Comparing the Performance of various Supervised Machine Learning Techniques for early Detection of Breast Cancer

Moses Kazeem Abiodun[1], Sanjay Misra[2], Joseph Bamidele Awotunde[3], Samson Adewole[4], Akor Joshua[1], Jonathan Oluranti[5]

[1]Department of Department of Computer Science, Landmark University, Omu Aran, Kwara State, Nigeria
[2]Department of Computer Science and Communication,Ostfold University College, Halden, Norway
[3]Department of Computer Science, University of Ilorin, Ilorin, Kwara State, Nigeria
[4]Global Technology Managament and Policy Research Group, Southern University and A & M College, Baton Rouge USA
[5]Center of ICT/ICE, Covenant University, Ota, Nigeria

moses.abiodun@lmu.edu.ng[1], ssopam@gmail.com[2], awotunde.jb@unilorin.edu.ng[3], akor.joshua@lmu.edu.ng[4]

Abstract. Cancer is a fatal disease that is constantly changing and affects a vast number of individuals worldwide. At the research level, much work has gone into the creation and improvement of techniques built on data mining approaches that allow for the early identification and prevention of breast cancer. Because of its excellent diagnostic abilities and effective classification, data mining technologies have a reputation in the medical profession that is continually increasing. Data mining and machine learning approaches can aid practitioners in conceiving and developing tools to aid in the early detection of breast cancer. As a result, the goal of this research is to compare different machine learning algorithms in order to determine the best way for detecting breast cancer promptly. This study assessed the classification accuracy of four machine learning algorithms: KNN, Decision Tree, Naive Bayes, and SVM in order to find the best accurate supervised machine learning algorithm that might be used to diagnose breast cancer. Naive Bayes has the maximum accuracy for the supplied dataset, according to the prediction results. This reveals that, when compared to KNN, SVM, and Decision Tree, Naive Bayes can be utilized to predict breast cancer.

Keywords: Breast Cancer, Machine Learning, Data Mining, Algorithm, Classification

1 Introduction

Cancer is a highly lethal disease that is constantly evolving and touches a vast group of individuals all around the globe. According to the World Health Organization, 8.2 million people have died from cancer (WHO). The occurrence of cancers such as colon, liver, lung, and breast cancer is common. Behavioral and food risk factors (lack of physical exercise, insufficient eating of fruits and vegetables, use of tobacco and alcohol, and a high Body Max Index (BMI)) account for about 30% of cancer-related fatalities [1]. Breast cancer is the most frequent disease in the world, with 2.26 million cases expected in 2020. (WHO, 2021). It's also the most frequent cancer in women in both developed and developing countries, and it's a huge public health concern (WHO, 2021). [2, 3].

According to a data published by National Cancer Observatory of Colombia's, the death rate for cancer of the breast in women in Colombia was 11.49 per 100,000 in 2014, with a chance of increasing over the next ten years. [4]. Breast cancer identification and prognosis provide a significant problem for researchers. Since the implementation of machine learning for breast cancer detection and prediction, which transformed the entire process, significant changes have been made. Different writers have put forth a lot of work at the research level to use revolutionary methodologies based on big data, artificial intelligence, machine learning and data mining. It is critical to notice advances in biomedical technologies, hardware, and software that enable the retrieval of data for the compilation of large quantities of data, as well as the development of computational intelligence-based algorithms for the efficient prediction and diagnosis of disease [20]. Large amounts of data play a crucial role in the development of technologies for breast cancer screening. Data mining is thought to be the discipline in charge of evaluating large amounts of information [21].

It can be utilized as a stand-in to aid decision-making for the successful and early detection of breast cancer. Supervised Machine Learning (SVM) is a statistical technique-based algorithm that allows systems to learn from data without having to be programmed. Machine Learning (ML) approaches have been utilized to advance forecasting models that assist resolution making in a variety of sectors, including biomedical and medical fields, over the years. Machine Learning can be utilized in cancer research to assess whether a tumour is cancerous or benign. The performance of these strategies can be assessed using a variety of criteria, including precision, accuracy, and recall. The goal of this study is to compare and contrast various data mining and machine learning strategies for accurate breast cancer diagnosis and detection. The remaining sections of this paper are section two, which reviews comparable works of literature, and section three, which covers the study's approach and performance indicators. Section four shows outcomes and discussion of the results. Lastly, Section five recapitulates the research outcomes and proposal for future work.

2 Review of Related Literature

Several researchers in the field of data mining, machine learning, artificial intelligence, and big data have made major contribution using various methods to provide for early and efficient detection of breast cancer. Based on the finding of several studies which are related to the study of breast cancer applying various datasets, early cancer identification increases the chances of survival by 98 percent [5]. [6] Describes a system for automatically diagnosing breast cancer that employs the association rules technique for attribute reduction and Neural Networks as a classification tool. The data set used during the validation phase, as well as during the training, was Breast Cancer data from Wisconsin. The three-fold cross-validation method was used.
According to the outcomes of the trials, the right grouping rate produced with the SVM model with AR has the maximum classification accuracy (98.00 percent) for eight characteristics and 96.14 percent for four attributes. The results suggest that the proposed method can be utilized to reduce feature space and save time during the training phase, resulting in more accurate and quick automatic classification systems. The same dataset was utilized in [7], where the authors proposed a model for predicting menacing ash cancer by combining the Naive Bayes algorithm, RBF network, and J48 algorithm. According to the data, the Naive Bayes (NB) approach is the most accurate, with a 97.3 percent accuracy, followed by the RBF network with a 96.77 percent accuracy, and finally the J48 algorithm with a 93.41 percent accuracy. During the experiment, tenfold cross validation was used.

The use of ML techniques to the Wisconsin records set for breast cancer prediction is given in [8]. Linear regression, multilayer perception MLP, GRU-SVM, NN, and SVM remained the five machine learning algorithms compared. The data set is split into two parts: 70% for training and 30% for the phase. The MLP algorithm produced the most accurate result of all the methods, with a precision of 99.04 percent. [9] Proposes a novel breast cancer screening method based on data mining techniques. The goal of this research work is to examine three categorization strategies by means of the Weka tool, which uses the IBK, SMO, and BF tree algorithms. The data set used matches to the Wisconsin Breast Cancer data set. The results suggest that the SMO had the highest accuracy level of 96.2 percent. A comparison of the Fuzzy C means and the K-means for breast cancer detection is reported in [10]. The paper compares the performance of Fuzzy C means and K-Clustering algorithms, as well as other computational measures' integration, which allow the aforementioned techniques to increase their grouping accuracy. The Fuzzy C means were found to be 97 percent accurate, while the K-means were shown to be 92 percent accurate.

A study using data mining techniques to forecast the reappearance of breast cancer is given in [11]. The research work suggested the usage of numerous arrangement algorithms such as Naive Bayes, KNN, C5.0, SVM, and EM, PAM, K-means, and Fuzzy C-means clustering methods, as well as EM, PAM, K-means, and Fuzzy C-means clustering methods. The C5.0 achieved the highest accuracy level of 81.03% in the testing. The authors of [12] used the SVM and K-NN machine learning procedures to diagnose respiratory pathologies using pulmonary acoustic signals from the RALE lung sound database, demonstrating that the K-Nearest Neighbour classifier

has a higher generalization capability than the Support Vector Machine. The K-Nearest Neighbour algorithms were found to be 98.26% accurate, while the Support Vector Machine was shown to be 92.19% accurate.

The K-Nearest Neighbour algorithms were found to be 98.26 percent accurate, while the Support Vector Machine was shown to be 92.19 percent accurate. On the Lima-diabetes dataset, a comparison study was conducted on the diagnosis of diabetes using neural networks, and it was discovered that compare with other neural network based classifiers, neural networks multilayer with the Levenberg-Marquardt (LM) algorithm excel over others. [13] Conducted a comparison study on the diagnosis of Parkinson's disease. On a dataset of 197 Parkinson's disease patients, the outcomes of Decision Tree, DM Neural, Neural Network, and Regression classification models were compared. In total, twenty-two criteria were examined, and neural networks outperformed the rest of the classification techniques with an accuracy of 92.9 percent. [14] Utilizes SEER cancer data to produce two comorbid data sets: one for breast and female genital cancers, and the other for prostate and urinal cancers.

3 Methodology

To forecast the development of breast cancer, this research presents a methodology based on the supervised machine learning algorithms Decision Tree (DT)[16], K-Nearest Neighbour (K-NN)[15], Naive Bayes (NB) and Support Vector Machine (SVM)[17][18]. In this study, the big amount of data is critical. A huge number of datasets geared toward illness analysis will be created. This study will make use of a breast cancer dataset. The data set was obtained from Kaggle, which provides a wide range of datasets for various reasons. To pre-process the dataset into an array form, train it, and develop a model, the preceding supervised machine learning algorithms are utilized. This model will be trained with the help of a Python library named "Scikit Learn." The model will be trained using the trained database to detect whether a patient has breast cancer. The algorithms' accuracy will be examined in order to determine the most effective algorithm for identifying breast cancer. NB Classifier, DT Classifier, KNN Classifier, and SVM were employed in this study. This section contains a full description of the many classifiers that are employed.

Assumptions made when using Decision Tree.

1) In the beginning, the entire training set is regarded as the root.

2) The values of features are categorical. If those values persist, the model is developed by converting them to discrete values.

3) Attribute values are used to recursively distribute records.

4) As the root or internal node, a statistical approach is employed to rank attributes.

System Flowchart

The system flowchart depicts the system's process flow across various stages. The flowchart in Figure 1 depicts the implementation of four alternative supervised machine learning algorithms for breast cancer prediction: Decision Tree, Support Vector Machine, Naïve Bayes, and K-Nearest Neighbour. Before proceeding to the last stage – stop – the system requires a dataset, which is generated and processed using the trained model.

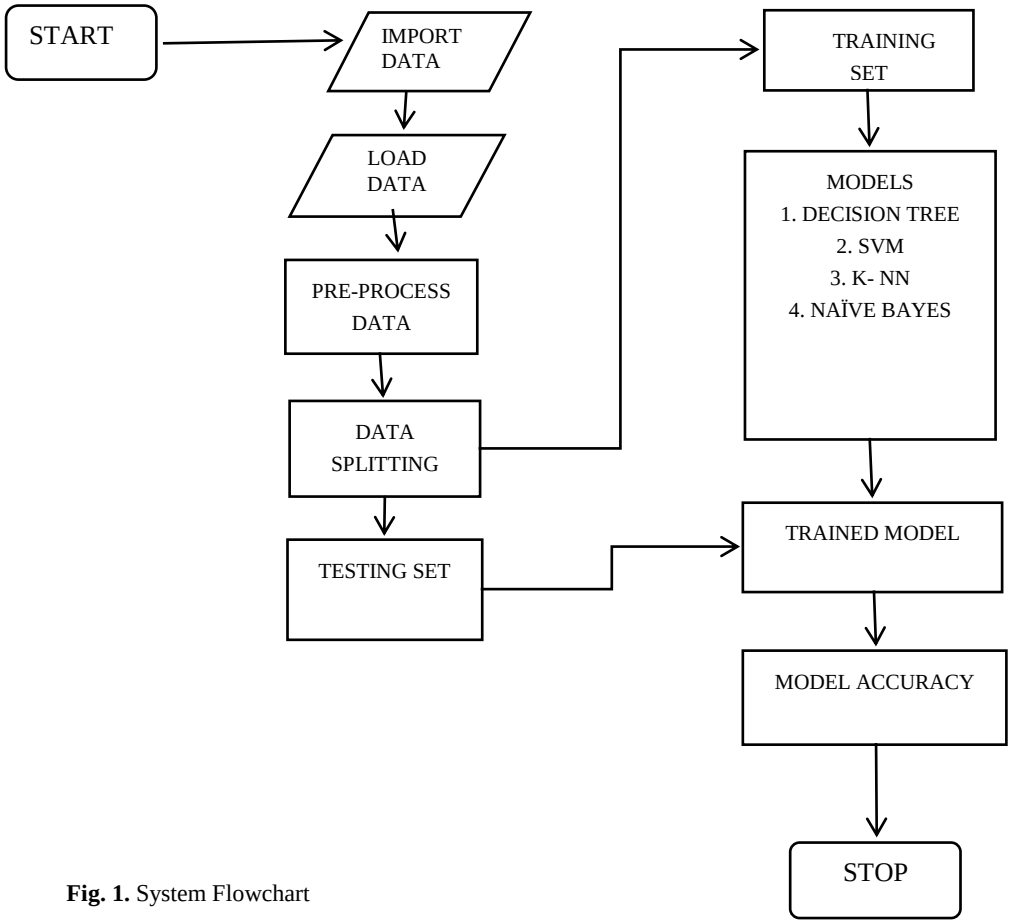

Fig. 1. System Flowchart

4 Implementation and Discussion

4.1 Development Tools

All through the development process of the system, a number of tools were used for the implementation of different aspects and functionalities of the system. Python programming language, Jupyter Notebook and Anaconda software was used to develop the model. This consisted of various libraries such as Pandas library, Numpy, Sklearn,

Matplotlib amongst many others. Also, the Anaconda package/platform was used for better integration of the model locally.

4.2 Import Libraries.

The first step for the diagnosis system is to import all the libraries that will be needed to develop the model. The libraries of all the algorithms used will need to be imported.

4.3 Import Dataset.

The next step is to import the dataset which will be used and processed by the various supervised machine learning algorithms to develop the model.

4.4 Training the Model.

The model is then trained with the imported dataset which contains 6 columns and 569 instances. During the training the model tries to learn and recognize patterns from the provided data for better accuracy. During this training much is taken in consideration to avoid under-fitting and over-fitting of the model as this can result into inaccurate predictions.

4.5 Algorithm Accuracy

The algorithms' accuracy is very important as it gives certainty to whether the algorithms can be used to predict breast cancer accurately. To achieve this, tests are performed on the model for correctness of results. Figure 2 shows the results of the accuracy gotten after the test was carried out.

Fig. 2. Algorithm Accuracy

The confusion matrix was use to further analysis the performance of the four models as show in Fig. 3. The results show that SVM perform better with an accuracy of 97.6%, precision of 94.6%, and sensitivity of 96.3%. follow by the Naive Bayes with an accuracy of 90.1%, precision of 92.4%, and sensitivity of 93.8%. the least of all the four classifiers is K-Nearest Neighbours with an accuracy of 88.9%, precision of 81.0%, and sensitivity of 93.0%. but the K-Nearest Neighbours perform better in term of sensitivity with 93.0%.

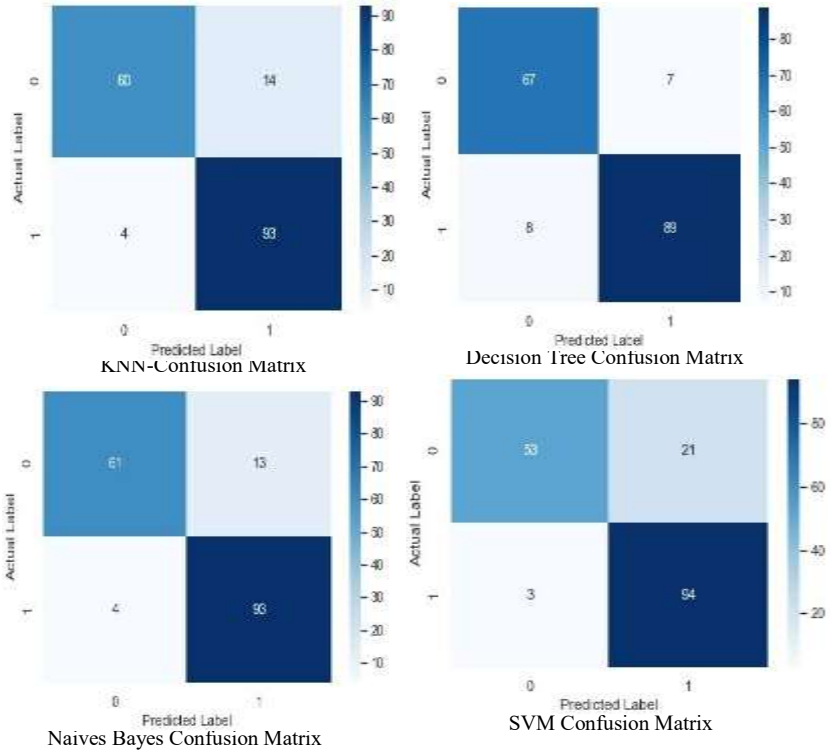

Fig. 3. Displays the confusion matrix for each of the classifiers used

A simulation of K-NN, DT, NB and SVM was conducted on breast cancer data for prediction. The Table 1 demonstrates classification performs of the four model suing accuracy, precision, and sensitivity.

Table 1. Results of Classification Algorithm

Method	Accuracy(%)	Precision(%)	Sensitivity(%)
K-Nearest Neighbours	88.9	81.0	93.0

Decision Tree	89.1	90.5	89.3
Naive Bayes	90.1	92.4	93.8
Support Vector Machine	97.6	94.6	96.3

Table 1 shows that in the diagnosis of predicting breasts, the Naïve Bayesian classification accuracy is higher than the classification accuracy of KNN, DT, and SVM classification algorithms, according to this study. With a precision of 90.1, it can be observed that this is the most accurate algorithm for classifying Naive Bayes data. KNN and decision tree are tied for second place with a degree of 88.9%. SVMs can occasionally achieve a precision of 87.6. The size of the data used in this study is a constraint. A modest number of samples are utilized for training and testing. Larger datasets should be used to analyze data relevant to the clinical environment.

Table 2. Comparison with related works

Models	**Methods**	**Accuracy (%)**
[14]	Random Forest	75.25
[18]	Naive Bayes	97.3
[12]	Support Vector Machine	92.2
[19]	Decision Tree	92.9
	Naive Bayes,	90.1,
	Decision Tree,	89.1,
	Support Vector Machine,	87.6,
Proposed method	K-Nearest Neighbour	88.9

Table 2 reveals the comparison of this work with others that has been done before. It shows that more research work will still need to be carried out to achieve a reasonable level of accuracy. The result is also affected by the size of the dataset. These results have laid the foundation that it is possible to predict breast cancer with a high degree of accuracy and therefore boost the impact of correcting them earlier to avoid terminal death of patients.

5 Conclusion

This study assessed the classification accuracy of four machine learning algorithms: K-NN, Decision Tree, Naive Bayes, and SVM. The goal of this comparison study was to determine the most accurate supervised machine learning algorithm that could be used to diagnose and forecast breast cancer effectively. Naive Bayes has the maximum accuracy for the supplied dataset, according to the prediction results. This reveals that, when compared to k-NN, SVM, and Decision Tree, Naive Bayes can be utilized to predict breast cancer. The size of the data used in this study is a constraint. A modest number of samples are utilized for training and testing. Larger datasets should be used to analyze data relevant to the clinical environment. It's vital to highlight that using machine learning techniques and algorithms to diagnose breast cancer is only for the purpose of improving it. In order to increase the accuracy of the model, we want to employ real-time data as well as a larger dataset in the future. Adding alternative algorithms, such as Random Forest and Logistic Regression, to improve the model's robustness can also be considered.

References

1. Oladipo, I. D., Babatunde, A. O., Awotunde, J. B., & Abdulraheem, M. (2020, November). An improved hybridization in the diagnosis of diabetes mellitus using selected computational intelligence. *Communications in Computer and Information Science*, 2021, 1350, pp. 272–285.
2. What Is Breast Cancer? Centers for Disease Control and Prevention. (2021). Retrieved 13 October 2021, from https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm.
3. Cancer. Who.int. (2021). Retrieved 13 October 2021, from https://www.who.int/news-room/fact-sheets/detail/cancer.
4. Awotunde, J. B., Folorunso, S. O., Bhoi, A. K., Adebayo, P. O., & Ijaz, M. F. (2021). Disease Diagnosis System for IoT-Based Wearable Body Sensors with Machine Learning Algorithm. In Hybrid Artificial Intelligence and IoT in Healthcare (pp. 201-222). Springer, Singapore.
5. S. A. Korkmaz, and M. Poyraz, "A New Method Based for Diagnosis of Breast Cancer Cells from Microscopic Images: DWEE—JHT," *J. Med. Syst.*, vol. 38, no. 9, p. 92, 2014.
6. Ed-daoudy, A., & Maalmi, K. (2020). Breast cancer classification with reduced feature set using association rules and support vector machine. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *9*, 1-10.
7. Chaurasia, V., & Pal, S. (2020). Applications of machine learning techniques to predict diagnostic breast cancer. *SN Computer Science*, *1*(5), 1-11.
8. A. F. M. Agarap, (2018), "On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset," ACM Int. Conf. Proceeding Ser., no. 1, pp. 5–9, doi: 10.1145/3184066.3184080.
9. Chaurasia, V., & Pal, S. (2017). A novel approach for breast cancer detection using data mining techniques. *International journal of innovative research in computer and communication engineering (An ISO 3297: 2007 Certified Organization) Vol*, *2*.
10. Dubey, A. K., Gupta, U., & Jain, S. (2018). Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(1), 18-29.

11. Ojha, U., & Goel, S. (2017, January). A study on prediction of breast cancer recurrence using data mining techniques. In *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence* (pp. 527-530). IEEE.
12. Palaniappan, R., Sundaraj, K., & Sundaraj, S. (2014). A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals. *BMC bioinformatics*, *15*(1), 1-8.
13. Das, R. K., Kasoju, N., & Bora, U. (2010). Encapsulation of curcumin in alginate-chitosan-pluronic composite nanoparticles for delivery to cancer cells. *Nanomedicine: Nanotechnology, Biology and Medicine*, *6*(1), 153-160.
14. Zolbanin, H. M., Delen, D., & Zadeh, A. H. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decision Support Systems*, *74*, 150-161.
15. S. Medjahed, T. Saadi, and A. Benyettou, "Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules," *International Journal of Computer Applications*, 2013, vol. 62, no. 1, pp. 0975 – 8887.
16. R. Sumbaly, N. Vishnusri, and S. Jeyalatha, "Diagnosis of Breast Cancer using Decision Tree Data Mining Technique," *International Journal of Computer Applications*, 2014, vol. 98, no. 10, pp. 0975 – 8887.
17. M. Elgedawy, "Prediction of Breast Cancer using Random Forest, Support Vector Machines and Naïve Bayes," *International Journal of Engineering and Computer Science*, 2017, vol. 6, no. 1, pp. 19884-19889.
18. Gana, N. N., Abdulhamid, S. I. M., Misra, S., Garg, L., Ayeni, F., & Azeta, A. (2020, December). Optimization of Support Vector Machine for Classification of Spyware Using Symbiotic Organism Search for Features Selection. In *International Conference on Information Systems and Management Science* (pp. 11-21). Springer, Cham.
19. Liu, W., Swetzig, W. M., Medisetty, R., & Das, G. M. (2011). Estrogen-mediated upregulation of Noxa is associated with cell cycle progression in estrogen receptor-positive breast cancer cells. PloS one, 6(12), e29466.
20. Awotunde, J. B., Folorunso, S. O., Bhoi, A. K., Adebayo, P. O., & Ijaz, M. F. (2021). Disease Diagnosis System for IoT-Based Wearable Body Sensors with Machine Learning Algorithm. In Hybrid Artificial Intelligence and IoT in Healthcare (pp. 201-222). Springer, Singapore.
21. Ogundokun, R. O., Sadiku, P. O., Misra, S., Ogundokun, O. E., Awotunde, J. B., & Jaglan, V. (2021, February). Diagnosis of Long Sightedness Using Neural Network and Decision Tree Algorithms. *Journal of Physics: Conference Series*, 2021, 1767(1), 012021.