transfer Learning, and Support Vector Machine. IEEE Access. 8. 204309-204324. 10.1109/ACCESS.2020.3036912.

[20] https://medium.com/@Rghv_Bali/perceptron-where-it-all-started-55d3508e38af

[21] https://www.linkedin.com/pulse/machine-learning-basics-support-vector-machines-amsal-gilani/

[22] Hernández-Julio, Y.F.; Díaz-Pertuz, L.A.; Prieto-Guevara, M.J.; Barrios-Barrios, M.A.; Nieto-Bernal, W. Intelligent Fuzzy System to Predict the Wisconsin Breast Cancer Dataset. Int. J. Environ. Res. Public Health 2023, 20, 5103. https://doi.org/10.3390/ijerph20065103

[23] https://www.usaii.org/ai-insights/understanding-knn-algorithm-and-its-role-in-machine-learning