NCCS Research Journal, Volume 3, No. 1 October 2024, ISSN NO: 2822-1605

Received Date: June 2024 *Revised: August 2024* *Accepted: October 2024*

Breast Cancer Prediction: A Comparative Study of Support Vector Machine and Logistic Regression

Nawaraj Paudel
Central Department of Computer Science and Information Technology,
Tribhuvan University
Email: nawarajpaudel@cdcsit.edu.np

Abstract

One of the most common malignancies among women worldwide is breast cancer and a key factor in raising survival rates is early identification. So, it is important to differentiate between malignant (cancerous) or benign (non-cancerous) tumors. Support Vector Machine (SVM) and Logistic Regression are popular machine learning models that has been widely used for binary classification problems including breast cancer prediction. This study explores the effectiveness of SVM and Logistic Regression in predicting breast cancer and compare their performances. This study uses Python programming to implement SVM and Logistic Regression to classify the Breast Cancer Wisconsin dataset from the UCI machine learning repository. Performance metrices such as recall, F1 score, accuracy, precision, and AUC-ROC have all been used to gauge how well these two algorithms work. Upon comparison, the result showed that SVM model outperformed Logistic Regression model on all the performance metrices.

Keywords: Breast Cancer, Logistic Regression, Support Vector Machine, Performance Metrics

Breast Cancer Prediction: A Comparative Study of Support Vector Machine and Logistic Regression

One of the most common diseases affecting women globally is breast cancer. It is the second largest disease that is responsible for women's death in the world. A precise and timely diagnosis is a crucial first step in recovery and care. This disease is caused by abnormal breast cells that proliferate and develop into tumors. If left untreated, tumors can spread throughout the body and become fatal. Breast cancer is still a complicated and common health issue that affects millions of people globally. Breast cancer is one of the many common malignant tumors that harm women. Breast cancer can grow and arise as a result of many internal and environmental factors. Poor lifestyle decisions, environmental circumstances, and social and psychological issues are linked to its prevalence. According to research, genetic abnormalities and family history account for 5% to 10% of breast cancer instances, whereas potentially modifiable variables account for 20% to 30% of cases. The cells of the breast are where breast cancer starts. A collection of cancer cells that has the potential to spread and destroy nearby tissue is called a malignant tumor (Obeagu & Obeagu, 2024).

Breast cancer diagnosis often relies on mammograms and biopsies. However, machine learning models can provide an additional layer of prediction to help clinicians make more informed decisions. For binary classification, two effective machine learning algorithms used are Support Vector Machine and Logistic Regression. Due to their adaptability and efficiency in a variety of applications with high-dimensional data and nonlinear relationships, these algorithms have been extensively used for binary classification tasks like distinguishing between malignant and benign tumors (Géron, 2017).

Based on a breast cancer dataset (Wolberg et al., 1993), this study focuses on applying SVM and Logistic Regression to classify malignant or benign tumors. Five metrics (Han et al., 2011), including accuracy, precision, recall, F1 score, and AUC-ROC, were used to compare these two algorithms. To determine which of these two algorithms is superior for predicting breast cancer, a comparison between them has finally been conducted.

Literature Review

Many researchers have recently been interested in applying machine learning algorithms for cancer prediction. This section summarizes some of the recent methods that have been widely used in cancer prediction. The authors of (Huang et al., 2017) evaluated the prediction performance of SVM and SVM ensembles using both small- and large-scale breast cancer datasets. For small-scale datasets, where feature selection should be done in the data pre-processing stage, linear kernel-based SVM ensembles based on the bagging method and RBF kernel-based SVM ensembles with the boosting method may be the better options, according to the experimental results based on accuracy, ROC, F-measure, and computational times of training. SVM ensembles based on boosting and RBF kernels outperformed the other classifiers on a big dataset. Authors (Jiang et al., 2023) compared SVM and Bayesian classification algorithms for breast cancer risk prediction. The test result showed that SVM outperformed the Bayesian classification algorithm in the actual target-tracking problem. For predicting the risk of breast cancer, the authors of (Jiang et al., 2023) contrasted SVM with the Bayesian classifier. According to the test results, SVM performed better in the real target tracking problem than the Bayesian classifier. A comparative analysis of data mining, deep learning, and machine learning algorithms for breast cancer prediction was reported by the authors in (Fatima et al., 2020). The main objective of this research was to assess and contrast several machine learning and data mining techniques that are currently in use to identify the most effective technique for managing large datasets with high prediction accuracy. Compiling the results of previous research on machine learning algorithms for breast cancer prediction was the main objective. The authors in (Islam et al., 2024) assessed and contrasted the classification accuracy, precision, recall, and F1 scores of five distinct machine learning techniques: XGBoost, Logistic Regression, Random Forest, Decision Tree, and Naive Bayes. They did this using a primary dataset of 500 patients from Dhaka Medical College Hospital. In this study, XGBoost outperformed other algorithms with an accuracy of 97%.

To improve breast cancer prediction using machine learning techniques, the authors of (Das et al., 2024) suggested an expert system called the "Machine Learning Based Intelligent System for Breast Cancer Prediction (MLISBCP)". The proposed approach makes use of the "Boruta" feature selection strategy to identify the most pertinent characteristics from the breast cancer dataset and the "K-Means SMOTE" oversampling method to address the class imbalance issue.

Accuracy, precision, recall, F1-score, and AUC-ROC score were used to assess MLISBCP's efficiency in comparison to a range of single classifier-based models, ensemble models, and models from the literature. This study concluded that the proposed model achieved the best accuracy of 97.53% when compared with other models.

The Wisconsin Breast Cancer Diagnostic dataset was utilized by the authors of (Khan et al., 2022) to identify breast cancer using a variety of machine learning algorithms. Various performance indicators were used to assess and compare the K-nearest neighbor, logistic regression, random forest, and decision tree algorithms. When the results are compared, it is shown that the logistic regression model yields the best results. Logistic regression has an accuracy of 98%, which is superior to the previously described method. The authors of (Zuo et al., 2023) evaluated several machine algorithms to determine which model was most effective at forecasting the recurrence of breast cancer. This research took eleven distinct machine learning (ML) algorithms to construct the prediction model. The area under the curve (AUC), accuracy, sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), and F1 score were used to evaluate the prognostic model's performance. Shapley Additive Explanation (SHAP) values were used to rank the feature importance and determine which machine learning model performed best. The AdaBoost algorithm was used to create the prediction model since it demonstrated the greatest prediction performance among the 11 algorithms when it came to accurately predicting the recurrence of breast cancer. Furthermore, it was discovered that the most crucial variables in the dataset for predicting the recurrence of breast cancer were CA125, CEA, Fbg, and tumor diameter.

A novel prediction model that utilizes machine learning techniques to accurately classify cases of breast cancer has been proposed by experimenting by the model using the WDBC breast cancer dataset. Based on accuracy, precision, recall, and f-measure, it was found that the proposed model performed better than other state-of-the-art machine-learning techniques (Wadhwa et al., 2023). To predict breast cancer, the authors in (Zhu, 2024) used the Light Gradient Boosting Machine (LightGBM) algorithm. The accuracy and speed of the LightGBM were both good. The bootstrap aggregating (Bagging) approach was used in this work to address the over-fitting issue. The study demonstrated how LightGBM can be used to create medical detection devices that are precise, quick, and affordable.

By evaluating the benefits and drawbacks of popular machine learning algorithms, authors in (Shengjie, 2024) developed and deployed a breast cancer prediction system that would increase the early detection rate of the disease and lower healthcare expenses. Furthermore, using the real development environment, authors developed a machine learning model appropriate for predicting breast cancer and conducts methodical testing and deployment. The findings of this study offered a novel technical method for the early detection of breast cancer in addition to significant experience in the use of machine learning in medical field.

With recall serving as the primary evaluation index, authors in (Chen et al., 2023) established various models to classify and predict breast cancer. The authors considered random forest, XGBoost, KNN, and logistic regression for classification. The goal was to serve as a reference for the early diagnosis of breast cancer. In order to assess and contrast the predictive impact of each model, this article also takes precision, accuracy, and F1-score evaluation markers into account. The Pearson correlation test was used to eliminate 15 features from the model's input in order to identify the ideal subset and raise the model's accuracy.

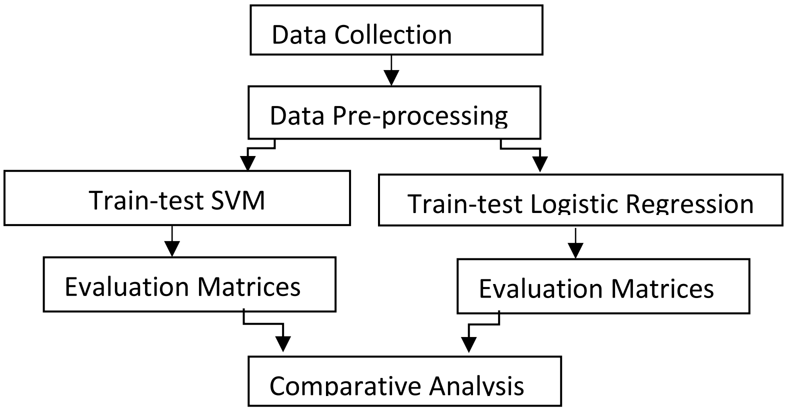
Using the Wisconsin breast cancer diagnostic dataset, authors in (Wei et al., 2023) provided a comparative examination of three machine learning models for breast cancer prediction: logistic regression, decision trees, and random forests. The results of the study demonstrated that, for the test dataset, the Random Forest model obtains the highest predicted accuracy of almost 95% and a cross-validation score of roughly 93%.

Methodology

Two machine learning models, SVM and Logistic Regression, were trained using a breast cancer Wisconsin data set and evaluated based on accuracy, precision, recall, f1-score, and AUC-ROC scores as shown in the figure below.

Figure 1:

Research Methodology



These two algorithms were implemented using Python programming language and its libraries such as numpy, pandas, matplotlib, seaborn, and sklearn. Python is a popular object-oriented, interpreted, high-level, dynamically-semantic programming language used for general-purpose work. Programmers may communicate their ideas in less lines of code because to its syntax, which was developed with the readability of code as a primary focus. Python is a programming language that facilitates faster work and more effective system integration. In recent years, Python has grown to be one of the most widely used programming languages worldwide. It has been applied to a wide range of tasks, including software testing, website development, and machine learning. Both developers and non-developers can use it.

Data Collection and Preprocessing

The Wisconsin Breast Cancer Dataset (Wolberg et al., 1993) is used in the study. It contains features that were taken from digital images of breast mass fine needle aspiration (FNA) procedures. The dataset contains thirty features including radius, texture, area, perimeter, smoothness, and compactness. To eliminate any omitted or unnecessary entries, the data is cleansed. The dataset is also normalized by applying feature scaling.

Model Development

Support vector machine and Logistic regression are employed as the primary predictive models in this research. Finding the optimal hyperplane in an N-dimensional space to partition data points into different feature space classes is the main objective of the SVM method. The hyperplane aims to keep as big a buffer as possible between the closest points of different classes. The dimension of the hyperplane is determined by the number of features. The equation of hyperplane is given as:

$W^T X + b = 0$

Here, **W** is a weight vector, **X** is input vector, and **b** is bias. The goal is to maximizing the margin. For linear SVM classifier, the output will be 1 if $W^T X + b \geq 0$ and 0 if $W^T X + b < 0$. Predictions and their probability are mapped using logistic regression using a logistic function known as the sigmoid function. An S-shaped curve that transforms any real value into a range between 0 and 1 is known as the sigmoid function. Moreover, the model predicts that the instance belongs to that class if the estimated probability produced by the sigmoid function exceeds a predetermined threshold on the graph. The model anticipates that the instance does not belong in the class if the calculated probability is less than the predetermined threshold.

$Logit(pi) = \frac{1}{1+\exp(-pi))}$ -------------------------------------- Equation 1

$\ln\left(\frac{pi}{1-pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$ --------------------------Equation 2

Here Logit(pi) is the dependent variable and X is the independent variable and βi are coefficients.

A training set and a testing set were created from the dataset in order to evaluate each model's performance.

In this case, the training set contained 70% of the data, whereas the testing set contained 30% of the data.

Performance Evaluation

Selecting an appropriate metric is essential when assessing machine learning (ML) models. After a machine learning algorithm has been put into practice, the next stage is to determine the model's effectiveness using metrics and datasets. Various machine learning algorithms are assessed using different performance indicators. The most common metrics are accuracy, precision, recall, F1-score, and AUC-ROC. The performance of both machine learning models has been assessed using these matrices. To calculate the value of different performance indicators, a confusion matrix is used. Confusion matrices, which are frequently used to assess the effectiveness of classification models, which seek to predict a categorical label for each input instance, are matrices that summaries the performance of a machine learning model on a set of test data. According to the model's predictions, they indicate the proportion of accurate and inaccurate instances. The number of instances that the model generated on the test data is shown in the matrix.

- True Positive (TP): When a positive outcome is accurately predicted by the model, the actual result is also positive.
- True Negative (TN): When a negative result is accurately predicted by the model, the real result is also negative.
- False Positive (FP): When a positive result is predicted by the model but the actual result is negative. Likewise referred to as a Type I mistake.
- False Negative (FN): When a positive result occurs instead of the expected negative one, the model predicted the wrong thing. Likewise referred to as a Type II mistake.

How often a machine learning model predicts the outcome accurately is measured by its accuracy. It is calculated by dividing the number of correct predictions by the total number of predictions.

Accuracy = $\frac{TP+TN}{TP+FN+FP+TN}$

The quality of a positive prediction produced by the model is referred to as precision. In other words, the proportion of observations that fall under the category of good emotion that are truly in that category.

Precision = $\frac{TP}{TP+FP}$

The frequency with which a machine learning model properly selects positive examples from among all of the real positive samples in the dataset is measured by a statistic called recall.

Recall = $\frac{TP}{TP+FN}$

The F-measure, which is the harmonic mean, combines the measurements of recall and precision.

F-measure = $\frac{2\times precision\times recall}{pression+recall}$

The area under the ROC curve is called the AUC-ROC score.

Figure 2.

Confusion Matrix of Support Vector Machine

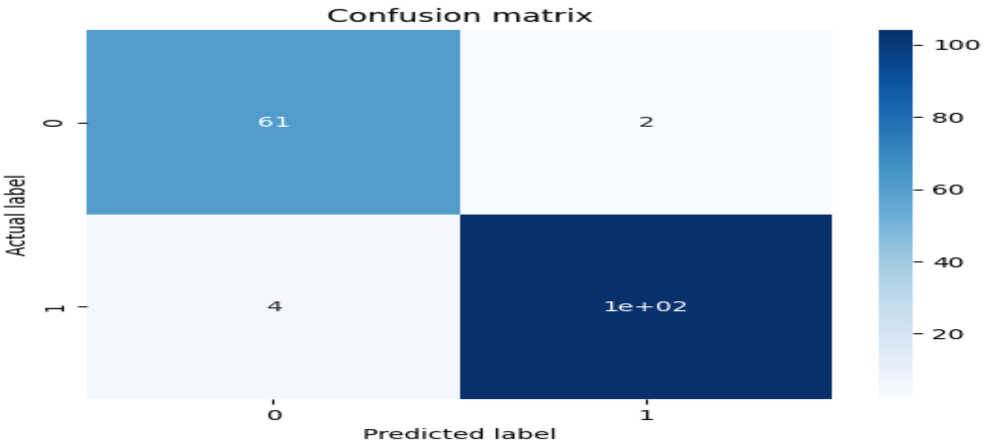
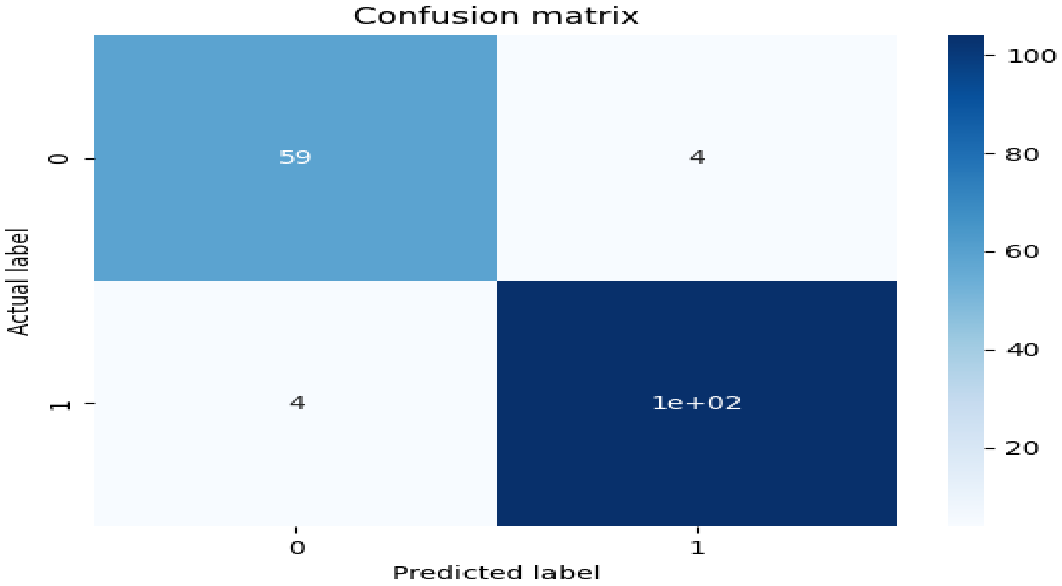


Figure 3.
Confusion Matrix of Logistic Regression



It summarizes the model's ability to provide relative scores that distinguish between positive and negative examples across all categorization levels. The AUC-ROC score has a range of 0 to 1, with 1 denoting ideal performance and 0.5 representing random guessing.

The Figure 2 and Figure 3 above show confusion matrix of the classification report that has been obtained after testing SVM and Logistic Regression respectively on the test dataset.

The Table 1 below shows accuracy, precision, recall, F1-score, and AUC-ROC scores of both SVM and Logistic regression.

Table 1.

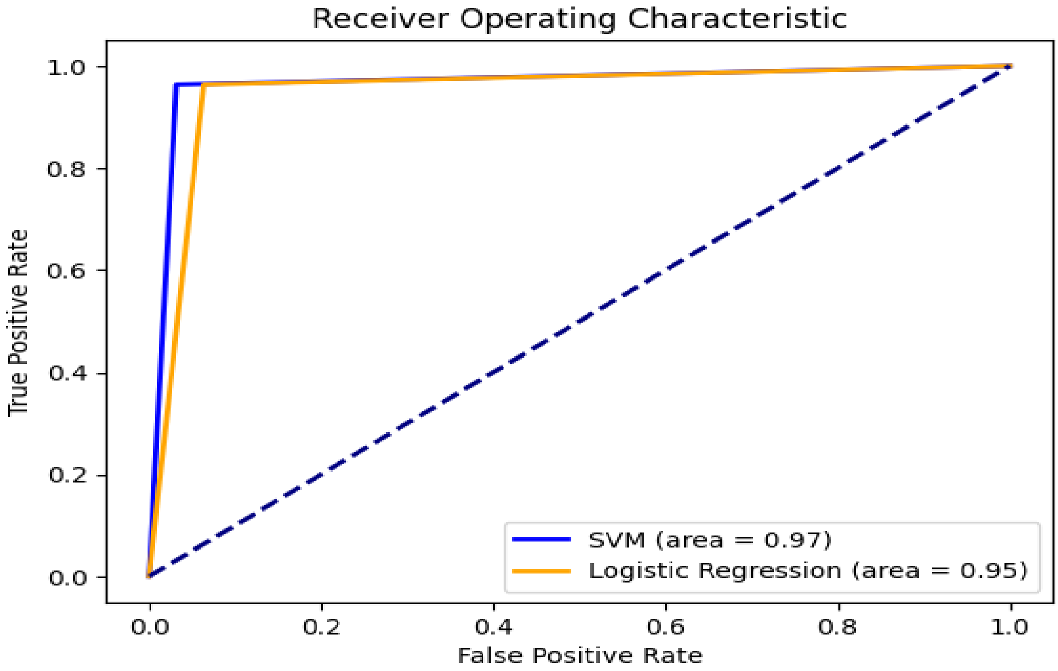
Accuracy, Precision, Recall, F1-score, AUC-ROC score of both models

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
SVM	0.965	0.981	0.963	0.972	0.966
Logistic Regression	0.953	0.963	0.963	0.963	0.950

The AUC-ROC curve for the models is displayed in Figure 4 below. This curve demonstrates that SVM is a more accurate predictor of breast cancer than logistic regression. The area under the ROC curve is known as the AUC-ROC score. It combines the relative scores that a model can generate to determine whether an occurrence is good or negative across all classification criteria. The AUC-ROC score has a range of 0 to 1, with 1 denoting ideal performance and 0.5 representing random guessing.

Figure 4.

AUC-ROC curve of SVM and Logistic Regression



Results And Discussion

The study highlights the importance of using SVM and logistic regression in clinical settings due to its simplicity and the ability to interpret the results easily. While other complex models may provide slightly better accuracy, SVM and logistic regression remain strong contenders for breast cancer prediction due to its transparency and ease of use.

- SVM achieved an accuracy of approximately 97% on the test dataset.
- The ROC-AUC score for SVM was found to be 0.966, indicating a strong predictive performance.
- Compared to logistic regression, SVM showed competitive results, especially in terms of interpretability and simplicity.

Conclusion

This study examined both SVM and logistic regression models for predicting breast cancer using the Breast Cancer Wisconsin (Original) dataset from the UCI machine learning repository. Performance of both of these algorithms were evaluated using F1 score, AUC-ROC, recall, accuracy, and precision. The accuracy, precision, recall, f1-score, and AUC-ROC scores of SVM were 0.965, 0.981, 0.963, 0.972, and 0.966 respectively. These scores of Logistic Regression were 0.953, 0.966, 0.963, 0.963, and 0.95 respectively. The final result of this study demonstrates that the SVM model slightly improves classification accuracy compared to Logistic Regression model in predicting breast cancer.

References

Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y., & Cai, G. (2023). Classification Prediction of Breast Cancer Based on Machine Learning. *Computational Intelligence and Neuroscience, 2023*. https://doi.org/10.1155/2023/6530719

Das, A. K., Biswas, S. K., Mandal, A., Bhattacharya, A., & Sanyal, S. (2024). Machine Learning based Intelligent System for Breast Cancer Prediction (MLISBCP). *Expert Systems with Applications*, *242*. https://doi.org/10.1016/j.eswa.2023.122673

Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. In *IEEE Access* (Vol. 8). https://doi.org/10.1109/ACCESS.2020.3016715

Géron, A. (2017). Hands-on Machine Learning. In *O'Reilly Media, Inc* (Vol. 53, Issue 9).

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann Publishers Inc.

Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). SVM and SVM ensembles in breast cancer prediction. *PLoS ONE*, *12*(1). https://doi.org/10.1371/journal.pone.0161501

Islam, T., Sheakh, Md. A., Tahosin, Mst. S., Hena, Most. H., Akash, S., Bin Jardan, Y. A., FentahunWondmie, G., Nafidi, H.-A., & Bourhia, M. (2024). Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI. *Scientific Reports*, *14*(1), 8487. https://doi.org/10.1038/s41598-024-57740-5

Jiang, J., Li, H., Qi, H., & Wu, W. (2023). Comparison between Bayesian and SVM model for breast cancer risk prediction. *Applied and Computational Engineering*, *5*(1). https://doi.org/10.54254/2755-2721/5/20230555

Khan, M. M., Islam, S., Sarkar, S., Ayaz, F. I., Kabir, Md. M., Tazin, T., Albraikan, A. A., & Almalki, F. A. (2022). Machine Learning Based Comparative Analysis for Breast Cancer Prediction. *Journal of Healthcare Engineering*, *2022*(1), 4365855. https://doi.org/https://doi.org/10.1155/2022/4365855

Obeagu, E. I., & Obeagu, G. U. (2024). Breast cancer: A review of risk factors and diagnosis. *Medicine*, *103*(3). https://journals.lww.com/md-journal/fulltext/2024/01190/breast_cancer__a_review_of_risk_factors_and.67.aspx

Shengjie, W. (2024). Design and implementation of breast cancer prediction system based on machine learning. *Theoretical and Natural Science*, *32*(1). https://doi.org/10.54254/2753-8818/32/20240871

Wadhwa, K., Singh, S., Sharma, A., & Wadhwa, S. (2023). Machine Learning-Based Breast Cancer Prediction Model. *International Journal of Performability Engineering*, *19*(1). https://doi.org/10.23940/ijpe.23.01.p6.5563

Wei, Y., Zhang, D., Gao, M., Tian, Y., He, Y., Huang, B., & Zheng, C. (2023). Breast Cancer Prediction Based on Machine Learning. *Journal of Software Engineering and Applications*, *16*(08). https://doi.org/10.4236/jsea.2023.168018

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). *Breast Cancer Wisconsin (Diagnostic)*.

Zhu, J. (2024). Breast cancer prediction based on the machine learning algorithm LightGBM. *Applied and Computational Engineering*, *40*(1). https://doi.org/10.54254/2755-2721/40/20230630

Zuo, D., Yang, L., Jin, Y., Qi, H., Liu, Y., & Ren, L. (2023). Machine learning-based models for the prediction of breast cancer recurrence risk. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02377-z