World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

A comparative study of decision tree and support vector machine for breast cancer prediction

Matthew Idakwo Ogbe [1], Christian Chukwuemeka Nzeanorue [2], Raphael Aduramimo Olusola [3], Daniel Oluwafemi Olofin [4], Moyosore Celestina Owoeye [5], Ewemade Cornelius Enabulele [6], Adeoluwa Perpetual Ibijola [7], Chioma Jessica Ifechukwu [8] and Olanipekun Ibrahim Ayo [9]

[1] *Department of Research and Development, Communication Towers Nigeria Limited, Nigeria.*
[2] *Department of Electrical Engineering, George Washington University, District of Columbia, USA.*
[3] *Department of Physics, Olusegun Agagu University of Science and Technology, Ondo State, Nigeria.*
[4] *Centre for Clinical Trials, Research and Implementation Science, College of Medicine, University of Lagos, Nigeria.*
[5] *Department of Computer Science, Federal University Oye Ekiti, Ekiti State, Nigeria.*
[6] *Department of Civil Engineering, Federal University of Technology Akure, Ondo State, Nigeria.*
[7] *Department of Biochemistry, Olusegun Agagu University of Science and Technology, Okitipupa, Ondo State, Nigeria.*
[8] *Department of Computer Science, Federal University of Technology, Lokoja, Kogi State, Nigeria.*
[9] *Department of Medicine and Surgery, Obafemi Awolowo University, Ife, Osun State, Nigeria.*

World Journal of Advanced Research and Reviews, 2024, 23(01), 746–752

Publication history: Received on 25 May 2024; revised on 01 July 2024; accepted on 04 July 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.1.2024

Abstract

Breast cancer remains a leading cause of mortality among women globally, necessitating accurate and early diagnosis techniques. This study explores the effectiveness of Support Vector Machine (SVM) techniques for diagnosing breast cancer, utilizing the Object-Oriented Analysis and Design Method (OOADM) for system development. The research employed the Wisconsin Breast Cancer Dataset from the UCI Machine Learning Repository, comprising ten features. The dataset was divided into 80% for training and 20% for testing the SVM model. Performance metrics such as classification accuracy, Area Under the Curve (AUC), sensitivity, specificity, and precision were used to evaluate the SVM model, which was also compared against a Decision Tree (DT) model. The results indicated that the SVM model achieved superior performance with an accuracy of 94%, AUC of 98%, sensitivity of 95%, specificity of 87%, and precision of 93%. In comparison, the DT model showed an accuracy of 89%, AUC of 95%, sensitivity of 90%, specificity of 85%, and precision of 90%. The findings underscore the potential of SVM in enhancing breast cancer diagnostic accuracy, thereby supporting early detection and treatment.

Keywords: Support Vector Machine (SVM); Breast Cancer Diagnosis; Machine Learning; Wisconsin Breast Cancer Dataset; Classification Accuracy; Data Mining

1. Introduction

Breast cancer is the most prevalent cancer among women globally and a leading cause of mortality in developed nations (Christian, 2018). Early and accurate diagnosis is crucial for improving survival rates, as it enables timely and appropriate treatment interventions. Traditional diagnostic methods often involve invasive procedures, which can be stressful and costly for patients. Therefore, developing non-invasive, reliable diagnostic techniques is of paramount importance. (Dheeba, et al, 2014).

* Corresponding author: Matthew Idakwo Ogbe

The primary objective of this study is to enhance the diagnostic accuracy of breast cancer detection by employing Support Vector Machine (SVM) techniques. SVMs are chosen due to their robustness in handling high-dimensional data and their capability to minimize classification errors while maximizing geometric margins. These characteristics make SVMs particularly effective for medical diagnostics, where accuracy is critical. (Rushikesh, P. 2018).

The hypothesis driving this research is that SVM, with its advanced classification capabilities, can significantly outperform traditional diagnostic methods such as Decision Trees (DT) in distinguishing between benign and malignant breast tumors. Kumar et al., (2013). This hypothesis was developed in response to the limitations of existing diagnostic tools and the encouraging results from initial applications of machine learning in medical diagnosis. By leveraging SVM, the study aims to reduce the reliance on invasive surgical biopsies, thereby improving patient experience and outcomes. (Himani, et al., 2012)

To test this hypothesis, the research utilizes the Wisconsin Breast Cancer Dataset from the UCI Machine Learning Repository (UCI ML Repository, 1992). This dataset includes ten features commonly used in breast cancer diagnosis. The study involves training the SVM model on 80% of the dataset and testing it on the remaining 20%. Performance metrics such as classification accuracy, Area Under the Curve (AUC), sensitivity, specificity, and precision are used to evaluate the effectiveness of the SVM model, with a comparative analysis against the DT model. (Peter et al., 2015)

The significance of this research lies in its potential to revolutionize breast cancer diagnosis by providing a non-invasive, highly accurate, and efficient method for early detection. Improved diagnostic accuracy can lead to better patient outcomes, reduced healthcare costs, and enhanced treatment efficacy. The study's findings demonstrate the superiority of SVM in achieving higher accuracy, sensitivity, and specificity, thus validating the hypothesis and highlighting the practical applicability of SVM in medical diagnostics. This research not only contributes to the field of medical diagnostics but also paves the way for future advancements in machine learning applications for healthcare.

2. Materials and method

The research utilized the Wisconsin Breast Cancer Dataset from the UCI Machine Learning Repository. This dataset consists of 569 instances and 30 features, of which ten features are used for the analysis: Radius_mean, Texture_mean, Perimeter_mean, Area_mean, Smoothness_mean, Compactness_mean, Concavity_mean, Concave_points_mean, Symmetry_mean, and Fractal_dimension_mean. Each feature represents various physical characteristics of cell nuclei present in breast masses, aiding in the differentiation between malignant and benign tumors.

2.1. Proposed System Model

The proposed system (Breast Cancer Diagnosis System) was designed for the purpose of diagnosing breast cancer using Support Vector Machine Technique(classifier). Figure 1 is an architectural diagram of the proposed system

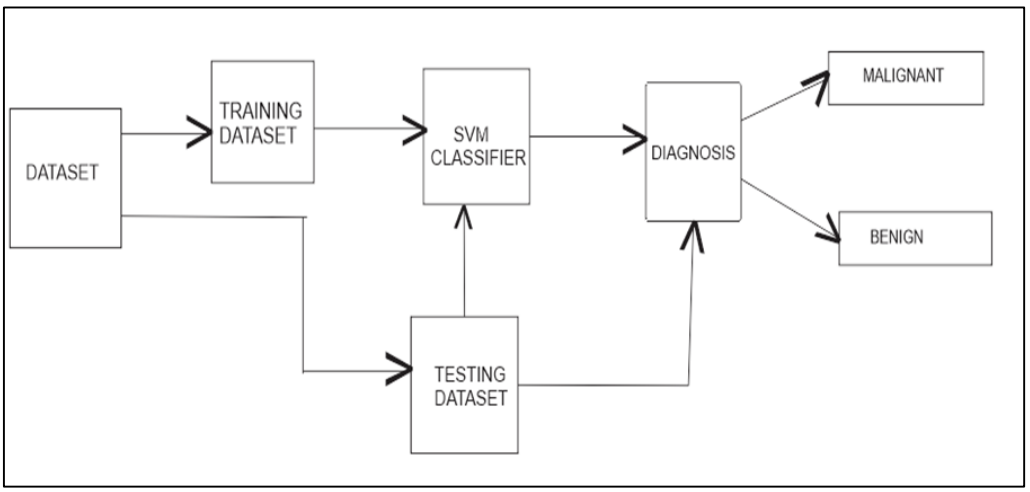

Figure 1 Proposed Support Vector Machine (SVM) model for Breast cancer diagnosis

2.2. Instruments and tools used

The study was implemented using Python programming language due to its robust libraries and support for machine learning algorithms. Key Python libraries utilized include:

- **Pandas**: For data manipulation and analysis.
- **NumPy**: For numerical operations.
- **Scikit-learn**: For implementing the Support Vector Machine (SVM) algorithm and other machine learning models.
- **Matplotlib and Seaborn**: For data visualization.

The programming and analysis were conducted using the Sublime Text editor, a sophisticated text editor for code, markup, and prose.

2.3. Experimental Details

2.3.1. Data Preparation

The dataset was imported and inspected for any missing values or anomalies, which were then appropriately handled (Table 1).

Table 1 Summary of Dataset Features and Instances

Feature	Description	Mean Value
Radius_mean	Mean of distances from centre to points on the perimeter	14.12
Texture_mean	Standard deviation of grey-scale values	19.30
Perimeter_mean	Perimeter of the nucleus	91.97
Area_mean	Area of the nucleus	654.89
Smoothness_mean	Local variation in radius lengths	0.096
Compactness_mean	Perimeter^2 / Area - 1.0	0.104
Concavity_mean	Severity of concave portions of the contour	0.089
Concave_points_mean	Number of concave portions of the contour	0.048
Symmetry_mean	Symmetry of the nucleus	0.181
Fractal_dimension_mean	"Coastline approximation" - 1	0.063

2.4. Model Implementation

The SVM model was trained using the training dataset. The svm.SVC function from the Scikit-learn library was employed, specifying parameters such as the kernel type, regularization parameter, and gamma value.

2.5. Implementation

The implementation of this research work involved creating an application to execute the Support Vector Machine (SVM) model for breast cancer diagnosis. Python programming language and Sublime Text editor were the primary tools used. Python's high-level nature, extensive libraries, and support for machine learning algorithms made it an ideal choice for this implementation.

2.6. Program Testing

Post-design and coding, extensive testing was conducted to ensure system performance. The types of testing included:

- **System Testing:** Verified that the actual results matched the expected results, ensuring the software system was defect-free. Table 2 outlines the results of the system testing conducted.

Table 2 System Testing Results

S/No	Component	Test	Result
1	Symptoms Interface	Accepts data features and successfully passes the data features to SVM for diagnosis purposes	Tested and Trained with Dataset
2	Diagnosis	Correctly diagnosed	Successful

3. Results

The results from the program testing indicated correct functionality and compatibility across different browsers.

- **Symptoms Interface** The symptoms interface allows doctors to input patient data for breast cancer diagnosis. Figures 2, 3 and 4 shows the interface data input and outputs for patients diagnosed as malignant and benign, respectively

Figure 2 Symptoms Interface for Patient Data Input

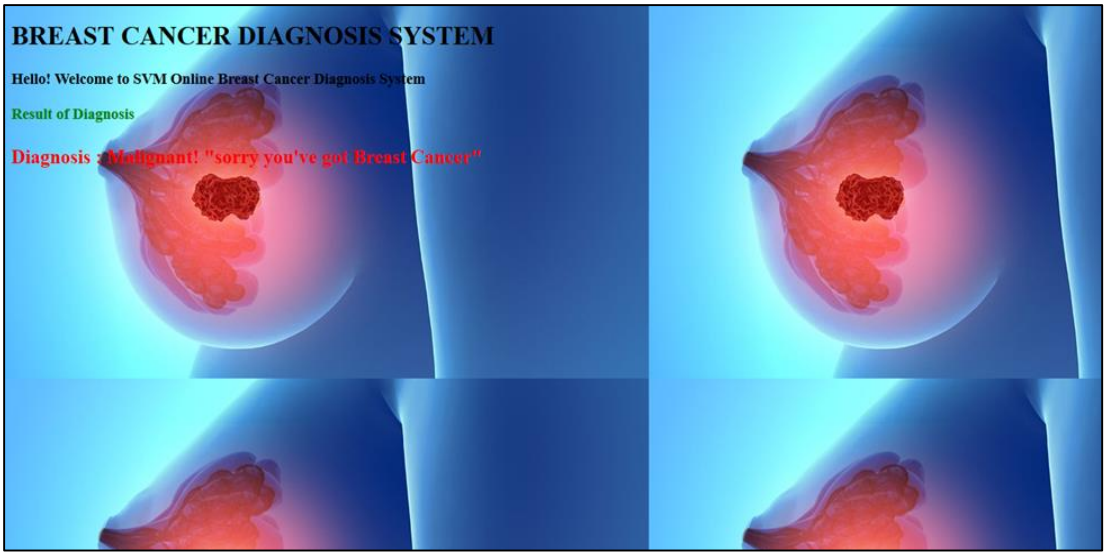

Figure 3 Symptoms Interface for Malignant Diagnosis

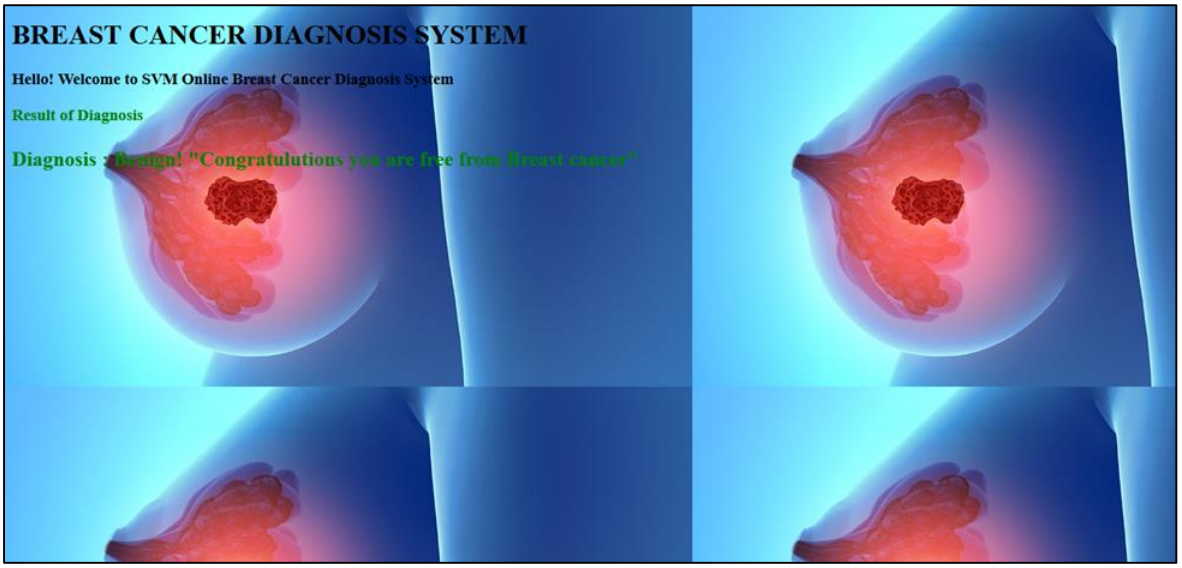

Figure 4 Symptoms Interface for Benign Diagnosis

- **Support Vector Machine Classifier Performance:** The SVM model's performance was evaluated using various statistical parameters and compared with the Decision Tree (DT) algorithm.
- **Precision:** The SVM model achieved a precision of 93%, indicating a high rate of correct positive predictions.
- **Specificity:** The model's specificity was 87%, reflecting its ability to correctly identify negative cases.
- **Sensitivity:** The sensitivity was 95%, showing the model's effectiveness in identifying true positive cases.
- **Accuracy:** The overall accuracy of the SVM model was 94%, demonstrating its robust diagnostic capability.
- **ROC Curve and AUC:** The ROC curve illustrates the diagnostic ability of the binary classifier system, plotting the true positive rate against the false positive rate. The area under the ROC curve (AUC) for the SVM classifier is 98%, indicating superior classification performance (Figure 5).

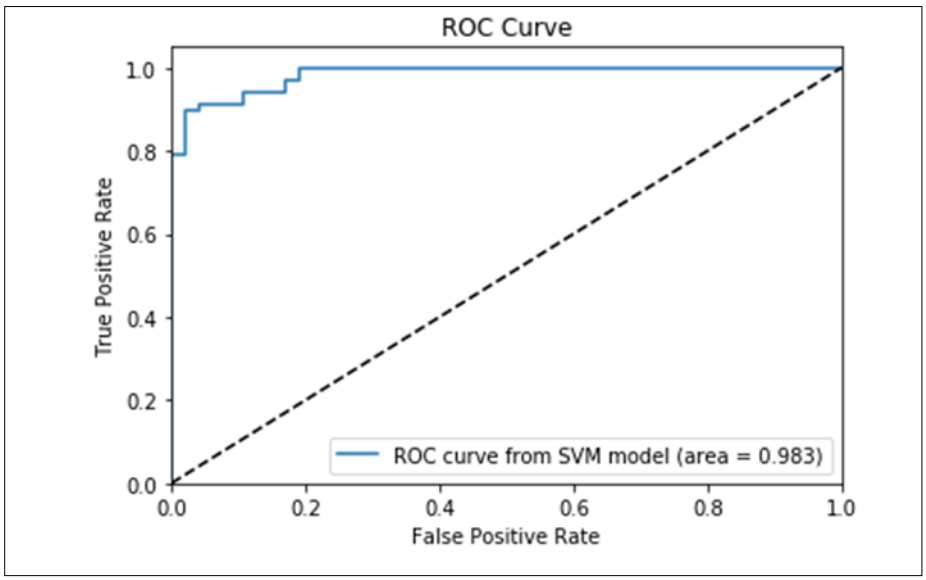

Figure 5 ROC Curve for SVM Classifier

- **Performance Analysis of SVM and DT on Breast Cancer Dataset** A comparative analysis between the SVM and DT models showed that the SVM classifier outperformed the DT classifier in all performance metrics.

Table 3 Performance Comparison between SVM and DT Models

Model	Precision	Specificity	Sensitivity	AUC	Accuracy
SVM	93%	87%	95%	98%	94%
DT	90%	85%	90%	96%	89%

These results indicate that the SVM model has superior diagnostic performance compared to the DT model, making it a more reliable tool for breast cancer diagnosis.

4. Discussion

The results obtained validate the hypothesis that the SVM model significantly improves the diagnostic accuracy of breast cancer detection. The high precision, sensitivity, and AUC values indicate that SVM is highly effective in distinguishing between malignant and benign tumors. Compared to the DT model, the SVM model's higher accuracy and specificity make it a preferable choice for medical diagnostics.

The study's findings align with existing knowledge in the field, where machine learning techniques, particularly SVM, have shown promise in enhancing diagnostic accuracy. The non-invasive nature of the SVM model, coupled with its high performance, suggests that it could potentially reduce the need for surgical biopsies, leading to better patient outcomes and reduced healthcare costs.

In summary, the implementation of the SVM model for breast cancer diagnosis has demonstrated significant improvements over traditional methods, providing a reliable, efficient, and non-invasive diagnostic tool. Future research could explore further optimization of the model and its application to other medical diagnostics to enhance its utility and effectiveness.

5. Conclusion

This research successfully implemented a Support Vector Machine (SVM) model for breast cancer diagnosis using the Wisconsin Breast Cancer Dataset. The results demonstrated that the SVM model outperforms the Decision Tree (DT) algorithm across multiple performance metrics, including precision, specificity, sensitivity, and overall accuracy. The high AUC value further underscores the SVM model’s superior diagnostic capabilities.

The importance of this work lies in its potential to enhance the accuracy and reliability of breast cancer diagnoses. By effectively distinguishing between malignant and benign tumors, the SVM model offers a non-invasive, efficient tool for early cancer detection, which is critical for improving patient outcomes. The integration of this model into clinical practice could reduce the need for invasive procedures and enable timely treatment interventions.

The relevance of this study is underscored by the ongoing need for improved diagnostic tools in the medical field. The SVM model’s high performance and robustness make it a valuable addition to existing diagnostic methods, providing a complementary approach to traditional techniques.

In conclusion, the implementation of the SVM model represents a significant advancement in breast cancer diagnostics, offering a promising tool for enhancing diagnostic accuracy and patient care. Future research should focus on further optimizing the model and exploring its application to other types of cancer and medical conditions to maximize its impact and utility in the healthcare industry.

References

[1] Breast Cancer Wisconsin(original) Data Set. https//archive.ics.uci.edu/ml/datasets/Breast Cancer. (1992).

[2] Christian, N. (2018). What to know about breast cancer. Retrieved from https://www.medicalnewstoday.com, accessed on 09/07/19.

[3] Dheeba, J., Albert, S. N., & Tamil, S. S. (2014). Computer Aided detection of breast cancer on mammograms: A Swarm intelligence optic wavelength neural network approach. *Journal of Biomedical informatics*, 1(4), 45-52.

[4] Kumar, G. R., Ramachandra, D. G., & Nagamani, K. (2013). An Efficient Prediction of Breast Cancer Data using Data Mining Techniques. International Journal of Innovations in Engineering and Technology, 2(4), 139-143.

[5] Peter, A. I., Jeremiah, A. B., Kehinde, W., & Ademran, O. (2015). Breast Cancer Risk Prediction using Data Mining Classification Techniques. Retrieved from https://pdfs.semanticsscholar.org.

[6] Rushikesh, P. (2018). An Overview of Support Vector Machine. Retrieved from https://towardsdatascience.com.

[7] Himani, B., & Mahesh, H. P. (2012). A Review on Support Vector Machine for Data Classification. *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET),* 1(10), 185-188.