Science World Journal Vol. 19(No 3) 2024 https://dx.doi.org/10.4314/swj.v19i3.24
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

WEB BASED APPLICATION FOR BREAST CANCER DETECTION USING SUPPORT VECTOR MACHINE

*[1]Dada O.S., [1]Nathaniel A.N., [2]Irunokhai E.A., [1]Nuhu M.A., [1]Jiji S.A.

[1]Department of Computer Science, Federal University of Kashere, Nigeria
[2]Department of Computer Science, Federal College of Wildlife, New-Bussa, Nigeria

*Corresponding Author Email Address: saintdada2000@gmail.com

ABSTRACT

Due to the life-threatening nature of breast cancer, which predominantly affects women, early detection is critical for improving patient outcomes. Traditional screening methods, such as mammography, clinical examinations, and biopsies, are widely employed in the healthcare sector. However, these approaches face challenges, particularly in the misclassification of tumors. In this study, we developed a web application utilizing a Support Vector Machine (SVM) algorithm to create a predictive model for breast cancer that accurately distinguishes between benign and malignant tumors. Feature selection was employed to identify the most informative and relevant variables in the dataset, thereby mitigating the curse of dimensionality and enhancing model performance. To ensure accessibility, the predictive model was integrated into a web application, allowing medical professionals to use the tool for informed decision-making. Experiments were conducted using the publicly available Wisconsin Diagnostic Breast Cancer (WDBC) dataset, with results demonstrating a notable improvement in accuracy compared to similar studies.

Keywords: Machine learning, Classification Model, SVM, Breast Cancer

INTRODUCTION

Breast cancer is a major health issue that affects significant number of women globally, presenting a substantial danger to their overall well-being and health (Akram et al., 2017). In 2018, 2.1 million incident cases of breast cancer were identified worldwide, which is the second most prevalent disease overall after lung cancer, accounting for roughly 12% of all incident cancer cases, with an estimated 627,000 fatalities occurring globally (Osei-Afriyie et al., 2021). Early detection plays a vital role in improving patient outcomes and survival rates (Rubin et al., 2011). Machine learning techniques have gained attention in recent years for their potential to assist in breast cancer prediction and diagnosis. One critical aspect of developing reliable and efficient machine learning models for breast cancer prediction is the selection of informative features.

Breast cancer datasets often contain a large number of attributes, many of which may be irrelevant, redundant, or noisy. Including these irrelevant features can introduce bias, reduce model performance, and increase computational complexity. Therefore, Feature Selection is an effective method for reducing the curse of dimensionality and certainly improving the accuracy of the predictive model (Assegie et al., 2022). The use of feature selection is to determine the necessary feature required for the training model and to remove irrelevant and duplicate features. According to the most recent global cancer statistics, 2.26 million new instances of breast cancer were recorded and diagnosed in 2020, making it the illness that kills the most women globally (Wilkinson & Gathani, 2022).

Detecting breast cancer at an early stage is crucial for successful treatment and improved patient outcomes. Early detection allows healthcare workers to make timely intervention, leading to a higher likelihood of successful treatment, reduced morbidity, and increased survival rates (American Cancer Society, 2021). Therefore, there is a pressing need for accurate and efficient methods to detect breast cancer at its earliest stages.

In this work we develop a web-based application for breast cancer prediction that incorporates machine learning techniques and feature selection methods. The proposed model leverages the strengths of existing studies while addressing their limitations. Therefore, recognizing the significance of feature selection in predicting breast cancer is essential. Feature selection enhances prediction accuracy by identifying the most informative features, reducing overfitting, and improving generalization. It also makes it easier for clinicians and researchers to understand the factors that influence predictions and gain valuable insights into the disease. By choosing a precise set of features, computational efficiency improves, enabling faster model training and deployment, which is especially important in time-sensitive clinical settings. Additionally, feature selection helps overcome data limitations, mitigates the impact of missing values or noise, and enhances reliability. This cost-effective approach reduces the need for extensive testing, benefiting both patients and healthcare systems. Yadav et al., (2019) conducted a comparative study of several machine learning algorithms (such as Decision Tree, Random Forest, Support Vector Machine, Naive Bayes Classifier, Artificial Neural Network, and K-nearest Neighbour) for breast cancer prediction. The accuracy, precision, and recall of all six ML techniques were compared. With the help of the Wisconsin dataset, they compared the performance of the aforementioned algorithms. However, Support Vector Machine and Random Forest achieved the highest accuracy of 97.2%, Naive Bayes Classifier has the highest precision and recall of 97.2% & 97.1%.

Doe et al., (2018), for instance, investigated the application of support vector machines (SVM) in breast cancer prediction and found promising results in terms of accuracy and sensitivity. Another study by Krishnamurthi et al., (2019) emphasized the importance of feature selection and data visualization in breast cancer prediction. The study highlighted the impact of selecting appropriate features on improving the accuracy and efficiency of predictive models for breast cancer detection.

Alfian et al., (2022) proposed a breast cancer prediction model based on support vector machines (SVM) and an extra-trees-based feature selection method. A dataset containing breast cancer risk variables was used. Alfian et al., (2022) demonstrated the effectiveness of their approach in accurately predicting breast cancer based on risk factors with an accuracy of 80.23%. Other machine learning algorithms, such as Support Vector Machines, Random Forests, Logistic Regression, Decision Trees, K-Nearest Neighbours, and Neural networks, have been used in breast cancer prediction research (Naji et al., 2021).

Deshwal and Sharma (2019) presented a grid search approach for identifying breast cancer using Support Vector Machine (SVM). The result obtained using grid search with Breast Cancer Dataset gives a much better result than the normal SVM Model. Their proposed system model compared the result with the grid search method and without grid search methods using the SVM Classification algorithms. However, the SVM Grid Search model performed better in terms of accuracy, precision, recall, and F1 score. They obtained an accuracy of 95%.

Subhani et al., (2019) focused on enhancing predictive models to achieve excellent performance in disease outcome detection using supervised machine learning techniques. On the WBCD dataset, they proposed and analyzed the implementations of three distinct machine learning classifiers, including logistic regression (LR), support vector machine (SVM), and k nearest neighbors. SVM performed best, with a 92.7% accuracy rate.

Support Vector Machine (SVM)

Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon, and Vapnik in COLT-92. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. In another terms, Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. Support Vector machines can be defined as systems which use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory (Burges, 1998**)** Support Vector Machine is a discriminative classifier that can be defined by a separating hyperplane. It is the generalization of maximal margin classifier which comes with the definition of hyperplane. In an n-dimensional space, the hyperplane is of (n-1) dimension with flat subspace that need not pass through the origin. The hyperplane is not visualized in higher dimension but the notion of an (n-1) dimensional flat subspace still applies [10]. If there doesn't exist any linearly separable hyperplane for any dataset, linear classifier can't be formed in that case. Kernel trick have to be applied to maximum-margin hyperplanes to develop nonlinear classifier. According this, nonlinear kernel function will be applied to the hyperplanes in replacement of dot product. Cubic, quadratic or higher-order polynomial function, Gaussian Radial basis function or Sigmoid function are forms of nonlinear kernel function. In p-dimensions, a hyperplane is described as follows.(Islam et al. 2017)

$$\boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \ldots + \boldsymbol{\beta_p x_p} = \mathbf{0} \qquad (1)$$

where **β0, β1, β2 … βp** are the hypothetical values and x_p are thedata points in sample space of p dimension

MATERIALS AND METHODS

In this section, the methodology for developing a web-based application for breast cancer prediction using machine learning and feature selection is presented. The section outlines the steps and techniques employed to address the research project objectives. This chapter provides a comprehensive methodology for the development of a web-based application for breast cancer prediction using machine learning and feature selection. It covers the system architecture, the flow chart, the use case diagram and also data collection and pre-processing, feature selection techniques, machine learning model selection, training and evaluation, web application development, deployment and integration, validation and testing, as well as limitations and mitigation strategies.

Data Collection and Preparation

The breast cancer named as Wisconsin Breast Cancer (WBC) data set is retrieved from UCI machine learning repository dataset

Table 1: Breast Cancer Dataset Description

Dataset	WBCD
No. of Attributes	32
No. of Instances	569
No. of Classes	2

The dataset contains features extracted from digitized images of fine needle aspirates (FNAs) of breast masses. Each FNA sample corresponds to a potential breast cancer case. The features are computed from a digitized image of a cytological slide, capturing various characteristics of the cell nuclei present in the image. The dataset contained 569 rows (i.e. instances) and 32 columns (i.e. Attributes) taken from needle aspirates from patients' breasts, of which 357 cases were identified as "benign" and the remaining 212 cases were classified as "malignant"

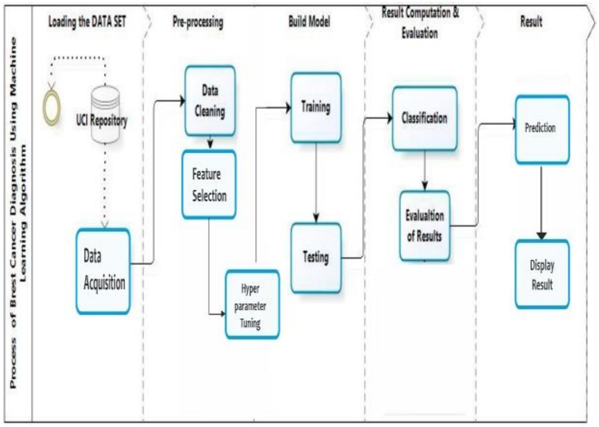


Figure 1: Overview of Proposed Methodology

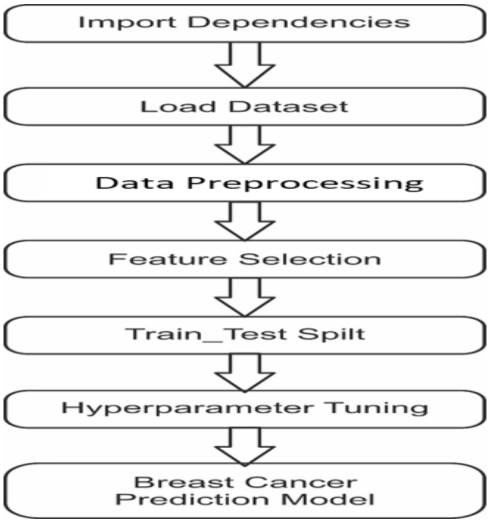


Figure 2: Flow Diagram of Proposed Methodology

Four performance metrics namely confusion matrix, precision, recall, and f1-score are used to evaluate the performance of the trained models. Then, their performances are discussed, analyzed and hypotheses are made. Fig. 4.1 shows the confusion matrix and Figure 4.2 represents the accuracy score, precision score, recall score, and f1-score for the SVM model when feature selection is applied on the WBCD

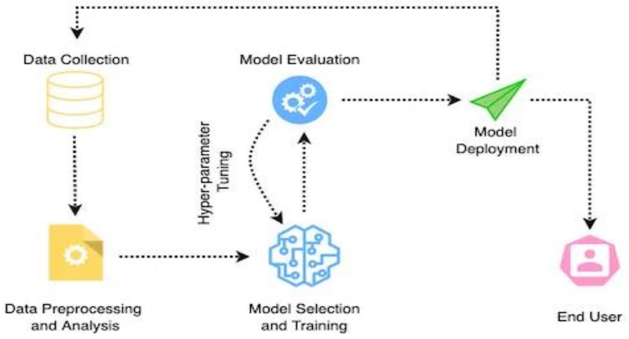


Figure 3: Development Lifecycle of Model Development

Confusion Matrix

The confusion matrix is a commonly used performance evaluation metric in machine learning and classification tasks. It provides a summary of the performance of a classification model by showing the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.

Terms associated with the Confusion matrix:

1. True Positives (TP): True positives are the cases when the actual class of the data point was True (1) and the predicted is also True (1).
2. True Negatives (TN): True negatives are the cases when the actual class of the data point was False (0) and the predicted is also False (0).
3. False Positives (FP): False positives are the cases when the actual class of the data point was False (0) and the predicted is True (1).
4. False Negatives (FN): False negatives are the cases when the actual class of the data point was True (1) and the predicted is False (0). False is because the model has predicted incorrectly and negatively because the class predicted was a negative one (0).

Model Deployment

The Model deployment is the process of integrating and making the developed machine-learning model accessible for real-world use. It involves implementing the model into a practical application or system where it can be utilized to make predictions or provide valuable insights. In the context of the project on breast cancer prediction, model deployment aims to create a web-based application that enables healthcare professionals or users to input relevant patient data and obtain predictions regarding the likelihood of breast cancer

Development Lifecycle of the Machine Learning Model to a Web Application

The development life cycle of the machine learning model encompasses various stages and processes that are essential for the successful integration and utilization of the machine learning model in a practical application ensuring its compatibility and seamless interaction with any kind of web browser platform. The diagram below provides an overview of the development lifecycle.

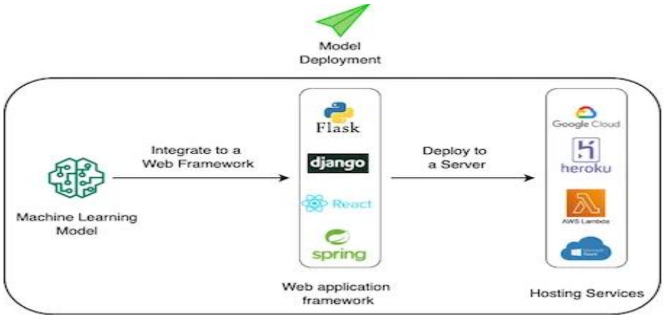


Figure 4: Machine Learning Model Integration to a Web Application

Web Application Workflow

The web application workflow in the context of a machine learning-based breast cancer prediction system involves a series of steps and interactions between different components to provide a seamless user experience.
High-level description of the workflow

1. User Interaction: The workflow begins with the user accessing the web application through a web browser. The user interface (UI) allows users to interact with the application, input relevant data, and initiate the breast cancer prediction process.

2. Data Input and Validation: The web application prompts the user to input relevant data required for breast cancer prediction which includes the specific features or attributes related to breast cancer diagnosis. The input data is validated to ensure it meets the required format and completeness criteria.

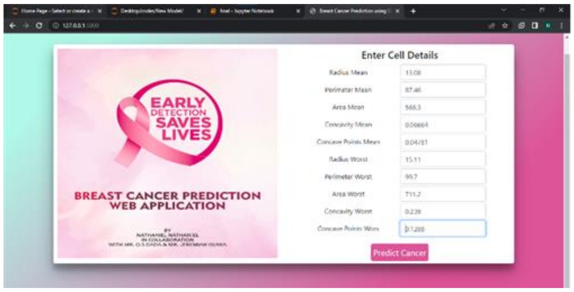
3. Data Pre-processing: Once the input data is validated, it undergoes pre-processing steps to ensure its compatibility with the machine learning model. This may involve data cleaning, normalization, feature scaling, or encoding categorical variables.

4. Feature Selection: The web application incorporates feature selection techniques to identify the most relevant features from the input data. This step helps reduce the dimensionality of the data and improves the efficiency and performance of the breast cancer prediction model

5. Prediction Request: After data pre-processing and feature selection, the web application sends a prediction request to the underlying machine learning model. This typically involves invoking an API or function that triggers the prediction process.

6. Machine Learning Model Execution: The machine learning model, which has been previously trained on a breast cancer dataset, processes the input data and generates a prediction. The model's algorithm analyses the input data, applies the learned patterns, and produces an output indicating the likelihood of breast cancer presence or not.

7. Result Display: The predicted outcome is communicated back to the web application, which then displays the results to the user


Enter Cell Details
EARLY DETECTION SAVES LIVES
BREAST CANCER PREDICTION WEB APPLICATION
BY
NATHANIEL NATHANIEL
IN COLLABORATION
WITH MR. O.S DADA & MR. JEREMIAH ISUWA
Predict Cancer

Radius Mean: 13.08
Perimeter Mean: 87.46
Area Mean: 566.3
Concavity Mean: 0.06664
Concave Points Mean: 0.04781
Radius Worst: 15.11
Perimeter Worst: 99.7
Area Worst: 711.2
Concavity Worst: 0.239
Concave Points Worst: 0.1288

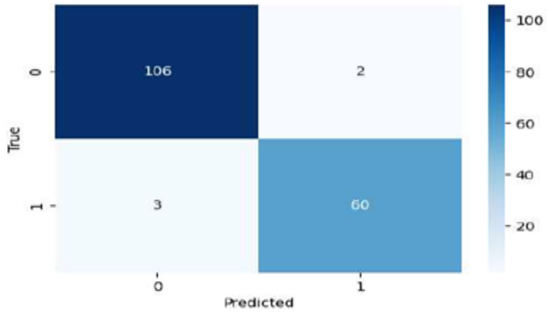
Figure 5: Front end of web application

RESULTS AND DISCUSSION

This section presents the analysis and discussion of the results obtained from the developed breast cancer prediction model. This chapter aims to provide a comprehensive analysis of the model's performance, evaluate its effectiveness in predicting breast cancer, and discuss the implications of the findings. The results analysis and discussion are based on the data collected and processed using Support Vector Machine with feature selection techniques.

Table 2: Confusion Matrix Description

Predicted Class	Actual Class	
Benign(0)	True Negative (TN)	False Positive(FP)
Malignant (1)	False Negative(FN)	True Positive (TP)



.

Figure 6: Confusion Matrix of Proposed Model

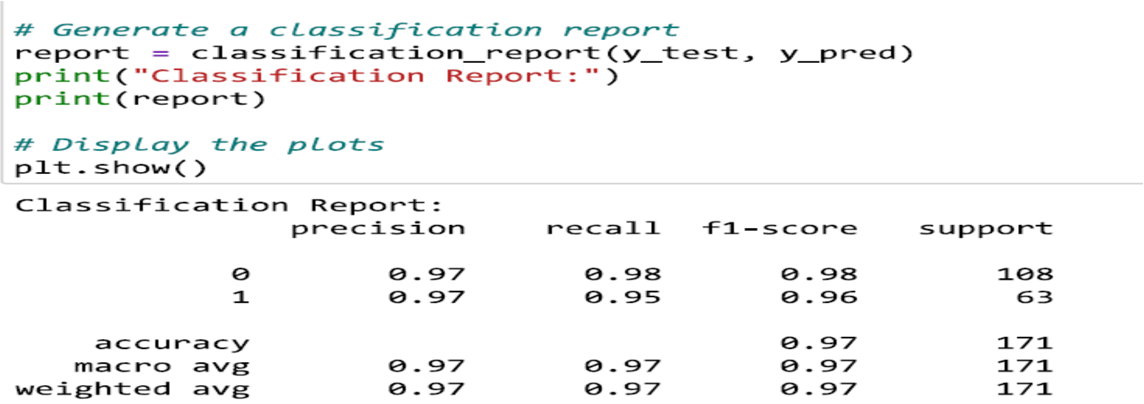
Accuracy = $\frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}}$ **(2)**

Precision = $\frac{\mathbf{TP}}{\mathbf{TP+FP}}$ **(3)**

F1 score = $\frac{2 * \boldsymbol{Precision}}{\boldsymbol{Precision} + \boldsymbol{Recall}}$ **(4)**

Support

The Support represents the number of occurrences of each class in the dataset. It provides insights into the distribution of instances across different classes, which is useful for understanding the model's performance on specific classes.



```
# Generate a classification report
report = classification_report(y_test, y_pred)
print("Classification Report:")
print(report)

# Display the plots
plt.show()
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.98      0.98       108
           1       0.97      0.95      0.96        63

    accuracy                           0.97       171
   macro avg       0.97      0.97      0.97       171
weighted avg       0.97      0.97      0.97       171
```

Figure 7: Performance of SVM model

Comparison between Similar Work and the Proposed Model

After completing the implementation of the SVM Model using Grid Search and Feature Selection for detecting breast cancer. From the dataset, the results can be compared from Figure 4.2 using the performance metrics discussed previously.

Table 3: Result Comparison between existing work and our proposed model

Authors	Method	No. Feature	Algorithms	Precision	Recall	F1-Score	Accuracy
Deshwal and Sharma (2019)	Grid Search	32	SVM	95%	95%	95%	95%
Proposed Model	Grid Search &Feature Selection	10	SVM	97%	97%	97%	97%

In this result comparison between the existing work of that of Deshwal & Sharma (2019) and our proposed model for breast cancer classification, the existing approach utilizes SVM Grid Search with no feature selection and achieves an accuracy of 95%. In contrast, our proposed model incorporates SVM Grid Search with Feature Selection and achieves a higher accuracy of 97%. The key difference lies in the feature selection step, which allows our model to focus on the most relevant features, reducing overfitting and enhancing generalization. Moreover, our proposed model offers improved interpretability, as it highlights the contribution of specific features to the classification decision.

Overall, our results demonstrate the superiority of the proposed model over the existing approach. With a 2% increase in accuracy, the feature selection step proves to be pivotal in enhancing the model's performance. Our proposed SVM Grid Search with a Feature Selection model presents a compelling advancement in breast cancer classification, providing robust and accurate predictions that can be beneficial in real-world applications.

CONCLUSION

In conclusion, this study presents a significant advancement in breast cancer prediction by integrating feature selection and machine learning techniques within a web-based application. By utilizing a Support Vector Machine (SVM) algorithm alongside feature selection, the proposed model effectively reduces dimensionality, enhances predictive accuracy, and improves overall model performance. The application provides medical professionals with an accessible tool for accurate classification of benign and malignant tumors, which is crucial for early detection and timely intervention. The results demonstrate the superiority of the proposed model over existing approaches, with a notable improvement in accuracy. This research underscores the importance of feature selection in improving the reliability, efficiency, and interpretability of predictive models, ultimately contributing to better outcomes in clinical settings.

REFERENCE

Akram, M., Iqbal, M., Daniyal, M. & Khan, A.U., 2017. Awareness and current knowledge of breast cancer. *Biological Research*, 50(1). Available at: https://doi.org/10.1186/s40659-017-0140-9.

Alfian, G., Syafrudin, M., Fahrurrozi, I., Fitriyani, N.L., Atmaji, F.T.D., Widodo, T., Bahiyah, N., Benes, F. & Rhee, J., 2022. Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method. *Computers*, 11(9), p.136. Available at: https://doi.org/10.3390/computers11090136.

American Cancer Society, 2021. Breast Cancer Facts & Figures. Available at: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2021-2022.pdf.

Assegie, T.A., Tulasi, R.L., Elanangai, V. & Kumar, N.K., 2022. Exploring the performance of feature selection method using breast cancer dataset. *Indonesian Journal of Electrical Engineering and Computer Science*, 25(1), p.232.

Breast Biopsy, 2021. Available at: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/breast-biopsy.

Carrera, P., Kantarjian, H.M. & Blinder, V.S., 2018. The financial burden and distress of patients with cancer: Understanding and stepping-up action on the financial toxicity of cancer treatment. Available at: https://doi.org/10.3322/caac.21443.

Deshwal, V. & Sharma, M., 2019. Breast Cancer Detection using SVM Classifier with Grid Search Technique. *International Journal of Computer Applications*, 178(31), pp.18–23. Available at: https://doi.org/10.5120/ijca2019919157.

Doe, J., Smith, A. & Johnson, B., 2018. Breast cancer prediction using support vector machine with the sequential minimal optimization algorithm. *International Journal of Computer Applications*, 179(19), pp.16–19.

Iddrisu, M., Aziato, L. & Dedey, F., 2020. Psychological and physical effects of breast cancer diagnosis and treatment on young Ghanaian women: a qualitative study. Available at: https://doi.org/10.1186/s12888-020-02760-4.

Krishnamurthi, R., Aggrawal, N., Sharma, L., Srivastava, D. & Sharma, S., 2019. Importance of Feature Selection and Data Visualization Towards Prediction of Breast Cancer. *Recent Patents on Computer Science*, 12(4), pp.317–328.

Naji, M.A., Filali, S.E., Aarika, K., Benlahmar, E.H., Abdelouhahid, R.A. & Debauche, O., 2021. Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis. *Procedia Computer Science*, 191, pp.487–492. Available at: https://doi.org/10.1016/j.ijin.2022.05.002.

National Breast Cancer Foundation, 2023. Clinical Breast Exam. Available at: https://www.nationalbreastcancer.org/clinical-breast-exam.

Osei-Afriyie, S., Addae, A.K., Oppong, S., Amu, H., Ampofo, E. & Osei, E., 2021. Breast cancer awareness, risk factors, and screening practices among future health professionals in Ghana: A cross-sectional study. *PLOS ONE*, 16(6), p.e0253373. Available at: https://doi.org/10.1371/journal.pone.0253373.

Rubin, G., Vedsted, P. & Emery, J., 2011. Improving cancer outcomes: better access to diagnostics in primary care could be critical. *British Journal of General Practice*, 61(586), pp.317–318. Available at:

https://doi.org/10.3399/bjgp11x572283.

Sharma, D., Kumar, R. & Jain, A., 2022. Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning. *Measurement: Sensors*, 24, p.100560. Available at: https://doi.org/10.1016/j.measen.2022.100560.

Subhani, S., Pravalik, K. & Shravya, C., 2019. Prediction of Breast Cancer Using Supervised Machine Learning Techniques. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8(6).

Wilkinson, L. & Gathani, T., 2022. Understanding breast cancer as a global health concern. *The British Journal of Radiology*, 95(1130).

Yadav, A., Jamir, I., Jain, R.R. & Sohani, M., 2019. Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction - A Review. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp.979–985. Available at: https://doi.org/10.32628/cseit1952278.

Yedjou, C.G., 2021. Application of Machine Learning Algorithms in Breast Cancer Diagnosis and Classification. Available at: ttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8612371/.