Contents lists available at *Dergipark*

Journal of Scientific Reports-B

journal homepage: *https://dergipark.org.tr/en/pub/jsrb*

E-ISSN: 2717-8625 ***Number 12, April 2025***

RESEARCH ARTICLE

Receive Date: 24.04.2025 *Accepted Date: 30.04.2025*

Performance of machine learning methods on breast cancer prediction

Ghazwa Alsaffaf [a], Soydan Serttaş [b,*]

[a]*Kütahya Dumlupınar University, Department of Computer Engineering, 43000, Kütahya, Türkiye, ORCID:0000-0001-9824-5951*
[b]*Kütahya Dumlupınar University, Department of Computer Engineering, 43000, Kütahya, Türkiye, ORCID:0000-0001-8887-8675*

Abstract

In the last 50 years, the effect of cancer disease on the annual number of deaths has increased significantly. This has led to an increase in research on early detection and diagnosis of cancer. Early diagnosis of cancer increases the chance of surviving the disease and reduces the possibility of recurrence of the disease. The technological advances in artificial intelligence and machine learning are used to analyse patient data, while at the same time reducing the likelihood of developing diseases. In this paper, 7 different machine learning algorithms commonly used in the literature are used for breast cancer diagnosis. These are: Logistic Regression (LR), K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Radial Basis Function (RBF) Kernel, Naive Bayes, Decision Tree (DT), and Random Forest (RF) algorithms. In our study, two separate datasets were used for breast cancer diagnosis. In the first dataset, Random Forest, SVM (RBF), and SVM (Linear) algorithms had the highest accuracy value of 96.5, while the K-Nearest Neighbours algorithm had the highest sensitivity value of 98.8, and the decision tree algorithm had the highest specificity value of 98.1. The K-Nearest Neighbour algorithm was also found to be the fastest algorithm, with 1.03 seconds. In the second dataset with different data, the K-Nearest Neighbours algorithm reached the highest accuracy value of 97.7 and was observed to be the second fastest algorithm with 1.48 seconds after the Gaussian Naive Bayes algorithm with 1.14 seconds.

* Corresponding author.
E-mail address: soydan.serttas@dpu.edu.tr

.

Keywords: Machine learning, classification algorithms, artificial intelligence, breast cancer prediction.

1. Introduction

The second most common cause of cancer-related deaths among women is breast cancer. The risk of breast cancer death for a woman is around 1 in 43, or 2.3% [1]. This study compares a number of machine learning methods for data analysis and breast cancer prediction early detection.

1.1. Breast Cancer

Breast cancer is a cancer of the breast and surrounding tissue. It is more common in women after skin cancer, but it can also affect men. After skin cancer, it is considered the most dangerous cancer in women's lives [2]. The funding of scientific research to develop treatments and early detection, as well as media coverage to increase awareness of breast cancer, have enhanced the diagnosis and treatment of the disease. According to the American Cancer Society, women will experience 49,290 new instances of ductal carcinoma in situ (DCIS) and 281,550 new cases of invasive breast cancer in 2021. The lifetime risk of breast cancer is 13% for women, and between 2010 and 2022, the annual death rate from breast cancer dropped by 1.2% [1].

1.2. Artificial Intelligence and Machine Learning

Artificial intelligence is the attempt to simulate human intelligence through devices, and in most cases, the device used is a computer. Artificial intelligence relies on three cognitive abilities to emulate human intelligence: learning, reasoning, and self-correction (learning from mistakes). Artificial intelligence works by analysing data and creating rules for analysing that data to derive a possible benefit from it. Machine learning: The term machine learning emerged when scientists wanted to know about the ability of computers to learn from data [3].

Machine learning is tested by inputting new data and testing its ability to reach correct results, and the computer learns from previous data [4]. After the technological revolution, the increasing impact of artificial intelligence (AI) and the importance of machine learning (ML) and artificial intelligence in our lives are having a pioneering way, especially in the treatment and diagnosis of diseases. Advances in machine learning have helped diagnose diseases by using large data sets to detect diseases early, especially in chronic diseases such as cancer [5].

1.3. Disease Identification

The machine learning method allows us to create models relating multiple variables to a disease. Machine learning algorithms analyse data, identify correlations between variables, and display the results. Clinicians now have access to vast amounts of data, including clinical symptoms, biochemical assays, and imaging device outputs, all of which are incorporated into machine learning models. There are several valuable data types to make an accurate medical diagnosis using machine learning, such as disease, environmental, and genetic data. It also has many benefits in research on risk factors and increases the efficiency of diagnosis [6-7].

2. Literature Review

In Jacob and Ramanai’s study, they compared the performance of various classification algorithms. The best algorithms were random forest and decision trees with 100% accuracy [8].

Abyan Farid Agharib compared six machine learning algorithms to analyze data of breast cancer patients to help diagnose the disease. Algorithms used: Linear regression, multilayer perception, nearest neighborhood, search,

softmax regression, support vector machine. Algorithms are compared based on their test accuracy, sensitivity, and specificity values. All algorithms showed a success of more than 90%, and the best result of the MLP algorithm showed an accuracy of 99.04% [9].

Sengar and others compared two machine learning algorithms. Using the Wisconsin diagnostic dataset, the same dataset we used in this paper, they compared the test accuracy of the logistic regression algorithm and the decision tree algorithm. Both algorithms showed more than 90% success results, showing the superiority of the decision tree algorithm with a 100% accuracy rate [10].

Jain and others used five classification algorithms to classify the type of breast cancer, K-Nearest Neighbor, Logistic Regression, Random Forest, SVM, and Decision Tree. With 96.52% and 98% eloquent effectiveness, Logistic Regression and K-Nearest Neighbor were the best indicators [11].

Ojha and Goel used a total of eight algorithms, four of which are classification algorithms: KNN, SVM, Naive Bayes, and C5.0, which is the algorithm used in data mining as a decision tree classifier that can be employed to generate a decision. Four of them are clustering algorithms, which are K-means, Expectation Maximization, Partitioning around Medoids, and Fuzzy c-means. C5.0 and SVM classifiers were the best prediction algorithms with an accuracy of 0.813, while the fuzzy mean clustering algorithms came out worse with an accuracy of 0.3711 [12].

In Özkan and Gündüz's study, they utilize the Breast Cancer Database, the exhibition of AI calculations in anticipating the shot at endurance following bosom malignant growth was investigated Surveillance Epidemiology and End Result (SEER). The algorithms used: Naive Bayes, J48 algorithm is used to classify different applications and performs accurate results of the classification, SVM and Multiobjective and Evolutionary Fuzzy Classifier (MEFC), the J48 algorithm showed a success rate of 93.02 and a speed of 84.39 seconds, which is considered the second fastest algorithm used [13].

In Kıyan and Yildirim's study, they tested diagnosing breast cancer using structural neural networks' performance, comparing the accuracy of different structural neural networks. Radial Basis Functions, Probabilistic Neural Networks, Generalized Regression Neural Networks (GRNN), The RBF and PNN structures showed the highest rate with 100% training accuracy. The GRNN structure test result showed the highest accuracy rate of 98.8%.

Based on the overall findings, GRNN appears to be the best neural network model for WBCD data classification [14].

Hazra and others' study showed that the Naive Bayes algorithm produces the highest accuracy with an average of 97.3978% with only five dominant features and a time of 0.102023 ms, which is the fastest algorithm comparing the other two classifiers (Support Vector Machine and Ensemble) [15].

Abdulla and others compared five machine learning algorithms. The SVM algorithm achieved the highest accuracy rate of 97% when combined with other algorithms such as Random Forest, Naive Bayes, and KNN. Convolutional neural networks (CNNs) the Deep Learning algorithm has reached 98% accuracy [16].

Shravya and others focused on creating prescient models to accomplish a decent rate utilizing regulated AI techniques. As a result of the comparison of three algorithms, k-nearest neighbor, logistic regression and SVM, the SVM algorithm gave the best result for breast cancer prediction with the highest accuracy rate of 92.7% [17].

Al-Azzam and Shatnawi compared the effectiveness and accuracy of supervised learning (SL) and semi-supervised learning (SSL) algorithms for breast cancer detection. SSL requires less data and is less expensive than SL. As a result of this study, SSL algorithms are almost as accurate as SL algorithms, where the predictions were correct for all malignant and benign tumours with a rate between 91% and 98%. KNN algorithms (SL = 98.4% & SSL = 97.4%) and logistic regression (SL = 97% & SSL = 98.4%) produced the best result. It is possible to replace supervised learning algorithms with semi-supervised learning algorithms [18].

Darwich and Islam provide suggestions for further study after weighing the advantages and drawbacks of every machine learning technique and dataset [19-20].

3. Material and Method

3.1. Machine Learning Classification

Logistic Regression is a fundamental classification method and is one of the quickest and uncomplicated classifications, and is convenient for interpreting results. It can apply to multiclass problems since it is a binary classification algorithm [21].

One kind of supervised machine learning technique is the K-Nearest Neighbors algorithm. KNN can handle challenging classification tasks and is simple to implement. Considering that there is no exceptional training phase, it is a lazy learning algorithm. It is a nonparametric learning algorithm, meaning that it has no prior knowledge or assumption about the underlying data. KNN algorithm needs more memory and more time to scan all data points [22]. Support vector machines outperform other classifiers like logistic regression and decision trees in terms of accuracy. Gene classification, handwriting recognition, facial identification, intrusion detection, email categorization, news articles, and web pages are just a few of the many uses for it. SVM is an algorithm with rather straightforward ideas. An SVM classifier is also referred to as a different classifier since it uses the hyperplane with the largest margin to separate the data points. Performance is improved by linear SVM training, which is quicker than non-linear ones (such as the RBF kernel) [23-25].

Naive Bayes is the simplest and fastest classification algorithm suitable for a large portion of data. A naive Bayes classifier is used in spam filtering, text classification, and recommendation systems. The classifier trains the model on a particular dataset and measures its functioning in the learning phase, and performance is evaluated based on various criteria such as accuracy, error, and recall [26].

An internal node's decision tree property, in which each leaf node indicates the outcome and the branch reflects a decision rule and a tree structure resembling a flowchart. This framework, which resembles a flowchart, aids in decision-making. Similar to a flowchart diagram, visualization readily imitates human-level thought processes. Decision trees are, therefore, simple to comprehend and analyze. Compared to the neural network algorithm, the training period is quicker. It is a non-parametric or distribution-independent approach that is independent of assumptions about probability distributions. High-dimensional data can be accurately processed by decision trees [27-28].

A supervised learning technique used for regression analysis or classification is called random forests. In contrast to other algorithms, it is versatile and simple to use. Random data samples are used to make the decision; each tree is estimated, and the best outcome is chosen by voting. It can be used for many things, such as feature selection and image rating. The choice is based on a divide-and-conquer strategy and uses a tree-clustering method (randomly partitioned data set) [29].

3.2. Datasets

The first dataset, named the Diagnostic Wisconsin Breast Cancer Database, contains information about breast cancer patients, determining whether their cancer diagnosis is malignant or benign, as prepared by researchers at the University of Wisconsin with expertise in databases and general surgery [30]. This dataset is widely used in breast cancer diagnosis using machine learning and statistical analysis techniques. The dataset consists of 569 samples in total. Of these samples, 212 represent malignant tumours (Malignant - M) and 357 represent benign tumours (Benign - B). The dataset contains 30 numerical features for each sample, in addition to 1 target variable (diagnosis) and 1 ID column containing the patient identification number.

The second dataset, named the Original Wisconsin Breast Cancer Database, is also a widely used dataset for breast cancer diagnosis [31]. The dataset, containing 699 samples in total, consists of 10 columns, each with 9 numerical features and 1 class label (benign or malignant). The features include morphological measurements

obtained from microscopic images of cells. Each feature is rated on a scale from 1 to 10. The class label indicates whether the tumour is benign or malignant.

3.3. Evaluation Metrics

Three evaluation metrics were used to compare the performance of the algorithms: accuracy, sensitivity, and specificity. Accuracy is one of the evaluation measures used to evaluate a classification model's overall performance. The ratio of projected samples to total samples is known as accuracy. When the distribution of classes is balanced, it is ideal [32]. Sensitivity indicates how accurately the model predicts examples belonging to the positive class. This metric expresses the rate at which the model correctly recognizes examples belonging to the positive class. It is especially used in applications where correctly detecting examples belonging to the positive class is important [33]. The rate at which the model accurately identifies instances from the negative class is known as specificity. This indicator demonstrates how well the model forecasts instances from the negative class. In applications where accurately identifying examples from the negative class is crucial, it is particularly utilized [33].

4. Results

As a result of pre-processing the breast cancer data set from the first database, a data set consisting of 30 traits and class values contains 18240 records. The second database provides ten items and 6990 record-classification assessments. We did the experiments and described them in the Jupyter notebook using the Python programming language. We separated the dataset into 75% for training and 25% for testing. In the first dataset, the number of records in the training set is 13680, and the amount of data used in the test set is 4560. In the second dataset, the number of records in the training set is 5243, and the amount of data used in the test set is 1748. The data set was analyzed using logistic regression, K-Nearest Neighbors, Naive Bayes, decision tree, random forest, and SVM. Each of the algorithms gave results with different success rates. In the first database, SVM Linear, SVM RBG, and random forests achieved an equal accuracy rate of 96.5%. The algorithm with the lowest rate was the Naive Bayes calculation with a pace of 92.3%. In the second database, K-Nearest Neighbor had the highest accuracy rate of 97.7%, and the algorithm with the lowest accuracy rate was the Decision Tree Classifier algorithm with 93.7%. Algorithms' run times vary widely in the application. We take the average of running seven algorithms. The fastest learning algorithm for the first database is the K-Nearest Neighbor algorithm with 1.03 seconds, and the slowest algorithm is the Random Forest Classifier algorithm with 33.43 seconds. The Gaussian Naive Bayes algorithm is the fastest in the second database with 1.14 seconds, and the logistic regression algorithm is the slowest with 16.67 seconds. In the following table, you can see the success rates of the algorithms and the average timing for each algorithm.

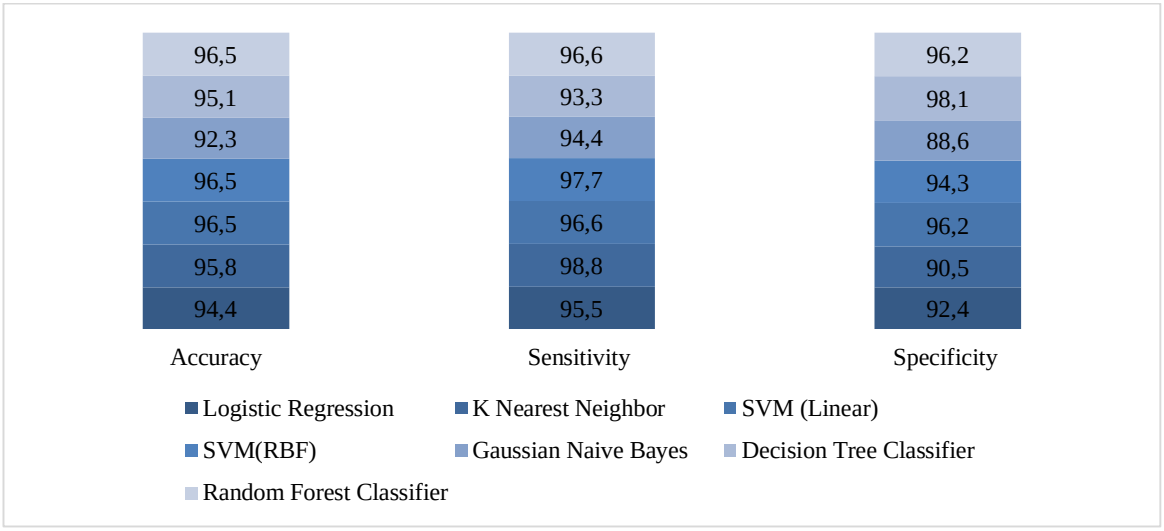

Fig. 1. Evaluation results for the first dataset.

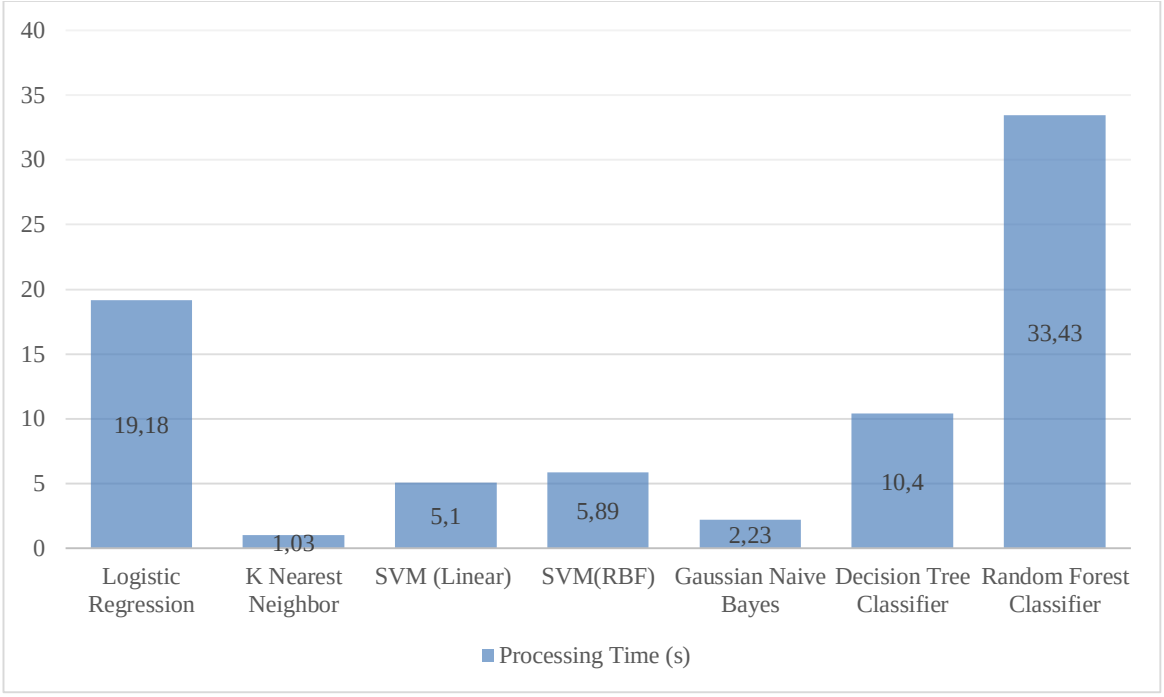

Fig. 2. Processing time results for the first dataset.

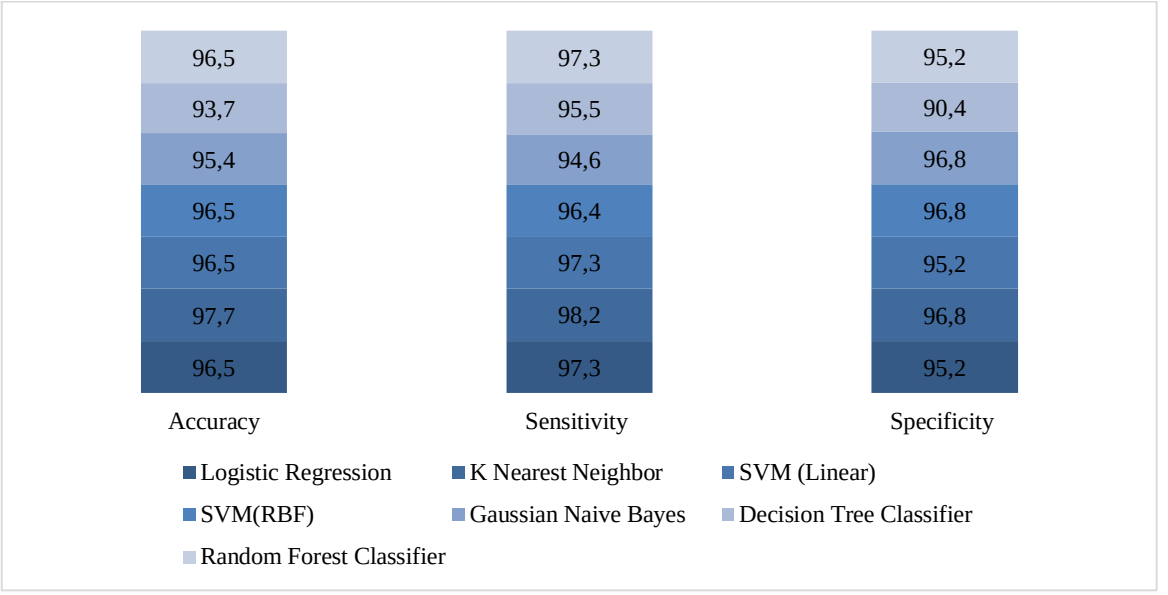

Fig. 3. Evaluation results for the second dataset.

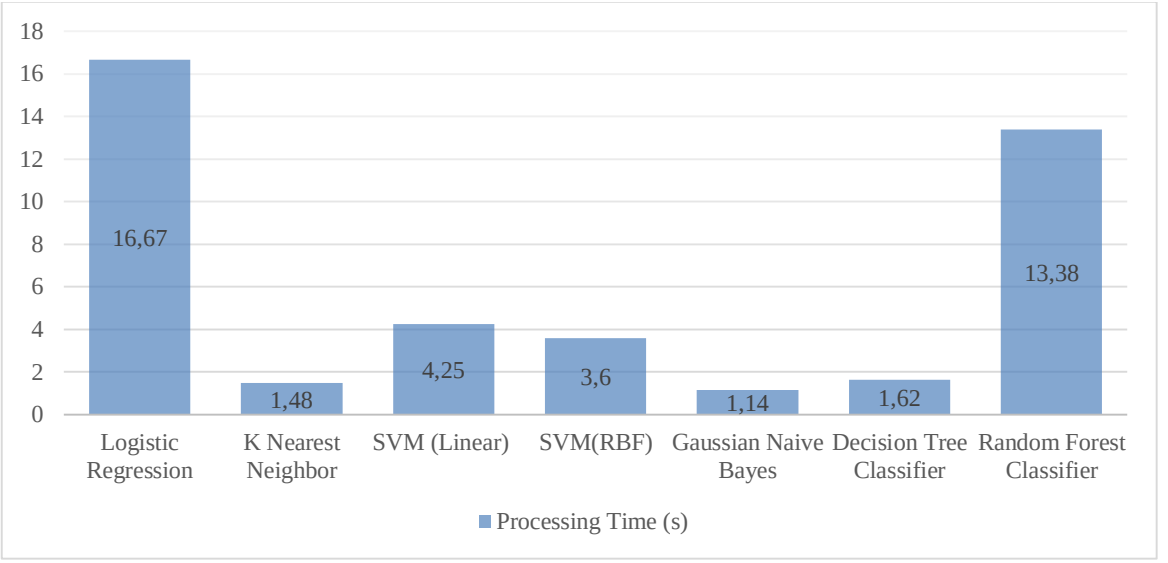

Fig. 4. Processing time results for the second dataset.

4. Conclusion

Breast cancer early detection is necessary to reduce mortality and increase the likelihood of recovery. This paper aims to apply the classification algorithms that can help identify breast cancer characteristics, clean the data, and identify traits that can help predict breast cancer. As a result of this study, the first data set, we found that the linear and non-linear SVM algorithms and the random forest algorithms nearby to the highest accuracy of 96.5%. For the second dataset, we found that the K-Nearest Neighbors algorithm achieved the highest accuracy of 97.7%. With this result, we conclude that it is possible to predict the likelihood of developing breast cancer using artificial intelligence and early detection, which helps in treatment and early prevention.

Acknowledgements

This study did not receive any dedicated funding from public, commercial, or non-profit organizations.

References

[1] American Cancer Society. Breast Cancer Facts & Figures 2024-2025. Atlanta: American Cancer Society, Inc. 2024, https://www.cancer.org/ (accessed Feb. 1, 2025).
[2] W. J. Archibald, R. E. Ziemer, J. S. Newman, "Ask mayo expert: Anemia workup in 1919," *Mayo Clinic Proceedings*, vol. 94, no. 9, pp. 1904, 2019.
[3] B. J. Copeland. "Artificial intelligence." https://www.britannica.com/technology/artificial-intelligence (accessed Dec. 14, 2024).
[4] "Machine learning." https://www.sas.com/en_us/insights/analytics/machine-learning.html (accessed Sep. 3, 2024).
[5] T. Davenport, R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare Journal*, vol. 6, no. 2, pp. 94-98, 2019.
[6] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Journal of Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89-109, 2001.
[7] J. Kiruba, R. Visalakshi, A. Vaishnavi, R. Ahalya, R. A. Keerthi, "Medical diagnosis using machine learning," *Indian Journal of Public Health Research and Development*, vol. 10, no. 4, pp. 1337, 2019.
[8] S. G. Jacob, R. G. Ramani, "Efficient classifier for classification of prognostic breast cancer data through data mining techniques," *World Congress on Engineering and Computer Science*, San Francisco, USA, 2012, vol. 1, pp. 978-988.
[9] Agarap A F M. "On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset," *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (ICMLSC'2018)*, Phu Quoc Island, Vietnam, 2018, pp. 5-9.
[10] P. P. Sengar, M. J. Gaikwad, A. S. Nagdive, "Comparative study of machine learning algorithms for breast cancer prediction," *Third International Conference on Smart Systems and Inventive Technology (ICSSIT'2020)*, Tirunelveli, India, 2020, pp. 796–801.
[11] T. Jain, V. K. Verma, M. Agarwal, A. Yadav, A. Jain, "A supervised machine learning approach for the prediction of breast cancer," *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Pondicherry, India, 2020, vol. 10, pp. 1-6.
[12] U. Ojha, S. Goel, "A study on prediction of breast cancer recurrence using data mining techniques," *7th International Conference on Cloud Computing, Data Science & Engineering Confluence*, Noida, India, 2017, pp. 527-530.
[13] G. Y. Özkan, S. Y. Gündüz, "Comparision of classification algorithms for survival of breast cancer patients," *Innovations in Intelligent Systems and Applications Conference (ASYU'20)*, Istanbul, Turkey, 2020, pp. 1-4.
[14] T. Kıyan, T. Yıldırım, "Breast cancer diagnosis using statistical neural networks," *Istanbul University Journal of Electrical & Electronics Engineering*, vol. 4, no. 2, pp. 1149-1153, 2004.
[15] A. Hazra, S. Kumar, A. Gupta, "A study and analysis of breast cancer cell detection using naïve Bayes, SVM and ensemble algorithms," *International Journal of Computer Applications*, vol. 145, no. 2, pp. 39-45, 2016.
[16] S. H. Abdulla, A. M. Sagheer, h. Veisi, "Breast cancer classification using machine learning techniques: A review," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 14, pp. 1970-1979, 2021.
[17] C. Shravya, K. Pravalika, S. Subhani, "Prediction of breast cancer using supervised machine learning techniques," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 6, pp. 2278-3075, 2019.
[18] N. Al-Azzam, I. Shatnawi, "Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer," The journal of Annals of Medicine and Surgery, vol. 62, pp. 53-64, 2021.
[19] M. Darwich, M. Bayoumi, "An evaluation of the effectiveness of machine learning prediction models in assessing breast cancer risk," *Informatics in Medicine Unlocked*, vol. 49, 101550, 2024.
[20] T. Islam, M. A. Sheakh, M. S. Tahosin, et al., "Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI," *Scientific Reports*, vol. 14, article 8487, 2024.
[21] M. Stojiljković. "Logistic regression in Python." https://realpython.com/logistic-regression-python/ (accessed Jan. 13, 2024).

[22] C. Sampaio. “Guide to the K-nearest neighbours algorithm in python and scikit-learn.” https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/ (accessed Feb. 02, 2025)
[23] J. C. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Journal of Advances in Large Margin Classifiers*, pp. 1-11, 2000.
[24] Quora. “When can I use Linear SVM instead of RBF, polynomial, or a sigmoid kernel?” https://www.quora.com/When-can-I-use-Linear-SVM-instead-of-RBF-polynomial-or-a-sigmoid-kernel (accessed Dec. 8, 2024).
[25] Z. Anw. “Difference between SVM Linear, polynmial and RBF kernel?” https://www.researchgate.net/post/Diffference_between_SVM_Linear_polynmial_and_RBF_kernel. (accessed Dec. 8, 2024).
[26] A. Navlani. “Naive Bayes Classification using Scikit-learn.” https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn (accessed Feb. 8, 2025).
[27] A. Navlani. “Decision Tree Classification in Python.” https://www.datacamp.com/community/tutorials/decision-tree-classification-python. (accessed Dec. 8, 2024).
[28] M. N. Dumont, R. Marée, L. Wehenkel, P. Geurts, “Fast multi-class image annotation with random subwindows and multiple output randomized trees,” *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, Lisboa, Portugal, 2009, vol. 2, pp. 196-203.
[29] A. Navlani. “Understanding Random Forests Classifiers in Python.” https://www.datacamp.com/community/tutorials/random-forests-classifier-python. (accessed Dec. 8, 2024).
[30] W. Wolberg, O. Mangasarian, N. Street, W. Street. Breast Cancer Wisconsin (Diagnostic) [Dataset], UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B (accessed Sep. 8, 2024).
[31] W. Wolberg. Breast Cancer Wisconsin [Dataset], UCI Machine Learning Repository. https://doi.org/10.24432/C5HP4Z (accessed Sep. 8, 2024).
[32] G. M. Foody, “Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient,” *PLoS ONE*, vol. 18, no. 10, 2023.
[33] H. H. Rashidi, S. Albahra, S. Robertson, N. K. Tran, and B. Hu, “Common statistical concepts in the supervised Machine Learning arena,” *Frontiers in Oncology*, vol. 13, 2023.