MJSAT

Malaysian Journal of Science and Advanced Technology

journal homepage: https://mjsat.com.my/

Performance Evaluation of Data Mining Classification Algorithms for Predicting Breast Cancer

Nyme Ahmed*[1], Rifat-Ibn-Alam[1], and Syed Nafiul Shefat[1]

[1]*Dept. of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh*

KEYWORDS

Performance Evaluation
Breast Cancer
Data Mining
Medical Science
Classification

ARTICLE HISTORY

Received 5 July 2022
Received in revised form 14 July 2022
Accepted 17 July 2022
Available online 24 July 2022

ABSTRACT

The most prevalent cause of death among women is breast cancer. At an early stage, predicting breast cancer enhances the probability of a successful cure. It requires a breast cancer prediction technology capable of classifying a breast tumor as dangerous malignant or harmless benign. This is especially true in the medical field, where classification methods are often used for finding and investigation to make decisions for the disease. This study examines the performance of six classification algorithms of data mining which are Logistic Regression classifier, Naïve Bayes classifier, Decision Tree, Random Forest Classifier, Support Vector Machine, and K-Nearest Neighbors on the Wisconsin Breast Cancer (original) dataset. The principal purpose is to measure the performance of each algorithm in terms of their accuracy, precision, sensitivity, and specificity. The findings indicate that the accuracy of Support Vector Machine has the greatest rate (97.20 %) and the lowest error rate when determining if a woman has a malignant or benign tumor.

1. INTRODUCTION

Nowadays, breast cancer is growing increasingly frequent among women. It should be predicted early on so that proper countermeasures may be taken. Breast cancer prediction utilizing various data mining approaches is a critical necessity. [1] The second biggest cause of death among women is Breast cancer, behind lung cancer. It is valued that breast cancer is responsible for around 12% of all new cancer cases, and 25% of all malignancies in females. [2] The development of malignant tumors, which are more often referred to as cancer, is a significant cause of death around the world. The problem is exacerbated in underdeveloped or developing countries, where there are not enough experienced doctors or physicians to undertake a cancer prognosis.

Even though early identification of breast cancer doubles the probability of survival, many women die of this disease. As a result, much research are now underway to develop techniques for predicting breast cancer in its early stages. The supervised learning approach calls for the development of a model for analyzing prior performance.

Statistical regression, classification, and association rules are all supervised learning approaches that are employed in medical and clinical research. This study employs categorization methods from the field of medicine. It begins by classifying the data set and then produces the optimal method for identifying and forecasting breast cancer. Prediction starts with the identification of symptoms in patients, followed by the identification of ill individuals among a huge number of healthy and sick patients [3].

The main objective of this study is to assess data from a dataset of breast cancer to predict the class properly in each case using six classification methods: Logistic Regression classifier, Naïve Bayes classifier, Decision Tree, Random Forest Classifier, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). This paper makes three major contributions: it identifies which classifier is best for breast cancer prediction, compares multiple data mining approaches on a breast cancer data set, and identifies the optimal performance-based strategy for disease prediction.

*Corresponding author:
E-mail address: Nyme Ahmed < nymeahmedhimu@gmail.com >.

There are five sections in this paper. Each section follows the structure outlined below: Introduction, followed by a brief survey of relevant works on the use of data mining in Brest Cancer prediction in section 2. Section 3 contains the methodology. Section 4 contains the findings of this research, as well as a discussion of the findings, and section 5 includes the conclusion.

2. LITERATURE REVIEW

Data mining relies heavily on classification to do its work effectively. Many studies using data mining on various medical datasets have been carried out to classify breast cancer. Several of them have a high degree of accuracy in their categorizations.

The goal of this study [4] was to find out how breast cancer and certain traits are related to lower the chance of dying from breast cancer. Both the Breast Cancer Coimbra Dataset and the Wisconsin Breast Cancer Dataset were utilized. A Decision Tree, Support Vector Machine, Random Forest, Neural Network, and Logistics Regression were employed to classify these datasets. To compare the five classification models, we used prediction accuracy, F-measure, and AUC. In a comparison study, the random forest model outperformed the other four models tested. On the Wisconsin Breast Cancer dataset, the authors of this study [5] evaluated five supervised machine learning methods: support vector machine, artificial neural networks, K-nearest neighbors, random forests, and logistic regression. Precision, F1 score, negative predictive value, false-negative and false-positive rates, and Matthews Correlation Coefficient are all used to measure the study's overall success. When it came to precision and F1 score for the identical set of input parameters, ANNs scored the highest at 98.57%, followed by SVMs at 97.14% and 0.98777. NB and KNN were two separate classifiers utilized by the authors to classify breast cancer in this paper [6]. The data they used was from the Wisconsin Breast Cancer Study. Cross-validation was used to compare the two new implementations and assess how accurate they were. With 97.51% accuracy in comparison to 96.19%, the KNN classifier was shown to be the most accurate.

To assess which machine learning algorithms, perform best on the Wisconsin Breast Cancer datasets, this paper [7] compared the performance of SVM, Naive Bayes, Decision Tree, and k Nearest Neighbors. Classifying data and determining the accuracy, precision, sensitivity, and specificity for each method was the study's primary objective for its authors. Research shows that SVM is the most accurate and error-free method for predicting outcomes. The purpose of this research [8] was to identify a suitable model for predicting the presence of breast cancer based on the clinical data of numerous patients. SVM, ANN, Naive Bayes classifier, and AdaBoost tree were some of the data mining models employed in this paper, and each one was thoroughly described in the article. Two widely used test data sets, the Wisconsin Diagnostic Breast Cancer and, the Wisconsin Breast Cancer were utilized to assess the performance of these models. With this approach, we could estimate the test error by 10 times. The results of this investigation indicated a wide-ranging trade-off between various methods and a full evaluation of the models.

Four machine learning algorithms were tested on the Wisconsin Breast Cancer dataset by the study's [9] author to see which one was the most effective. Support Vector Machine, Naive Bayes Decision Tree, and k Nearest Neighbors are the four algorithms that were used in this study. They find that, of these four algorithms, the Support Vector Machine algorithm is the best one for analyzing how likely it is that someone will get cancer. [10] The author used four machine learning algorithms to identify breast cancer in this study. SVM and K-Nearest Neighbors are just a few of the algorithms that fall under this category (K-NN). Their findings show that SVM is the most accurate of these four algorithms, with a 97% accuracy rate. In order to classify the Wisconsin Breast Cancer dataset, the author of this research [11] employed two of the most prevalent machine classification techniques. The Support Vector Machine and the Artificial Neural Network were utilized to achieve this. There's no doubt that the Support Vector Machine method yields both the best performance and accuracy. Using big data, the authors of this work [12] addressed the problem of breast cancer prediction in a large data setting. Their research focused on gene expression and DNA methylation data. The goal was to use each dataset on its own and with other datasets to scale up the machine-learning algorithms that were used to put things into categories. They employed SVM, decision tree, and random forest classification algorithms to create nine breast cancer prediction models. This study examined the accuracy and error rate of three different data sets: one each from GE and DM and a combination of the two. They also tested two platforms (Spark and Weka) to determine how they handled large data sets. On the GE dataset, the scaled SVM classifier in the Spark environment was more accurate and made less mistakes than the other classifiers. According to the authors of this paper [13], predictive information was extracted from routinely gathered demographic, clinical, and biochemical data of breast cancer patients using an ML-based decision support system paired with random optimization. A DSS model was constructed using training set data. The model's performance was then evaluated on a testing set, which yielded a C-index for progression-free survival of 0.84 and an accuracy of 86%.

This paper [14] explored SVM (radial basis kernel), ANN, and Naive Bayes using the WDBC Dataset. In this paper, the performance of different machine learning approaches and methods for selecting and extracting features were compared to find the best approach. We came up with the notion of combining dimensionality reduction with artificial intelligence. After reducing the number of features using LDA, this study used the reduced feature dataset to train an SVM to detect breast cancer. The proposed approach had 98.82% accuracy, 98.41% sensitivity, and 99.07% specificity. [15] compared the performance of three machine learning algorithms which are Support Vector Machine, K-nearest neighbors and Decision tree to see which classifier is better for classifying breast cancer. A dataset from the Wisconsin Breast Cancer (Diagnostic) Registry was also used in this study.

The primary objective of this research was to compare a wide range of classifiers to determine which one provided the best accuracy. Quadratic support vector machines were shown to have the highest accuracy (98.1 percent) and the lowest false discovery rates in this study. This article [16] examined the performance of supervised learning classifiers such as Naive Bayes, SVM-RBF kernel, Decision trees, RBF neural networks, and basic CART to determine the most effective classifier for breast cancer datasets. With a score of 96.84% on the Wisconsin Breast Cancer datasets, the SVM-RBF kernel was the most accurate classifier.

The authors of this paper [17] used three distinct data mining classification algorithms to predict breast cancer. When predicting cancer, it relied on several factors. For better prediction, however, the emphasis was on precision and processing efficiency. The investigations determined that Nave Bayes was superior to Decision Trees and K-Nearest Neighbor since it required the least amount of time to calculate (0.02 seconds) and provided the best level of accuracy. C4.5, Naive Bayes, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) classifiers were all examined in this paper [18] in order to determine which one was the most effective in classifying data. SVM was shown to be the most accurate classifier, with a 96.99 percent success rate.

3. METHODOLOGY

This work follows the quantitative-experimental methodology. Because the dataset we are using here is a combination of numeric data and we are going to identify numeric results also. The quantitative method employs statistical, logical, and mathematical techniques to generate numerical data. The main goal of our study is to analyze the performance of data mining classification algorithms for predicting breast cancer using a numeric dataset.

The dataset of breast cancer was obtained from the Wisconsin Breast Cancer Diagnostic database, which can be accessed through Kaggle. There are 569 rows and 32 characteristics in this table. In this situation, the Class Attribute is "diagnostic," which has two associated Class Levels: "M" and "B." The letter 'M' stands for Malignant and the letter 'B' for Benign. The number of malignant cases is 212, whereas the number of benign cases is 357. In our study, benign instances are classified as positive, whereas malignant ones are classified as negative.

Preprocessing of data is a phase in the data mining and data analysis process that involves taking raw data and transforming it into an arrangement that can be interpreted and analyzed by computers and different algorithms. The dataset has 13 missing values for several attributes. These missing values are handled in preprocessing phase. Other than that, the dataset has all the values available. On the other hand, the 'id' column has all unique values. So this column is not necessary for classifications. The class attribute is 'diagnosis' where it has two possible labels- The letters 'M' and 'B' stand for Malignant and Benign, respectively. The preprocessing step additionally encodes these string values for the class label using the values 0 (B) and 1 (M).

The dataset has been divided into two sections. These are the training and testing stages. About 75% of the data are put to use in training, while the remaining 25% are put to use in testing, which is picked randomly.

Six classification algorithms are applied to the training set to construct a model named Logistic Regression, Naïve Bayes, Decision Tree, Random Forest Classifier, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Based on those models, prediction is done on the test set. The flowchart of the proposed work is given below-

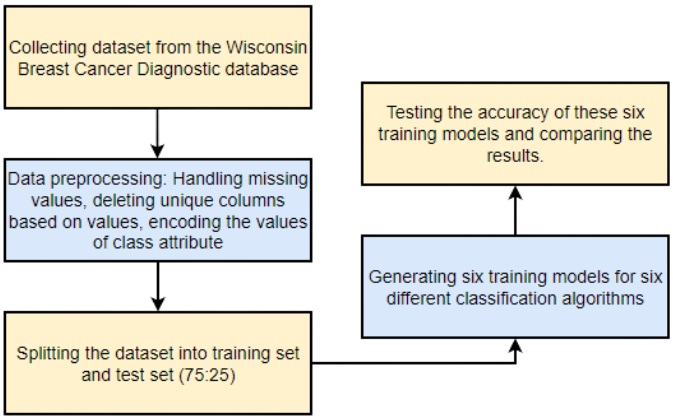

Fig. 1. Flowchart of proposed work.

Furthermore, a brief distinction is made on which algorithm is the most accurate one. After that, the predicted result of the most accurate algorithm is compared with the actual result. These six algorithms are selected because they are the most widely used classification algorithms in the breast cancer prediction arena by previous researchers.

4. Results and Discussions

As indicated before, this research employs six distinct data mining classification algorithms to recognize the tumor type, whether malignant or benign by using the Wisconsin Breast Cancer Dataset. We used 75% of the dataset as training data and 25% as test data.

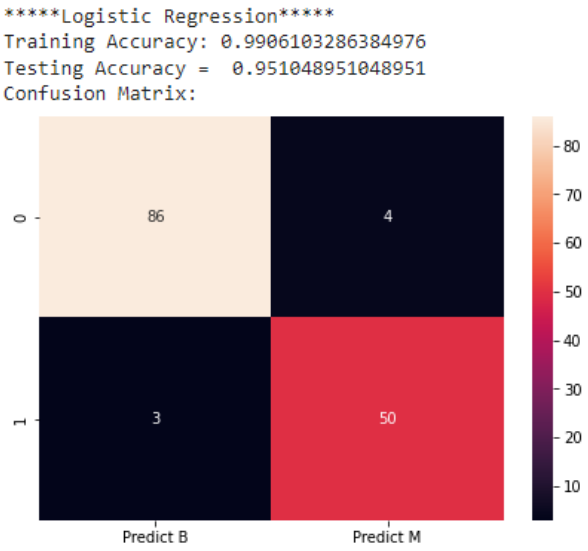

Fig. 2. Accuracy and Confusion Matrix for Logistic Regression.

Logistic Regression was applied to the training set to develop a model. Fig 2 demonstrates that the accuracy of this model in the training set is 99.06%. Then this model is used to predict the outcome on the test set and its accuracy on the test set is 95.10%. Hence the confusion matrix is also given where True Positive (TP) = 86, False Positive (FP) = 4, False Negative (FN) = 3 and True Negative (TN) = 50.

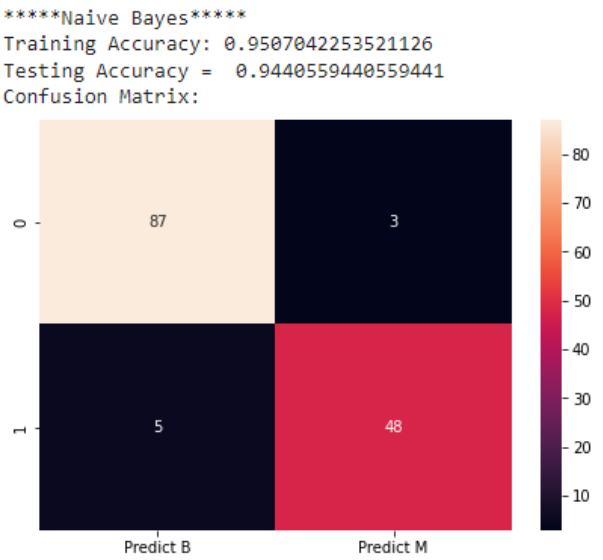

Fig. 3. Accuracy and Confusion Matrix for Naïve Bayes.

Here, Naïve Bayes was applied to the training set. Fig 3 shows that the accuracy of this model in the training set is 95.07%. Then this model is used to predict the outcome on the test set and its accuracy on the test set is 94.41%. Hence the confusion matrix is also given where TP= 87, FP= 3, FN= 5 and TN= 48.

```
*****Decision Tree Classifier*****
Training Accuracy: 1.0
Testing Accuracy =  0.9370629370629371
Confusion Matrix:
```

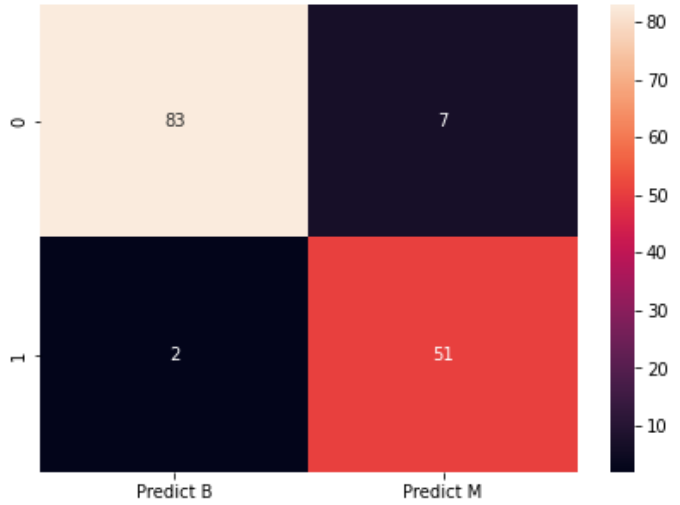

Fig. 4. Accuracy and Confusion Matrix for Decision Tree.

Decision Tree is used to build a model using the training set. As shown in fig 4, this model's accuracy in the training set is 100%. The model is then used to predict the result of the test set, with an accuracy of 93.71%. The confusion matrix is also demonstrated where TP=83, FP=7, FN=2, and TN=51.

```
*****Random Forest Classifier*****
Training Accuracy: 0.9953051643192489
Testing Accuracy =  0.965034965034965
Confusion Matrix:
```

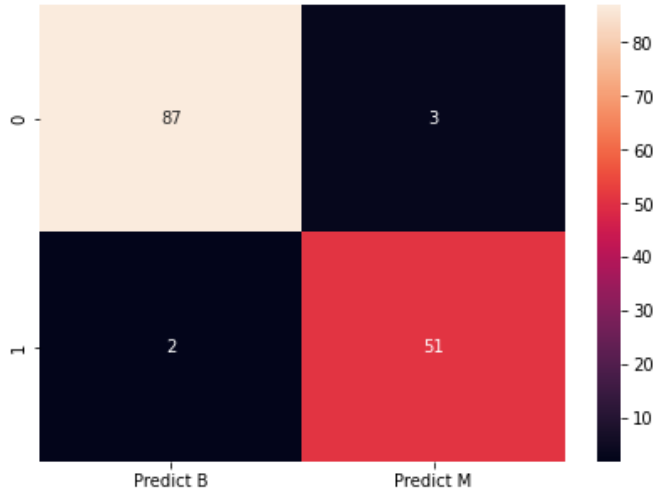

Fig. 5. Accuracy and Confusion Matrix for Random Forest.

To build a model, the Random Forest classifier is applied to the training data. As shown in fig 5, this model's accuracy in the training set is 99.53%. The model is then used to predict the result on the test set, with an accuracy of 96.50 %. The confusion matrix is also provided where TP=87, FP=3, FN=2, and TN=51.

```
*****Support Vector Machine*****
Training Accuracy: 0.9859154929577465
Testing Accuracy =  0.972027972027972
Confusion Matrix:
```

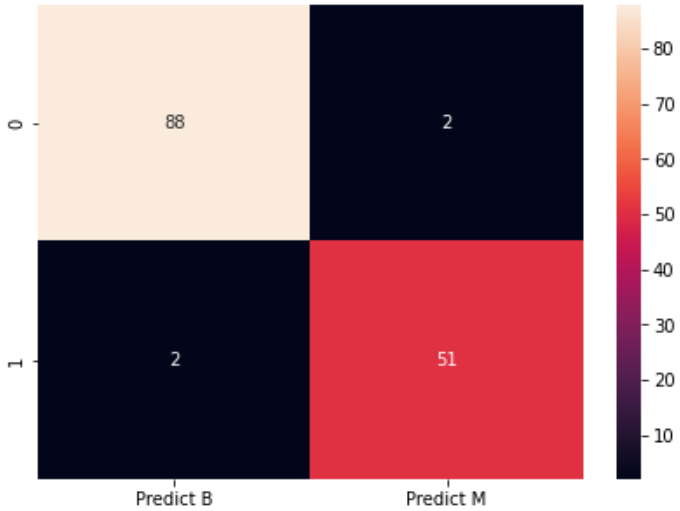

Fig. 6. Accuracy and Confusion Matrix for Support Vector Machine.

On the training set, a Support Vector Machine (SVM) is used to develop a model. As shown in fig 6, this model's accuracy in the training set is 98.59 %. The model is then used to forecast the result on the test set, with an accuracy of 97.20 %. The confusion matrix is also given where TP= 88, FP= 2, FN= 2, and TN= 51.

```
*****K-Nearest Neighbors*****
Training Accuracy: 1.0
Testing Accuracy =  0.951048951048951
Confusion Matrix:
```

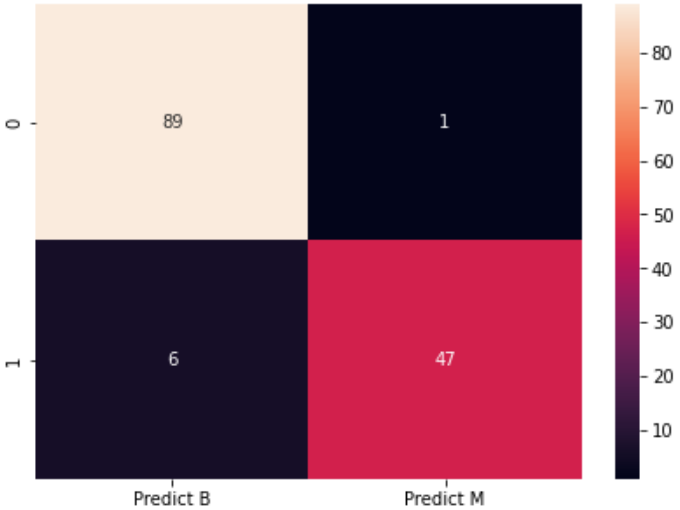

Fig. 7. Accuracy and Confusion Matrix for K-Nearest Neighbor.

To build a model, the K-Nearest Neighbors method was used for the training data. As shown in fig 7, this model's accuracy in the training set is 100%. Then, this model is utilized to predict the result on the test set, with a 95.10% accuracy. The confusion matrix is are also illustrated where TP=89, FP=1, FN=6, and TN=47.

Table 1. Distinction of the confusion matrix of all six algorithms

Algorithm	True Negative (TN)	False Positive (FP)	False Negative (FN)	True Positive (TP)	Total Support
Logistic Regression	86	4	3	50	143
Naïve Bayes	87	3	5	48	
Decision Tree	83	7	2	51	
Random Forest	87	3	2	51	
SVM	88	2	2	51	
KNN	89	1	6	47	

Table 1 depicts the value of the confusion matrix for all six algorithms in the testing set. These data are collected from fig 2-7. Confusion matrixes are used to summarize the outcomes of predictions on categorization problems. This matrix displays the many ways in which your classification model becomes confused while it is attempting to forecast. TP is for predicted positive and is correct, FP stands for predicted positive and is incorrect, FN means predicted negative and is incorrect, and TN indicates predicted negative and is correct. Here, KNN got the highest number of rows for TP and also the lowest number of rows for TN.

Table 2. Distinction between all six algorithms

Algorithm	True Positive Rate	False Negative Rate	True Negative Rate	False Positive Rate	Precision	Recall	F1 Score	Accuracy	Error Rate
Logistic Regression	0.94	0.06	0.96	0.04	0.93	0.94	0.93	95.10%	4.9%
Naïve Bayes	0.91	0.09	0.97	0.03	0.94	0.91	0.92	94.41%	5.59%
Decision Tree	0.96	0.04	0.92	0.08	0.88	0.96	0.92	93.71%	6.29%
Random Forest	0.96	0.04	0.97	0.03	0.94	0.96	0.95	96.50%	3.5%
SVM	0.96	0.04	0.98	0.02	0.96	0.96	0.96	97.20%	2.8%
KNN	0.89	0.11	0.99	0.01	0.98	0.89	0.93	95.10%	4.9%

Table 2 illustrates all the values of true positive rate, false positive rate, true negative rate, false positive rate, precision, recall, f1-score, accuracy, and error rate of the six algorithms applied on the dataset. These values are calculated from the confusion matrix of table 01. This table depicts that, Support Vector Machine outperforms all the algorithms with the highest accuracy of 97.20% whereas the Decision Tree shows the least accuracy (93.71%).

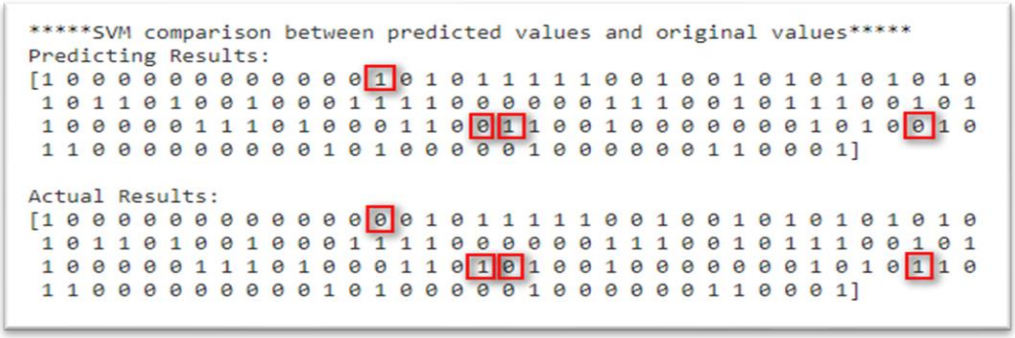

Fig. 8. Comparison between the Predicted and Actual results using SVM.

As Support Vector Machine (SVM) has the greatest accuracy for diagnosing breast cancer than other five algorithms, hence a slight contrast is created between the predicted values and the real ones which is shown in fig 8. As mentioned earlier, there are two class levels linked with the class attribute "diagnostic," which are "M" and "B.". These string values for the class label are encoded using the values 0 (B) and 1 (M). In fig 8, some forecasted values do not match the actual values since this method is not 100% accurate and have an error rate (as in Table 2).

5. CONCLUSION

This research investigated six different data mining classification algorithm for breast cancer prediction which are Logistic Regression classifier, Naïve Bayes classifier, Decision Tree, Random Forest Classifier, Support Vector Machine, and K-Nearest Neighbors. Each of the six methods was presented in its fundamental characteristics and operation. Support Vector Machine (SVM) achieved the greatest accuracy (97.2%), whereas Decision Trees was found to have the lowest accuracy (93.71%). In the medical industry, diagnosis is an exceedingly expensive and time-consuming procedure. This research implies that data mining classification algorithms may act as a medical assistant for breast cancer reorganization, which can be highly valuable for doctors in the case of cancer treatment. The generated Support Vector Machine (SVM) model outperformed all previous algorithms in terms of consistency, and it has the potential to improve in the field of breast cancer prediction and its treatment. Based on the outcomes of this research, we can infer that data mining classification algorithms are capable of automatically diagnosing diseases with high accuracy.

REFERENCES

[1] Khourdifi, Y. and Bahaj, M., 2018, December. Applying best machine learning algorithms for breast cancer prediction and classification. In 2018 International conference on electronics, control, optimization and computer science (ICECOCS) (pp. 1-5). IEEE.

[2] Asri, H., Mousannif, H., Al Moatassime, H. and Noel, T., 2016. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, pp.1064-1069.

[3] Jamal, A., Handayani, A., Septiandri, A.A., Ripmiatin, E. and Effendi, Y., 2018. Dimensionality reduction using pca and k-means clustering for breast cancer prediction. Lontar Komput. J. Ilm. Teknol. Inf, 9(3), p.192.

[4] Li, Y. and Chen, Z., 2018. Performance evaluation of machine learning methods for breast cancer prediction. Appl Comput Math, 7(4), pp.212-216.

[5] Islam, M., Haque, M., Iqbal, H., Hasan, M., Hasan, M. and Kabir, M.N., 2020. Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science, 1(5), pp.1-14.

[6] Amrane, M., Oukid, S., Gagaoua, I. and Ensari, T., 2018, April. Breast cancer classification using machine learning. In 2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT) (pp. 1-4). IEEE.

[7] Asri, H., Mousannif, H., Al Moatassime, H. and Noel, T., 2016. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, pp.1064-1069.

[8] Wang, H. and Yoon, S.W., 2015. Breast cancer prediction using data mining method. In IIE Annual Conference. Proceedings (p. 818). Institute of Industrial and Systems Engineers (IISE).

[9] Bharat, A., Pooja, N. and Reddy, R.A., 2018, October. Using machine learning algorithms for breast cancer risk prediction and diagnosis. In 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C) (pp. 1-4). IEEE.

[10] Khourdifi, Y. and Bahaj, M., 2018, December. Applying best machine learning algorithms for breast cancer prediction and classification. In 2018 International conference on electronics, control, optimization and computer science (ICECOCS) (pp. 1-5). IEEE.

[11] Bayrak, E.A., Kırcı, P. and Ensari, T., 2019, April. Comparison of machine learning methods for breast cancer diagnosis. In 2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT) (pp. 1-3). IEEE.

[12] Alghunaim, S. and Al-Baity, H.H., 2019. On the scalability of machine-learning algorithms for breast cancer prediction in big data context. IEEE Access, 7, pp.91535-91546.

[13] Ferroni, P., Zanzotto, F.M., Riondino, S., Scarpato, N., Guadagni, F. and Roselli, M., 2019. Breast cancer prognosis using a machine learning approach. Cancers, 11(3), p.328.

[14] Omondiagbe, D.A., Veeramani, S. and Sidhu, A.S., 2019, April. Machine learning classification techniques for breast cancer diagnosis. In IOP Conference Series: Materials Science and Engineering (Vol. 495, No. 1, p. 012033). IOP Publishing.

[15] Obaid, O.I., Mohammed, M.A., Ghani, M.K.A., Mostafa, A. and Taha, F., 2018. Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. International Journal of Engineering & Technology, 7(4.36), pp.160-166.

[16] V. Chaurasia and S. Pal, "Data Mining Techniques : To Predict and Resolve Breast Cancer Survivability," vol. 3, no. 1, pp. 10–22, 2014.

[17] Chintan Shah; Anjali G. Jivani "Comparison of data mining classification algorithms for breast cancer prediction",pp.1- 4,2013.

[18] Y. Christobel, A., & Sivaprakasam, "An empirical comparison of data mining classification methods," Int. J. Comput. Inf. Syst., vol. 3, no. 2, pp. 24–28, 2011.