# Assessment of Factors Influencing the Survival of Breast Cancer Patients using a Machine Learning Approach

**Shivani Motarwar, Dixshant Kumar Jha**

***Abstract:*** ***Breast cancer is one of the deadliest diseases, claiming approximately 627,000 lives worldwide in 2018–2019. Therefore, early detection of breast cancer through automation in the prediction of the disease will help the medical industry to cure this disease at an early stage and thereby reduce the risk of death drastically. In the present study, the Breast Cancer Wisconsin (Diagnostic) Data Set has been taken from the University of California Irvine (UCI) Machine Learning Repository. The dataset (n=699) contained a total of 30 predictor parameters and one dependent parameter. The dependent variable referred to the type of cancer tissue, i.e., benign or malignant. To predict the type of cancer tissue present in the patient, prediction models were built using 1) Logistic Regression (LR), 2) Decision Tree Classifier (DTC), 3) Random Forest Classifier (RFC), 4) K Nearest Neighbor (KNN), 5) Support Vector Machine (SVM), and 6) Ada Boost Classifier (ABC). To improve the accuracy of the model, a correlation matrix was used and the top 8 features were selected. To improve the accuracy even further, the Synthetic Minority Oversampling Technique (SMOTE) was used to eliminate the problem of class imbalance, and then accuracy was compared before and after SMOTE. The Precision, Recall, and F1 scores are the performance metrics that have been taken into consideration for selecting the best model for the analysis. The results of the study reveal that the KNN algorithm gives the highest accuracy of 95.321% after the SMOTE technique is applied to each of the six algorithms. It has been revealed that while SMOTE aids in the accuracy of some algorithms, it affects the performance of others. This research may be turned into realistic tools that can be utilized in the medical field to more accurately predict the stage of disease for better treatment management.***

***Keywords: Breast cancer, database, K Nearest Neighbors, Machine learning, random forest, SMOTE.***

## I. INTRODUCTION

Breast cancer is one of the major causes of cancer death among women and it contributes about 2.6% [1] globally which is next to lung cancer. In 2018, it resulted in 2 million new cases and 627,000 deaths [2]. Breast cancer happens when breast cells start to develop abnormally.

These breast cells start to divide much quicker than the other, unaffected ones and form a lump [3]. The count of these cells may further increase, as they may spread to the other unaffected areas. To date, healthcare professionals haven't been able to figure out the exact causes of breast cancer. However, there are a few risk factors, like age, genetic factors, consumption of alcohol, weight, physical activity, etc. that may increase the chances of a person getting breast cancer [4]. With the help of technology, there has been much advancement in the healthcare sector, thus, benefiting mankind. In addition to that, data science is an emerging field that has vast applications in the medical field. Further, machine learning has proved to be a remarkable tool for healthcare professionals to detect the presence of cancer in patients, in its early stages. The accuracies of four machine learning algorithms 1) Support Vector Machine (SVM), 2) Decision Tree Classifier (DTC), 3) Naive Bayes, and 4) k Nearest Neighbors (KNN) using the Wisconsin breast cancer dataset were tested and validated [5]. Among them, the SVM algorithm gave the highest accuracy of 97.13%. In addition to that Random Forest Classifier (RFC) also showed better performance and adaptation as reported by other researchers. Similarly, few other studies compared the performance of three algorithms such as KNN, Logistic Regression (LR), and SVM on the Wisconsin Breast Cancer dataset and it was revealed that the KNN algorithm gave the highest accuracy of 99.28% among them [7],[8], [9],[10]. However in the present study, six machine learning algorithms were used to compare and validate the existing dataset such as 1) LR, 2) DTC, 3) RFC, 4) KNN, 5) SVM, and 6) AdaBoostClasssifier (ABC). All the above machine learning algorithms were compared based on three indicators 1) Precision, 2) Recall, and 3) F1 score values. Two datasets, Wisconsin Breast Cancer and the Breast Cancer Coimbra dataset were used by many researchers [6], [11]. Only a few research have been carried out to compare different datasets and algorithm performance [11]. This study employs a technique known as Synthetic Minority Oversampling Technique (SMOTE) to apply multiple machine learning models based on various parameters and provide a suitable solution for class imbalance. Their results were compared. This provided a better understanding of the impact of utilizing SMOTE when dealing with datasets that had a class imbalance. Machine learning models that are extremely accurate are required in the medical domain since they are directly connected to the health of the patients. In the machine learning industry, a class imbalance is a typical binary classification problem.

Manuscript received on January 09, 2022.
Revised Manuscript received on February 14, 2022.
Manuscript published on February 28, 2022.

* Correspondence Author
**Shivani Motarwar,** School of Electronics Engineering, Vellore Institute of Technology, Chennai (Tamil Nadu), India
**Dixshant Kumar Jha*,** School of Electrical Engineering, Vellore Institute of Technology, Chennai (Tamil Nadu), India E-mail: harshjha5122001@gmail.com

Class imbalance occurs when the dataset contains an unequal number of instances of each of the two classes. The main objectives of this study are to (1) build a machine learning model that correctly detects the presence of breast cancer in a patient, with minimum possible errors and (2) resolve this problem with a technique called SMOTE and subsequently compare the performances of six Machine Learning algorithms, comprising of 1) LR, 2) DTC, 3) RFC, 4) KNN, 5) SVM, and 6) ABC.

## II. MATERIALS AND METHODS

The dataset used in the present study is the Wisconsin Breast Cancer [12], which contains 32 features, as shown in Table 1 (A). The first step which was carried out in this study was to identify and remove those features which were of no use as they would increase the training time of the model. The first feature of the dataset, the "id" column was identified to be an unimportant attribute and hence, it was dropped from the dataset. The second attribute of the dataset, i.e., the 'diagnosis' column was observed to have labels for the malignant ("M") and benign ("B") cases. To convert these labels to numerical data i.e., to make them machine-readable, the method of Label Encoding was used. Using Label Encoder class from the *scikit- learn* library, 1 was assigned to the "M" cases and 0 to the "B" cases. As a next step, all the other attributes were normalized to bring them on the same scale.

**Table 1 (A): Description of nominal variables in the breast cancer dataset**

| Table 1 (A) Description of nominal variables in the breast cancer dataset | | | |
|---|---|---|---|
| **ID** | **Concavity Mean** | **Smoothness SE** | **Perimeter Worst** |
| Diagnosis | Concave Points Mean | Compactness SE | Area Worst |
| Radius Mean | Symmetry Mean | Concavity SE | Smoothness Worst |
| Texture Mean | Fractal Dimension Mean | Concave Points SE | Compactness Worst |
| Perimeter Mean | Radius SE | Symmetry SE | Concavity Worst |
| Area Mean | Texture SE | Fractal Dimension SE | Concave Points Worst |
| Smoothness Mean | Perimeter SE | Radius Worst | Symmetry Worst |
| Compactness Mean | Area SE | Texture Worst | Fractal Dimension Worst |

The dataset has 458 "B" instances (65.5%) and 241 "M" (34.5%) cases. As the number of "B" cases is roughly twice that of the number of "M", the problem of class imbalance arises; this may ultimately reduce the accuracy (Fig.1). Hence, in this study, the technique of SMOTE is implemented, which eliminates the problem of class imbalance.

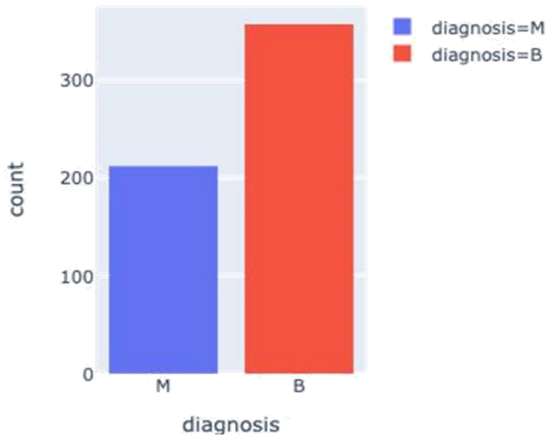

**Fig. 1 Count of "M" and "B" cases in the present study**

Further, a heat map using the Seaborn library was plotted to determine those features in the dataset which were highly correlated with the output variable, i.e., "diagnosis" as shown in Fig. 2. This was done to increase modeling accuracy and decrease the problem of overfitting. It was observed that eight features were highly correlated with the "diagnosis" target variable:1) Concave Points Mean 2) Radius Worst 3) Perimeter Worst 4) Radius Mean 5) Perimeter Mean 6) Area Mean 7) Concavity Mean and, 8) Area Worst. Following this, the dataset was split into 70% training and 30% testing data using a train, test, and split function from the scikit-learn library [13].

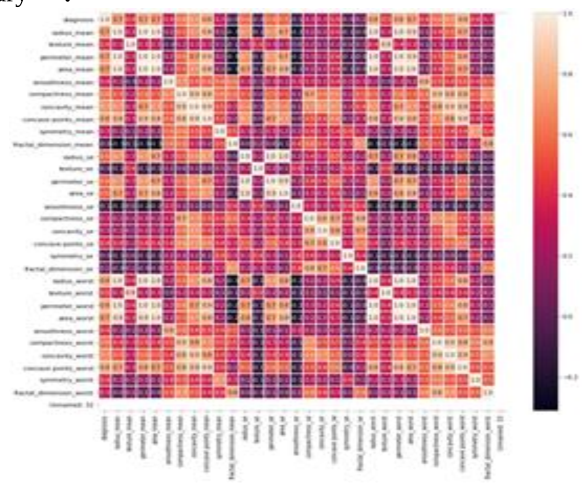

**Fig. 2 Correlation between the features of the Wisconsin Breast Cancer dataset**

Then we chose six different Classification Machine Learning Algorithms, namely 1) LR, 2) DTC, 3) RFC, 4) KNN, 5) SVM and 6) ABC. These six models were trained without SMOTE initially, then with SMOTE, and their performance was assessed using several performance indicators in both scenarios. SMOTE does not inherently improve the accuracy of machine learning algorithms, but it does assist in overcoming Class Imbalance, resulting in more reliable and accurate outcomes.

## III. MACHINE LEARNING ALGORITHMS

All the Machine Learning algorithms which have been implemented in this study are listed as follows.

### A. Logistic Regression (LR)

The LR model is one of the most elementary and powerful machines learning techniques and was devised by Joseph Berkson [14]. It is a Parametric Regression, which means that it uses linear equations for predictions. The LR model gives a categorical outcome for the input independent variables and is further classified into three categories as Binary Logistic Regression, Multinomial Logistic Regression, and Ordinal Logistic Regression. The Logistic function or sigmoid function is an S-shaped curve with the equation:

$$f(x) = \frac{L}{1 + e^{-k}(x - x_0)}$$

### B. Decision Tree Classifier (DTC)

The DTC is one of the most basic and widely used supervised machine learning algorithms, and it differs from other supervised machine learning algorithms in that it can be used to solve both classification and regression problems. Based on the training data, this algorithm determines the value of the target or dependent variable. The DTC algorithm begins with the root node. The values of the root and record attributes are then compared, and the appropriate branch is followed. The same process is repeated for the next node until the final output node i.e., the Leaf Node of the tree is reached. In the case of the DTC, it is important to choose the best attributes for the nodes. For this, a technique called Attribute Selection Measure is used, which makes use of Information Gain and Gini Index, which can be calculated as follows:

**Information Gain**=Entropy (S)-(Weighted Avg.)
*Entropy (each feature)

**Gini Index**= 1-$\Sigma_J P_J^2$

### C. Random Forest Classifier (RFC)

The RFC is an extensively used supervised machine learning technique for both classification and regression issues. It's an Ensemble Learning technique, which means it takes a divide and conquers approach to learning. It's made up of several decision trees that are selected using attribute selection indicators. By using the greatest number of votes collected from the decision trees, these individual trees are then examined for predicting the final output. In the case of classification issues, the final output class is determined by the class that obtains the most votes. For regression issues, on the other hand, the average of all the decision tree outputs is taken into account when determining the final result. The RFC accuracy increases as the number of decision trees examined for voting increases.

### D. K- Nearest Neighbors Algorithm (KNN)

The KNN algorithm was originally devised by Evelyn Fix and Joseph Hodges [15]. It is one of the Supervised Machine Learning algorithms used for both, Classification as well as Regression problems. The KNN algorithm stores the actual dataset and when it receives new data, it classifies the data into the respective category, based on the similarities. In this algorithm, k number of neighbors is chosen, and then the Euclidean distances between them are calculated, based on which, the nearest neighbors are obtained. Finally, the new data point is classified in the category having the maximum number of nearest neighbors.

### E. Support Vector Machine (SVM)

The SVM is one of the most widely used machine learning algorithms which is a supervised learning algorithm that can be used for Classification and Regression problems. SVM, is, however, mostly used for Classification problems [16]. In SVM, a hyperplane is used for classifying the data points into two classes. A hyperplane can be defined as a decision boundary in „n‟ dimensions, which acts as a partition between the two classes. A hyperplane having the maximum margin is always preferred. Further, SVM can be categorized into Linear SVM and Non-Linear SVM. In Linear SVM, a straight line can be used for separating the data points into the two classes, whereas, in Non-Linear SVM, a straight line cannot be used for classifying the data points. It is better to use SVM for a small dataset (having not more than a thousand data points), as the efficiency of SVM is low for large datasets.

### F. Ada Boost Classifier (ABC)

The ABC is a type of ensemble boosting method and is being used in Machine Learning. It consists of many individual classifier algorithms, the outputs of which are combined to get the final result. In this way, a strong classifier with high accuracy is obtained by combining the outputs of the constituent weak classifiers. The total number of accurately detected positive observations from all projected positive observations is known as precision. From all the actual positive observations, recall is the total number of properly detected positive observations. The Harmonic Mean of Precision and Recall is the F1-Score. These performance criteria were taken into account while deciding on the appropriate model for the analysis.

## IV. RESULTS AND DISCUSSION

To visualize the correlation between the features of the Wisconsin Breast Cancer dataset, the heatmap was plotted using the Seaborn library as shown in Fig. 2. Those features which showed the maximum correlation with the output variable were taken into consideration. This helped to maximize the accuracy of the model. As observed from the heatmap, such features include 1) Concave Points Mean, 2) Radius Worst, 3) Perimeter Worst, 4) Radius Mean, 5) Perimeter Mean, 6) Area Mean, 7) Concavity Mean, and 8) Area Worst. Further, these features were used for training to predict accurately the model for the study.

### A. Accuracy

After training the models, for comparison of their performance before and after the implementation of SMOTE, various performance metrics like Accuracy, Precision, Recall, and F1-Score have been analyzed. The accuracies of each of the 6 models, both, before and after SMOTE are given in Table 1(B). It can be seen that KNN gave the best accuracy score of 95.321% after the implementation of SMOTE.

The RFC also gave a fairly high accuracy score of 94.152%. However, the DTC gave a comparatively lower accuracy score of 88.888% in comparison to the other models.

**Table 1 (B): Accuracy percentage of the six models tested**

| Table 1 (B) Accuracy percentage of the six models tested | | | | | | |
|---|---|---|---|---|---|---|
| **Accuracy** | **Percentage** | | | | | |
| | **LR** | **DTC** | **RFC** | **KNN** | **SVM** | **ABC** |
| Before SMOTE | 94.15 | 91.23 | 94.15 | 94.74 | 94.74 | 92.98 |
| After SMOTE | 92.4 | 88.89 | 94.15 | 95.32 | 92.98 | 93.57 |

### B. Precision, Recall and F1-Score

The Precision, Recall, and F-1 scores obtained for each of the 6 Machine Learning algorithms used, both, before and after implementing SMOTE are given in Table 1(C). The precision results obtained for KNN, which gave the highest accuracy score, are 0.97(B) and 0.93(M). The KNN shows a recall of 0.95 (B) and 0.96 (M). The F-1 scores for KNN are 0.96 and 0.94 for “B” and “M” cases, respectively. As a result of the study, it was shown that KNN delivers a better model outcome for breast cancer patients, which might aid in the development of treatment management plans through early detection.

**Table 1 (C): Precision, Recall, and F-1 Scores Obtained for The Experimental Dataset**

| Table 1 (C) Precision, Recall, and F-1 scores obtained for the experimental dataset | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ML Algorithm | Before SMOTE | | | | | | After SMOTE | | | | | |
| | Precision | | Recall | | F1-Score | | Precision | | Recall | | F1-Score | |
| | B | M | B | M | B | M | B | M | B | M | B | M |
| LR | 0.94 | 0.95 | 0.97 | 0.9 | 0.95 | 0.92 | 0.97 | 0.86 | 0.9 | 0.96 | 0.94 | 0.91 |
| DTC | 0.99 | 1 | 1 | 0.99 | 1 | 0.99 | 0.94 | 0.82 | 0.88 | 0.91 | 0.91 | 0.87 |
| RFC | 0.98 | 1 | 1 | 0.97 | 0.99 | 0.98 | 0.97 | 0.9 | 0.93 | 0.96 | 0.95 | 0.93 |
| KNN | 0.94 | 0.98 | 0.99 | 0.91 | 0.97 | 0.95 | 0.97 | 0.93 | 0.95 | 0.96 | 0.96 | 0.94 |
| SVM | 0.94 | 0.95 | 0.97 | 0.91 | 0.96 | 0.93 | 0.97 | 0.88 | 0.91 | 0.96 | 0.94 | 0.91 |
| ABC | 0.99 | 1 | 1 | 0.99 | 1 | 0.99 | 0.96 | 0.9 | 0.93 | 0.94 | 0.95 | 0.92 |

[Benign (B) cases and Malignant (M) cases]

## V. CONCLUSION

The application of the SMOTE to deal with the problem of Class Imbalance in the Wisconsin Breast Cancer dataset is presented in this research. The KNN approach, out of the six machine learning algorithms analyzed in this study, has the highest accuracy of 95.321% after SMOTE implementation. Apart from accuracy, performance indicators such as Precision, Recall, and F-1 scores were examined for all six algorithms before and after SMOTE. This analysis aided us in achieving greater overall accuracy. The model may be used in the real world to assist medical professionals in making accurate illness forecasts and recommending suitable treatment approaches to save human lives.

## ACKNOWLEDGEMENT

The authors are grateful to the Dean and Professor & Pro-Vice-Chancellor, of the School of Electronics Engineering and School of Electrical Engineering, VIT Chennai Campus for their constant support and guidance. We also thank our mentor Dr. Dilip Kumar Jha for the constant support, guidance, and critical review of the manuscript.

## REFERENCES

1. Key Statistics for Breast Cancer. Available online: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html
2. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A. 2018. Global cancer statistics: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.
3. Breast cancer. Available online: https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470
4. Lifestyle-related Breast Cancer Risk Factors. Available online: https://www.cancer.org/cancer/breast-cancer/risk-and-prevention/lifestyle-related-breast-cancer-risk-factors.html
5. Hiba, A., Hajar, M., Hassan Al, M., Thomas, N. 2016. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, Procedia Computer Science, Volume 83, Pages 1064-1069, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.04.224.
6. Yixuan, Li., Zixuan, C. 2018. Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction, Applied and Computational Mathematics. Vol. 7, No. 4, pp. 212-216.
7. Islam, Md & Haque, Md & Iqbal, Hasib & Hasan, Md Munirul & Hasan, Mahmudul & Kabir, Muhammad Nomani. 2020. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN Computer Science. 1. 290. 10.1007/s42979-020-00305-w.
8. Madhu, Kumari, Singh, V. 2018. Breast Cancer Prediction system. Procedia Computer Science. 132. 371-376.
9. Ibrahim Obaid, Omar & Mohammed, Mazin & Abd Ghani, Mohd Khanapi & Mostafa, Salama & Al-Dhief, Fahad. 2018. Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer. International Journal of Engineering and Technology. 7. 160-166. 10.14419/ijet.v7i4.36.23737.
10. Chaurasia, V., Pal, S., Tiwari, B. 2018. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology; 119-126. doi:10.1177/1748301818756225.
11. Ajay Kumar, Sushil, R., Tiwari, A. K. 2019. “Comparative Study of Classification Techniques for Breast Cancer Diagnosis,” International Journal of Computer Sciences and Engineering, Vol.7, Issue.1, pp.234-240.
12. “UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
13. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
14. Cramer, J.S. 2002. The Origins of Logistic Regression. Tinbergen Institute Working Paper No. 2002-119/4, Available at SSRN:

https://ssrn.com/abstract=360300 or http://dx.doi.org/10.2139/ssrn.360300

15. Fix, Evelyn; Hodges, Joseph L. 1951. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (PDF) (Report). USAF School of Aviation Medicine, Randolph Field, Texas.
16. Ray, S. 2017. Understanding Support Vector Machine (SVM) algorithm from examples (along with code) https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.

## AUTHORS PROFILE

**Shivani Motarwar,** Undergraduate Student studying Electronics and Computer Engineering, School of Electronics Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu

**Dixshant Kumar Jha,** Undergraduate Student studying Electrical and Electronics Engineering, School of Electrical Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu