INTERNATIONAL JOURNAL OF APPLIED METHODS IN ELECTRONICS AND COMPUTERS

www.ijamec.org

International Open Access

Volume 13
Issue 01
March, 2025

Research Article

Examining the Relationship of Breast Cancer Data With Survival Chance and Comparison of Algorithms on Breast Cancer Prediction

Ali Murat TIRYAKI [a], ***Ahmet Can MERMER*** [a], ***Bora UĞURLU*** [a,*]

[a] *Department of Computer Engineering, Canakkale Onsekiz Mart University, Canakkale, Türkiye*

ARTICLE INFO

Article history:
Received 19 February 2025
Accepted 23 March 2025

Keywords:
Breast Cancer
Machine Learning
XGBoost

ABSTRACT

This article compares the performance of machine learning algorithms on breast cancer data. The aim is to predict the survival status of breast cancer patients and contribute to the development of clinical decision support systems. Using a dataset obtained from the National Cancer Institute, XGBoost, Random Forest, Support Vector Machines (SVM), and Logistic Regression algorithms were compared. Data preprocessing steps were applied, correlation analysis was performed, and it was determined that the XGBoost algorithm showed the best performance with hyperparameter optimization. The metrics obtained after hyperparameter optimization of the XGBoost algorithm show an overall accuracy of 92%. Optimization has resulted in high performance for class 0 (precision 92%, recall 98%), but the recall for class 1 remains at 54%. The article discusses the effect of data imbalance on the results and offers suggestions for future studies.

1. Introduction

Breast cancer is the most common type of cancer among women worldwide and is one of the leading causes of cancer-related deaths in women [1]. For this reason, early diagnosis and development of effective treatment methods are of great importance. Traditional methods such as mammography and biopsy, which are widely used in breast cancer diagnosis, have some limitations, such as the fact that the method used can lead to overdiagnosis and is invasive, causing difficulties in finding ideal detection and treatment methods for the patient and can impose additional burdens on the health system [2,3]. For these reasons, more objective, non-invasive and rapid diagnostic tools are needed [4].

In recent years, artificial intelligence algorithms have offered great potential in the field of health [5]. This potential also promises new methods and hopes in critical issues such as breast cancer diagnosis and prognosis [6]. Machine learning techniques, especially developments in image processing and big data analysis, play an important role in determining the complex relationships and risk factors associated with breast cancer [7]. The main motivation of this project is to evaluate this potential, accelerate breast cancer diagnosis, make treatment processes more effective, and ultimately increase the quality of life and survival rates of patients.

There are various studies in this field in literature. Some of them are given below. In [8], breast cancer prediction was made using different machine learning approaches. Random Forest (RF) showed better performance than other techniques (accuracy 80%, sensitivity 95%, specificity 80%) and Gradient Boosting (AUC=0.59) showed better performance than neural networks.

In [9], it was aimed to develop a machine learning model that combines ultrasound images and clinical data to predict the Ki-67 value, which indicates the rate of tumor cell division in breast cancer patients. In the study, clinical and ultrasound images collected from 228 breast cancer patients receiving neoadjuvant chemotherapy were used. Using the XGBoost algorithm, a prediction model was created by combining the "delta-radiomic" features obtained from ultrasound images with clinical data. The performance of the model was measured by the ability to correctly identify patients with Ki-67 values ≥ 15%. It was found that the developed model showed high accuracy

* Corresponding author. E-mail address: *boraugurlu@comu.edu.tr*
DOI: 10.58190/ijamec.2025.117

rates in both training and test datasets. It was also stated that the model was suitable for clinical use and could help doctors make treatment decisions. In conclusion, in this study, an effective machine learning model was developed to predict the Ki-67 value in MC by combining ultrasound images and clinical data.

In [10], a machine learning approach is presented to predict breast cancer at an early stage by using genomic data (gene expression profiles). It is aimed to make cancer predictions with higher accuracy and at an early stage by using genetic information in addition to traditional clinical data. Various machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes, Decision Tree and K-Nearest Neighbor (KNN) were used in the study. The results obtained show that Decision Tree and SVM algorithms show the best performance in predicting cancer. The study shows that genomic data can be a valuable tool in breast cancer prediction and this approach can provide significant benefits for early diagnosis and treatment planning.

In [11], it was shown that machine learning algorithms can support healthcare professionals by providing accuracy and speed in breast cancer diagnosis. While algorithms such as Logistic Regression provide high accuracy, feature selection has increased the performance of algorithms such as Light Gradient Boosting. In particular, the importance of features such as tumor size, age and metastasis has been emphasized. In [12], machine learning models were developed in Taiwan with a multicenter approach to predict breast cancer survival with clinical data. The best performing model was the "Artificial Neural Network model", and it was determined that factors such as cancer stage, tumor size, age at diagnosis, and surgery affected survival.

In this study, it is aimed to predict the survival status of breast cancer patients using machine learning techniques. For this purpose, an analysis was made by combining data obtained from different data sources such as demographic, clinical, survival data, and hormone receptors [8], and then the results obtained using XGBoost, Random Forest, Vertical Support Vector, Logistic Regression classification algorithms, which have been proven to be effective in literature, were compared to increase the model performance. At the end of the study, it is aimed to obtain successful results focused on the XGB algorithm in breast cancer diagnosis with machine learning and to contribute to the development of clinical decision support systems. This article consists of four sections. In the second section, used datasets and its properties, algorithms such as Logistic Regression, Support Vector Machines, Random Forest and Extreme Gradient Boosting (XGBoost), used libraries can be found. Besides, there are data preprocessing and hyper parameter optimization techniques. The third section is about results and discussion. The last chapter is the conclusion.

2. Material and Methods

2.1. Dataset

The dataset used was obtained from the November 2017 update of the National Cancer Institute Surveillance, Epidemiology, and End Results program, which provides information on population-based cancer statistics. The dataset included female patients with infiltrating duct and lobular carcinoma breast cancer. Patients with unknown tumor size and patients with survival months less than 1 month were excluded from the dataset, resulting in 4024 patients being included in the dataset. The dataset is a mixed structure containing demographic and medical data of patients. The parameters in the data set are "Age", "Race", "Marital Status", "T Stage", "N Stage, "6th Stage", "Differentiante", "Grade", "A Stage", "Tumor Size", "Estrogen Status", "Progesterone Status", "Reginol Node Positive", "Reginol Node Examined", "Survival Months ,Status" [13]. Necessary information about the parameters will be given in the section below and which parameters should be used will be determined in section 2.2 by examining the correlation matrix of our data set. Our target variable is "Status" and the result of our model is "1=Dead" or "0=Alive".

Age: This parameter contains the patient's age at the time of breast cancer diagnosis. Age is a significant demographic factor influencing breast cancer risk and prognosis. Although age alone may not determine the outcome, breast cancer is often more aggressive in younger patients (under 40) and survival rates tend to be lower in those diagnosed at an older age (over 70). Additionally, the choice of treatment methods and hormonal changes related to age can also impact survival [14].

Race: This parameter captures the racial or ethnic background of the patient. Studies have shown that breast cancer incidence and survival rates can vary significantly among different races, likely due to genetic factors, lifestyle differences, and socioeconomic disparities.[15].

T Stage: This parameter indicates the size and extent of the tumor. Studies show that more advanced cancer stages are associated with lower survival rates [16].

N Stage: This parameter provides information on the spread of cancer to the lymph nodes. It is crucial as it indicates the extent of cancer dissemination. Numerous studies in literature utilize this parameter in conjunction with the "T Stage" [16].

Differentiate: This parameter provides information on how similar cancer cells are to normal cells. Cells that are well-differentiated tend to grow more slowly. The literature shows it is commonly used alongside other parameters in breast cancer studies [17,18]. It has similar meanings to the "Grade" parameter. Therefore, it will be re-evaluated based on the results obtained to avoid erroneous outcomes.

Grade (Aggressiveness Class): This parameter includes the grade value determined based on the microscopic features of the tumor. It indicates the aggressiveness of the cells [17,18]. It has a similar meaning to the "Differentiate" parameter. Therefore, to avoid erroneous results, its use will be re-evaluated based on the findings obtained.

Tumor Size: This parameter represents the largest diameter of the tumor. It is one of the most crucial factors for determining the cancer prognosis [19].

A Stage: This parameter indicates whether the cancer has spread from the breast to nearby tissues or lymph nodes, or to distant parts of the body such as the lungs, liver, or bones. These stages are crucial in determining the extent of cancer progression [20].

6th Stage: The 6th staging system for breast cancer categorizes the extent of the disease's spread. As the stages progress from IIA to IIIB, both tumor size and lymph node involvement increase, signifying that the cancer has advanced further [21].

Estrogen Status: This parameter indicates whether estrogen receptors (ER) are present on tumor cells. The estrogen hormone can travel throughout the body and stimulate cancer cell growth. Tumors that are ER positive are more likely to respond well to hormone therapy and generally have a better prognosis [22].

Progesterone Status: This parameter identifies whether progesterone receptors (PR) are present in tumor cells. Like estrogen, progesterone can fuel cancer cell growth. However, research indicates that fewer tumors respond to this hormone compared to estrogen receptors [23].

Regional Node Examined: This parameter measures the number of regional lymph nodes that are examined to detect cancer spread. Screening these lymph nodes helps to accurately stage the cancer and understand its extent. Although it is not a prognostic factor on its own, it gains significance when evaluated alongside lymph node positivity. Essentially, as the number of lymph nodes examined increases, the likelihood of finding a positive lymph node also rises. The more regional nodes scanned, the greater the indication that the cancer has spread and is classified as high-risk [19,24].

Regional Node Positive: This parameter indicates that cancer cells have spread to the regional lymph nodes. It is crucial for understanding the extent of cancer dissemination and its response to treatment. This attribute, along with the T-stage, is vital in assessing the degree of cancer spread [19,24].

Survival Months: This parameter indicates the number of months a patient has survived since the diagnosis. It has a strong correlation with the "Status" variable.

Status: This parameter indicates whether the patient is alive or dead at the end of the study, providing two possible outcomes[8]. It serves as the target variable for the project. Accurate prediction of patient survival requires training and validating the model on this data.

2.2. Algorithms

When dealing with complex datasets and seeking high accuracy in predictive modeling, selecting the right algorithm is crucial. Four notable algorithms stand out for their unique strengths and applicability: Logistic Regression, Support Vector Machines (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost). Each of these methods offers distinct advantages and potential trade-offs that can significantly influence the outcome of your machine learning projects.

Logistic Regression (LR) is a widely used, interpretable, and probabilistic method for data mining and classification. However, it has some limitations with non-linear data, unbalanced datasets, and small samples. In such cases, correction techniques and alternative algorithms may be required [25].

Support Vector Machines (SVM) is a powerful machine learning technique used for classification. SVM tries to find the best separation between classes by mapping the input data to a higher dimensional space to solve complex problems. The advantages of SVM include good generalization ability, finding unique global optimum solutions, and working effectively using a small number of support vectors. The disadvantages are that computational complexity increases in large datasets and the choice of kernel function can significantly affect the results. It is also more complex than some other methods in terms of interpretability [26].

Random Forest is a powerful ensemble learning method that combines multiple decision trees. Trained with random subsets of data and features, these trees provide more accurate and stable predictions with majority voting in classification and averaging in regression. While it offers advantages such as high accuracy, stability, resistance to over-learning, and feature importance determination, it also includes high computational cost and interpretability difficulties. Especially in large datasets and real-time applications, resource usage and optimization are important [27].

Extreme Gradient Boosting (XGBoost) is a high-accuracy and scalable machine learning algorithm based on decision trees. It is versatile and can be applied to various tasks, providing interpretability. However, it carries the risk of overfitting, and tuning its parameters can be challenging. Therefore, it should be used with caution [28].

To compare these four algorithms, it is important to note that both Random Forest and XGBoost can create highly accurate models for both linear and nonlinear relationships. Logistic Regression, on the other hand, is particularly effective for modeling linear relationships. SVM can also handle nonlinear data well using kernels, but its performance may degrade on large datasets [25,26,27,28].

2.3. Libraries

In this study, several libraries in the Python programming language were utilized for machine learning and data science tasks. Pandas was employed for data analysis and preprocessing [29], while Seaborn and Matplotlib were used for data visualization and model result representation [30,31]. Scikit-learn facilitated the division of data into test, train, and validation datasets, and was also used to implement the KVM, Random Forest, and Logistic Regression algorithms, along with displaying metrics [32]. XGBoost was applied for the execution of the XGBoost algorithm [33], and Skopt was utilized for the hyperparameter optimization of XGBoost [34].

2.4. Data Analysis and Preprocessing

During the data analysis phase, as depicted in Figure 1, it was confirmed that there was no missing data.

```
cancer.isna().sum()

Age                       0
Race                      0
Marital Status            0
T Stage                   0
N Stage                   0
6th Stage                 0
differentiate             0
Grade                     0
A Stage                   0
Tumor Size                0
Estrogen Status           0
Progesterone Status       0
Regional Node Examined    0
Reginol Node Positive     0
Survival Months           0
Status                    0
dtype: int64
```

Figure 1. Missing Data Detection Using isna() and sum() Functions

Following the determination, the parameters of the dataset were analyzed, and the data types within these parameters were identified. This analysis was presented in Figure 2.

```
cancer.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4024 entries, 0 to 4023
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Age                     4024 non-null   int64
 1   Race                    4024 non-null   object
 2   Marital Status          4024 non-null   object
 3   T Stage                 4024 non-null   object
 4   N Stage                 4024 non-null   object
 5   6th Stage               4024 non-null   object
 6   differentiate           4024 non-null   object
 7   Grade                   4024 non-null   object
 8   A Stage                 4024 non-null   object
 9   Tumor Size              4024 non-null   int64
 10  Estrogen Status         4024 non-null   object
 11  Progesterone Status     4024 non-null   object
 12  Regional Node Examined  4024 non-null   int64
 13  Reginol Node Positive   4024 non-null   int64
 14  Survival Months         4024 non-null   int64
 15  Status                  4024 non-null   object
dtypes: int64(5), object(11)
memory usage: 503.1+ KB
```

Figure 2. Displaying Data Types in Parameters with the Info () Function

As a result of this determination, the ordinal, that is, the categorical variables that can be ranked (T stage, N stage, 6th Stage, Differentiate, Grade) were provided to be in hierarchical order by applying the mapping process as in Figure 3. After these processes, nominal type categorical variables with more than 2 different values were reshaped so that each value has a separate parameter and a value of 0 or 1 using the one-hot-encoder function in Python, and categorical variables with a value of 2 were reshaped to have a value of 0 or 1 using the label-encoder function.

```
cancer['Grade'].value_counts()

Grade
2                       2351
3                       1111
1                        543
 anaplastic; Grade IV     19
Name: count, dtype: int64
```

```
mapping = {
    "1": 0,
    "2": 1,
    "3": 2,
    " anaplastic; Grade IV": 3
}
cancer['Grade'] = cancer['Grade'].map(mapping)
```

Figure 3. Mapping Process with Value Counts and Map() Functions in Python

The updated correlation matrix is illustrated in Figure 4, highlighting numerous non-linear relationships within the dataset. Notably, the parameter with the strongest correlation to the target variable is “Survival Months,” exhibiting a moderate negative relationship with a value of -0.476. This suggests that as the patient's survival time increases, their likelihood of survival decreases, indicating a crucial connection. Apart from "Survival Months," there are four other parameters with moderate correlations. These are "Reginal Node Examined" (0.347), "6th Stage" (0.257), "N Stage" (0.255), "Estrogen Status" (-0.184), and "Progresterone Status" (-0.177). Relationships outside these values are considered weak. It's important to remember that the correlation matrix only shows linear relationships and might be misleading for non-linear ones. Hence, all data will be used in the initial stage of our model.

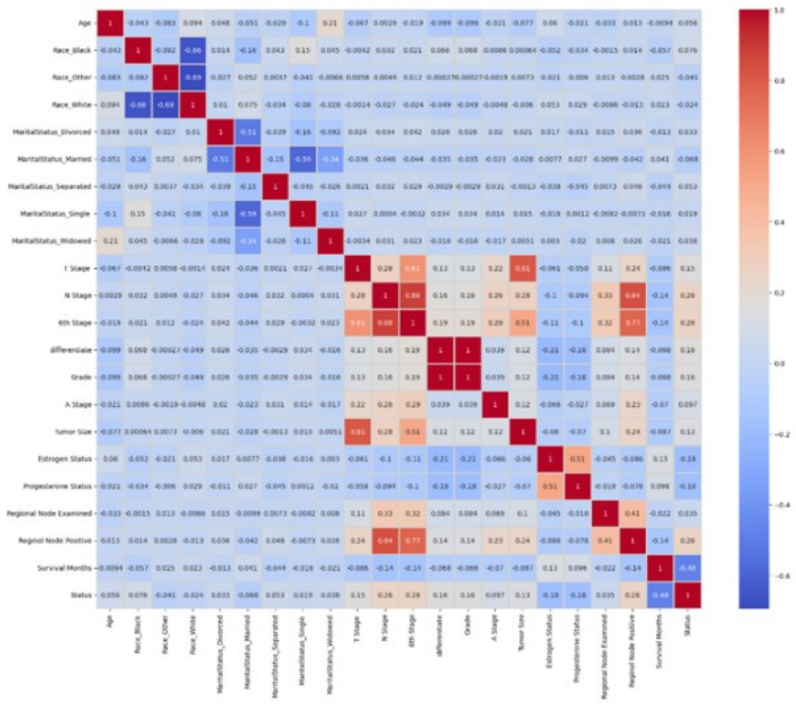


Figure 4. Correlation Matrix

Depending on the context, the model can be optimized and evaluated at a later stage. Following these steps, we divided our dataset into training and testing subsets.

2.5. XGBoost Parameters and Optimization

XGBoost has default parameter values if no customization is made using the xgboost library. For instance, the default is 0.3 for learning_rate and 6 for max_depth [32]. However, to achieve higher accuracy, hyperparameter optimization should be performed. The parameter optimizations made in this direction are given below [34].

The learning rate controls how much the model will "learn" with each iteration. A smaller learning rate may require slower learning and more iterations but generally provides better generalization. A larger learning rate provides faster learning but carries the risk of missing the optimal solution. We will try to find the optimal learning rate by trying different values such as [0.01, 0.1, 0.2]. Tree depth (max depth) determines the maximum depth of each decision tree. Deeper trees can provide better fit to the training data (overfitting), but their generalization ability is reduced. Shallower trees can lead to underfitting. The values [3, 5, 7] are commonly used and give good results. The Param Grid parameter generates all possible combinations of the learning rates and max depths lists. It uses the product function from Python's itertools module, which is a grid search approach that aims to find the best set of parameters by trying different combinations of learning rates and tree depths. The N Estimators parameter, set to 1000, signifies the total number of trees in the XGBoost model. Initially assigning a high value allows for determining the optimal number of trees using the early stopping mechanism. This number will be fine-tuned later with early stopping.

3. RESULTS AND DISCUSSION

Precision, Recall, F1-Score, Accuracy, Macro Avg and Weighed AVG metrics were used to interpret the models, and the definitions of these metrics are given below.

The True Positive (TP) count represents the examples correctly predicted as positive by the model, while the False Positive (FP) count indicates examples incorrectly predicted as positive. Conversely, the True Negative (TN) count denotes examples accurately identified as negative, whereas the False Negative (FN) count reflects examples that were inappropriately predicted as negative despite being positive [35].

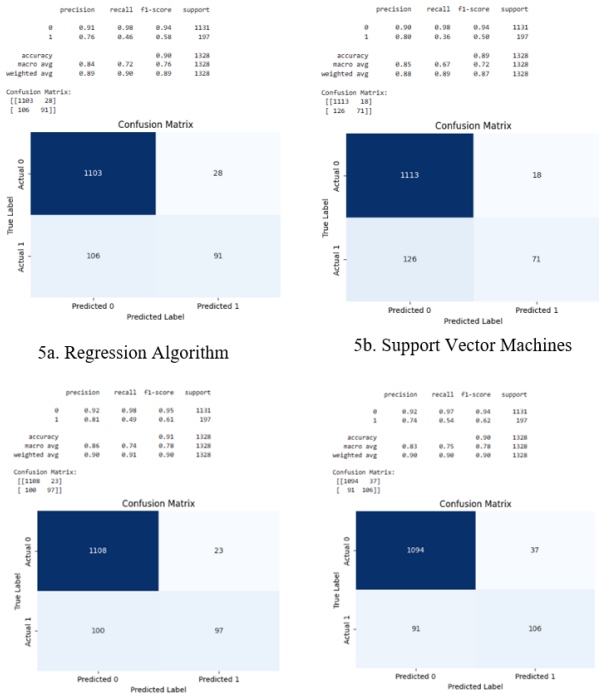
Precision measures the accuracy of the positive predictions made by the model. It answers the question, "How many of the examples predicted as positive are actually positive?" [36]. Recall (Sensitivity) measures the proportion of actual positive examples that were correctly identified by the model. It can also be viewed as the model's ability to detect all positive instances [36]. F1-score: It is the harmonic means of Precision and Recall. It provides a single metric that takes both precision and sensitivity into account. It is especially useful in imbalanced datasets [37]. Accuracy shows how many of the predictions are correct. It is the ratio of all correct predictions to the total number of samples [38]. Precision, Recall and F1 score formulas are given Formula 1, 2 and 3 respectively [39].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$F_1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{TP}{TP + \frac{FN+FP}{2}} \quad (3)$$

Support measures the number of actual instances in each class, which helps understand the weighting of the metrics for each class. Macro Average averages the metrics across all classes, giving equal weight to each class, while Weighted Average calculates the average metrics for each class but weighs them by the "support" value, offering a more meaningful assessment in imbalanced datasets. The metrics and confusion matrix obtained using the Logistic Regression, Support Vector Machines, Random Forest and XGBoost algorithm are given in Figure 5 respectively.



5a. Regression Algorithm

5b. Support Vector Machines

5c. Random Forest

5d. XGboost

Figure 5. Four Different Algorithm Metrics and Confusion Matrices

When we use the Logistic Regression algorithm, we obtain 1103 true negatives, 106 false negatives, 91 true positives and 28 false positives, reaching the metrics in Figure 5a. The metrics and confusion matrix obtained using the Support Vector Machines Algorithm are given in Figure 6. When we use the Support Vector Machines algorithm, we obtain 1113 true negatives, 126 false negatives, 71 true positives and 18 false positives, reaching the metrics in Figure 5b.

The metrics and confusion matrix obtained using the Random Forest Algorithm are given in Figure 7. When we use the Random Forest algorithm, we obtain 1108 true negatives, 100 false negatives, 97 true positives and 23 false positives, reaching the metrics in Figure 5c.The metrics and confusion matrix obtained without hyperparameter optimization in the XGBoost algorithm are given in Figure 5d. In this, when we use the XGBoost algorithm with default values, we obtain 1094 true negative, 91 false negative, 106 true positive and 37 false positive, reaching the metrics.

The metrics and confusion matrix obtained by performing hyperparameter optimization in the XGBoost algorithm are given in Figure 6.

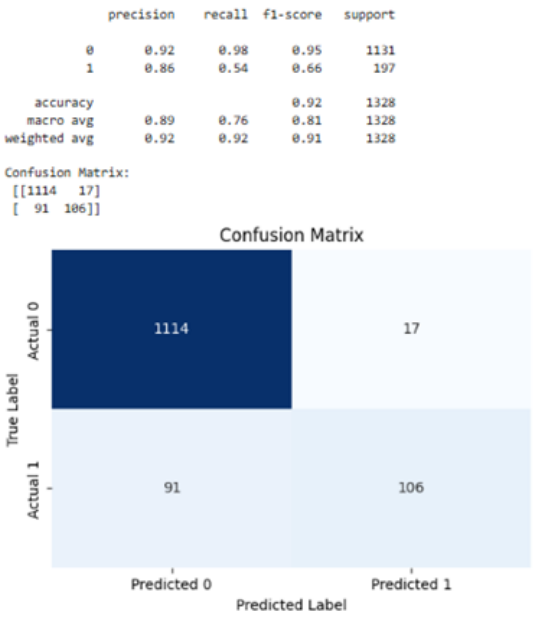


Figure 6. Metrics and Confusion Matrix Obtained When XGboost Algorithm with Hyperparameter Optimization

When we use the XGBoost algorithm by performing hyperparameter optimization, we obtain 1114 true negatives, 91 false negatives, 106 true positives and 17 false positives, reaching the metrics in Figure 6.

At the end of the study, by removing 5 parameters with low correlation values from the data set individually, their effects on class 1 were observed via the XGB algorithm. As seen in Table 1, no increase in values was found when I removed any parameter. This shows us that the main reason for our low results is data imbalance.

Table 1. Parameters with Low Correlation Values

Extracted Data	Precision	Recall	F1 Score
None	0.86	0.54	0.66
Race	0.86	0.51	0.64
Regional Node Examined	0.86	0.52	0.65
Marrial Status	0.85	0.50	0.63
A Stage	0.74	0.50	0.60

The primary goal of the study was to develop a more effective model using the XGBoost algorithm compared to other algorithms. Upon examining and comparing the similarities and differences, it becomes evident that some previous studies [8],[9],[10],[11],[12], mentioned in the introduction section, align with this study in terms of objectives. However, they either achieved their best results using different algorithms, employed different datasets, or aimed to achieve different outcomes.

4. Conclusions

When we examine the confusion matrices, we see that the Vertical Support Vector is the algorithm that can predict the minority class at the lowest rate due to its working principle. (0.36). XGB and Random Forest algorithms give better results with a very small difference in accuracy and F1 score values based on class 1. (accuracy: 0.91-0.92), (F1: 0.61-0.66). The reason for this difference is the hyperparameter optimization we made. The unoptimized XGB result is shown in Figure 8.

When we take the target variable as reference in our data set, a serious imbalance is seen as in Figure 7. This is the main reason why the metric values we obtained for class 1 are low.

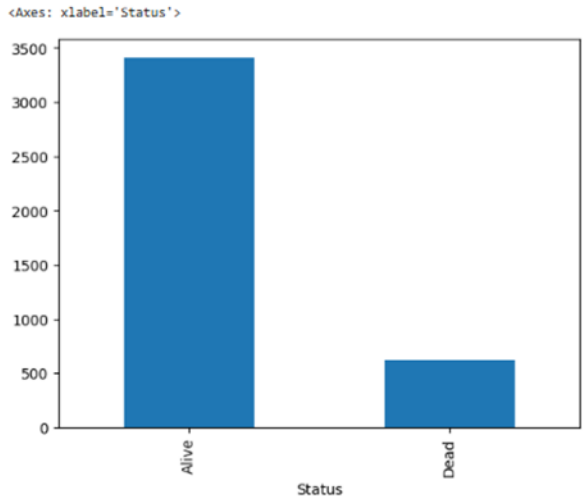


Figure 7. "Status" Data Imbalance

This study highlights key findings through the analysis of breast cancer data and the comparison of various machine learning algorithms. The XGBoost algorithm emerged as a promising method for breast cancer prediction, achieving higher accuracy and F1 scores than other algorithms. However, the results are somewhat limited due to the imbalance in the data set. Future studies with more balanced data are expected to yield more accurate results in this field. This study underscores the potential of machine learning techniques in the diagnosis and prognosis of breast cancer and marks a significant step towards the development of clinical decision support systems.

Declaration of Ethical Standards

The authors confirm that this study adheres to all ethical standards, including proper authorship attribution, accurate citation, appropriate data reporting, and the publication of original research.

Credit Authorship Contribution Statement

Conceptualization of the research was managed by Ali Murat Tiryaki. The data collection study was carried out by Ahmet Can Mermer. The methodology was handled collaboratively by Ali Murat Tiryaki and Ahmet Can Mermer. Evaluation and analysis of the results was handled by Bora Uğurlu. Writing of the original draft was equally contributed by Ahmet Can Mermer and Bora Uğurlu. Review and editing of the manuscript were carried out by Bora Uğurlu. Supervision was undertaken by Ali Murat Tiryaki.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Funding / Acknowledgements

No funding or research grants were received during the preparation of this study.

Data Availability

The dataset used was obtained from the November 2017 update of the National Cancer Institute Surveillance, Epidemiology, and End Results program, which provides information on population-based cancer statistics.

References

[1] World Health Organization. Breast cancer.2024.
[2] Welch HG, Prorok PC, O'Malley AJ, Kramer BS. "Breast-Cancer Tumor Size, Overdiagnosis, and Mammography Screening Effectiveness". New England Journal of Medicine, 375(15), 1438-1447, 2016.
[3] American Cancer Society. "Invasive Breast Cancer.",2021.
[4] Harbeck N, Gnant M. "Breast cancer". The Lancet, 389(10074), 1134-1150, 2017.
[5] Topol E.J. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25, 44-56.2019.
[6] Ehteshami Bejnordi B, Veta M, van Diest PJ, et al. "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer". JAMA, 318(22), 2199-2210, 2017. DOI:10.1001/jama.2017.14585
[7] Zhang B, Shi H, Wang H. "Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach". Journal of Multidisciplinary Healthcare, 16, 1779-1791, 2023. DOI:10.2147/JMDH.S410301
[8] Rabiei R, Ayyoubzadeh SM, Sohrabei S, Esmaeili M, Atashi AR. Prediction of Breast Cancer using Machine Learning Approaches. J Biomed Phys Eng.;12(3):297-308, 2022. doi: 10.31661/jbpe.v0i0.2109-1403.
[9] Lu, Y., Yang, F., Tao, Y., & An, P. An XGBoost Machine Learning Based Model for Predicting Ki-67 Value ≥ 15% in T2NxMo Stage Primary Breast Cancer Receiving Neoadjuvant Chemotherapy Using Clinical Data and Delta-Radiomic Features on Ultrasound Images and Overall Survival Analysis: A 5-Year Postoperative Follow-Up Study. Technology in Cancer Research & Treatment, 23, 1-12,2024. DOI:/10.1177/15330338241265989
[10] Sharma, Saurabh and Shah, Neel and Singh, Rishiraj and Lokare, Reena, Machine Learning Approach for Predicting Breast Cancer Using Genomic Data.Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST) 2020, DOI:10.2139/ssrn.3571724
[11] La Moglia, Alan, and Khaled Mohamad Almustafa. "Breast cancer prediction using machine learning classification algorithms." *Innovation in Biomedical Engineering*, 2024. https://doi.org/10.1016/j.ibmed.2024.100193.
[12] Nguyen, Quynh Thi Nhu, et al. "Machine learning approaches for predicting 5-year breast cancer survival: A multicenter study." *Cancer Science*, cilt 114, sayı 10, 2023, ss. 4063–

4072. DOI:/10.1111/cas.15917
[13] Breast Cancer dataset in Kaggle https://www.kaggle.com/datasets/reihanenamdari/breast-cancer/data. Erişim Tarihi: Aralık 25, 2024
[14] Høst, H., & Lund, E. Age as a prognostic factor in breast cancer. Cancer, 57(11), 2217-2221,1986 DOI:10.1002/1097-0142(19860601)57:11<2217::AID-CNCR2820571124>3.0.CO;2-T
[15] Walker, B., Pollard, E., Howard, S. P., Jones, V. M., O'Connor, K. L., Durbin, E. B., Hull, P. C., Jones, S. R., Adegboyega, A., Wang, X., Owen, W. A. B., Szabunio, M. M., Williams, L. B., & Moore, J. X.. The Role of Race/Ethnicity on the Association between Neighborhood Deprivation and Breast Cancer Outcomes among Kentucky Breast Cancer Patients years 2010-2022. Cancer Epidemiology, Biomarkers & Prevention. Advance online publication.2024. DOI:10.1158/1055-9965.EPI-24-1139
[16] Cleator, S., Makris, A., & Powles, T. .Response to letter “Analysis of breast cancer survival by clinical response to neoadjuvant chemoendocrine therapy” by Bogaerts et al. Annals of Oncology, 17, 352-353,2006.
[17] Dağlar G., Yüksek Y.N, Gözalan A.U, Tütüncü T, Güngör Y, Kama N.A. "The prognostic value of histological grade in the outcome of patients with invasive breast cancer." Turkish Journal of Medical Sciences, cilt 40, sayı 1,, ss. 7–15, 2010.
[18] Henson, D. E., et al. "Relationship among outcome, stage of disease, and histologic grade for 22,616 cases of breast cancer. The basis for a prognostic index." Cancer, cilt 68, sayı 10,, ss. 2142-2149,1991.DOI: 10.1002/1097-0142(19911115)68:10<2142::aid-cncr2820681010>3.0.co;2-d
[19] Narod, S.A. Tumour Size Predicts Long-Term Survival among Women with Lymph Node-Positive Breast Cancer. Curr. Oncol. 19(5), 249-253, 2012. doi:10.3747/co.19.1043
[20] American Cancer Society. "Understanding a Breast Cancer Diagnosis.",2021.
[21] Koh J, Kim MJ. Introduction of a New Staging System of Breast Cancer for Radiologists: An Emphasis on the Prognostic Stage [Erratum]. Korean J Radiol. 2019;20(1):69-82.
[22] Belete, A.M., et al. "The Effect of Estrogen Receptor Status on Survival in Breast Cancer Patients in Ethiopia. Retrospective Cohort Study." *Breast Cancer - Targets and Therapy*, cilt 2022, 2022, ss. 153-161. doi:10.2147/BCTT.S365295.
[23] Li, Z., Wei, H., Li, S., Wu, P., & Mao, X. The Role of Progesterone Receptors in Breast Cancer. Drug Design, Development and Therapy, 16, 305–314, 2022 DOI:/10.2147/DDDT.S336643
[24] Australian Institute of Health and Welfare (AIHW) & National Breast Cancer Centre (NBCC). Breast cancer survival by size and nodal status in Australia. Cancer Series no. 39. Cat. no. CAN 34. Canberra: AIHW; 2007.
[25] Maalouf, M.. Logistic regression in data analysis: An overview. International Journal of Data Analysis Techniques and Strategies, 3(3), 281-299 , 2011.DOI:10.1504/IJDATS.2011.041335
[26] Awad, M., & Khanna, R. Support Vector Machines for Classification. In Efficient Learning Machines (pp. 39-66). Springer,2015. DOI: 10.1007/978-1-4302-5990-9_3
[27] Kulkarni, V. Y., & Sinha, D. K. Random Forest Classifiers: A Survey and Future Research Directions. International Journal of Advanced Computing, 36(1), 1144-1153,2013
[28] Zeravan Arif Ali, Ziyad H. Abduljabbar, Hanan A. Tahir, Amira Bibo Sallow, & Saman M. Almufti. Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review. Academic Journal of Nawroz University (AJNU), 12(2), 320-334, 2023. DOI: /10.25007/ajnu.v12n2a1612
[29] Pandas Development Team. .pandas-dev/pandas: Pandas. Zenodo. DOI:/10.5281/zenodo.3509134
[30] Waskom, M. L. Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021,2021. DOI:/10.21105/joss.03021
[31] Hunter, J. D. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95,2007. DOI:/10.1109/MCSE.2007.55
[32] Scikit-learn developers. Scikit-learn: Machine Learning in Python. scikit-learn.org
[33] Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).
[34] Scikit-optimize contributors. Scikit-optimize: Sequential model-based optimization with a SciPy API. scikit-optimize.github.io
[35] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.
[36] Blair, D. C. (1979). [Review of Information Retrieval, by C.J. Van Rijsbergen]. Journal of the American Society for Information Science, 30(6), 374-375
[37] Chinchor, N. (1992). MUC-4 Evaluation Metrics. In Fourth Message Understanding Conference (MUC-4). Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992 (pp. 22-29).
[38] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In International Joint Conference on Artificial Intelligence (Vol. 14, No. 2, pp. 1137-1145).
[39] Geron A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, O'Reilly Media, Inc.