Breast Cancer Survival Prediction from Imbalanced Dataset with Machine Learning Algorithms

Aditi Kajala[*1], Sandeep Jaiswal[2]
[1]School of Engineering & Technology,
Mody University of Science and Technology,
Lakshmangarh-332311
Sikar-Rajasthan
India
[2]School of Engineering & Technology,
Mody University of Science and Technology,
Lakshmangarh-332311
Sikar-Rajasthan
India
*E-mail: aditikajala@gmail.com

Article Info
Page Number: 167 - 172
Publication Issue:
Vol 71 No. 3 (2022)

Abstract
Breast cancer has surpassed heart disease as the leading cause of mortality among women. Analysis of the duration of the death of an individual after breast surgery can be used to forecast a patient's chances of surviving for a given period. Standard statistical approaches give predictions without elucidating the meaning of the forecast or the relationships between many factors that may affect the patient's survival. With SEER, a publicly available dataset, Shapely Additive Explanation (SHAP) feature of Machine learning algorithms is used to get the representation of predictions. Under-sampling and oversampling approaches are used to balance the imbalanced dataset. Support Vector Machine (SVM) model and Random over sampler outperformed all other machine learning methods and dataset balancing strategies respectively. The SVM model achieved the values of 1 for the precision and 0.9935 for the Area Under Curve (AUC) score.

Article History
Article Received*: 12 January 2022*
Revised*: 25 February 2022*
Accepted*: 20 April 2022*
Publication*: 09 June 2022*

Keywords: - SHAP, balanced dataset, undersampling, oversampling, machine learning models, Decision Tree, Random Forest.
Subject Classification: 68T30, 62P10.

1. Introduction

When cells in any organ begin to grow out of control, cancer develops[1]. Breast cancer survival analysis is estimating the disease risk that may aid patients and clinicians in deciding whether or not to pursue future adjuvant treatment[2,3]. Breast cancer prognosis or survival analysis is crucial in several ways. For starters, it provides patients with information on how their sickness may progress in the future. Second, the prognosis is also helpful for breast cancer treatment as based on the result of prognosis ,a patient may be assigned better treatment [4,5]. Patients with a poor prognosis, maybe considered for intensive treatments as compared to others [6]. Lastly,

forecasting based on the survival analysis can also aids policymakers in comparing death rates among hospitals and institutions by [7].

Using machine learning algorithms it is possible to get the complete detail of the prediction because of their explainability and transparency [8].In this study, we used Shap values to get insights into ML algorithms. The performance of any model also depends on the used dataset. The dataset is said to be imbalanced if the ratio of the proportion of different classes is not the same. The performance of any ML algorithm will also be inappropriate when an imbalanced dataset is applied so it is required to balance the dataset before applying it [9] .

The paper is organized in the sections as follows: Section 2 presents the previous work, section 3 consists the methods and techniques applied in the work section 4 shows the results of the research and section 5 gives the conclusion.

2. Previous Work

The interactions of clinical factors were examined and used to predict the mortality risk combination [10] combining Multifactor Dimensionality Reduction(MDR), Receiver Operating Curve(ROC) dichotomous methods, and logistic regression for the patients of hemodialysis. These algorithms such as decision tree(DT), Random Forest(RF) could identify the important factors like cancer stage, tumor size, the number of total axillary lymph nodes removed that affect mortality[11]. Overall performance of ML algorithms do not show any improvement in the result over conventional statistical and faced the need of methods for data preprocessing, selecting important features [12]. Age at the time of diagnosis of breast cancer is an independent factor for the occurrence of the disease but it had played an important role in the survival and receptiveness to cancer-related therapy. For the prognosis of distant metastasis patterns, it has been observed that patients aged older than 80 years had a lower rate of treatment acceptance [13]. Machine learning algorithms can generate the prediction with explanation and transparency [14]. Due to an imbalanced dataset, machine learning models can be more biased towards the majority class[9].

3. Methods and Techniques

3.1. **Dataset:** The Surveillance, Epidemiology, and End Results (SEER) dataset is used for the presented study. There are 4024 records of breast cancer participants with 16 attributes in the dataset, with 3408 alive cases and 616 death cases [15].

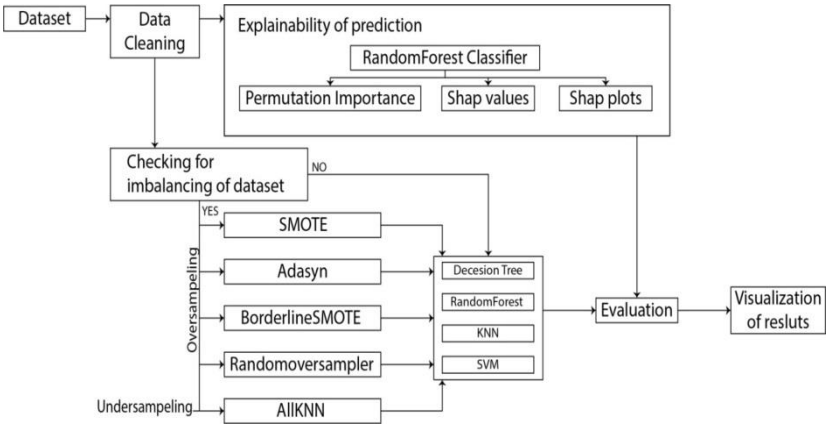

Figure 1: The step-by-step experimental method's framework

3.2. Implementation: The framework in the figure 1 depicts the experiment's step-by-step method. The entire implementation is done in Python using the Kaggle kernel with Graphical Processor Unit (GPU). The GPU specification is as follows: Nvidia Tesla P100, 16 GB GPU RAM, 1.32 GHz clock and performance is 9.3 TFLOPS.

3.2.1. **Data Cleaning:** The label encoding method is used to convert all category attributes into numerical values during data cleaning.

3.2.2. Explainability: The Explainability of machine learning algorithms' prediction helps to understand the impact of various attributes in prediction. It can be explained in graph like structures as Partial Plot or summary plot of Shap values. These graphs describe the importance of each attribute on prediction .The explanation and importance of every attribute can also be described in terms of numerical value by computing the Permutation Importance. In the presented study, Random Forest algorithm is used as a classifier, and the prediction of the model is explained by using Python library and packages to plot Partial Plot, Summary plot of Shap values and to compute Permutation Importance.

3.2.3. Balancing Dataset: By data sampling methods, some samples of either minority(class with less number of samples) or majority(class with more samples) are increased or removed preserving the required related information[9]. When samples of minority class are replicated to balance the dataset, it is called oversampling. In under-sampling, samples from majority class are removed to balance the dataset. In this study, four oversampling techniques[16]: Synthetic Minority Oversampling Technique (SMOTE), Adaptive synthetic Sampling(ADASYN), Random oversample[17], and SMOTE with Borderline and AllKNN under-sampling techniques are applied to balance the dataset.

3.2.4.Breast Cancer Survival Analysis by Machine Learning Algorithms:
For survival analysis, Decision Tree, Random Forest, K-Nearest Neighbors(KNN), and Support Vector Machine algorithms are used to classify a breast cancer patient for five year survival . For each model, the dataset is split into training and test set with function train_test_split keeping the ratio of 70:30 from Python library sklearn.model_selection.

3.2.5. Performance Evaluation of ML algorithms: Every ML algorithm is evaluated without applying any data sampling techniques as well as with the data sampling techniques mentioned in section 3.2.3 with the SEER dataset. The best value for the parameters of applied ML algorithms is computed by grid search. The performance of every algorithm is measured by computing precision value and Area under Curve (AUC) .These values are computed from confusion matrix and ROC (Receiver Operating Characteristics) curve respectively for every algorithm.

4. Results
Table 1 showed the permutation importance of various features with the used dataset .Every feature is assigned some weight that depicted the impact of that feature on prediction on average. Towards the top of the table features are important and could affect the prediction. Features

towards the bottom of the table are less important. The contribution of 15th record from the dataset on the prediction is depicted in partial plot shown by figure 2.It is showed that this patient has more chances of survival due to young age, N Stage value with 0 Grade 1 and Estrogen Status 0. A summary plot in figure 3 showed the effect of attributes with positive and negative shap values depicted by pink and blue colors. Pink showed more impact and blue showed less impact of that feature on the prediction. It is observed that Age, Regional Node examined and Grade are more important feature than Regional Node Positive, N stage and 6th stage. In Table 2 Decision Tree, Random Forest, KNN and SVM with five data sampling techniques are compared for the dataset. From the result presented in the table 2, it is observed that among all data sampling techniques Random over sampler performed better and AllKNN poorly performed with every algorithm. One possible reason of this may be since in AllKNN all samples which were misclassified by KNN are removed from majority class to balance the dataset .SVM among all algorithms showed better performance achieving highest precision value and AUC score.

Table 1: Permutation Importance of features

Weight	Feature
0.0857±0.0090	Survival Months
0.0076 ± 0.0053	Age
0.0022 ± 0.0074	Regional Node Examined
0.0010 ± 0.0036	Grade
0.0006 ± 0.0027	Race
0.0000 ± 0.0028	T Stage
-0.0002 ± 0.0008	A Stage
-0.0012 ± 0.0034	Estrogen Status
-0.0018 ± 0.0026	Progesterone Status
-0.0030 ± 0.0068	Reginol Node Positive
-0.0030 ± 0.0022	Tumor Size
-0.0052 ± 0.0042	6th Stage
-0.0099 ± 0.0028	N Stage

Figure2: Partial Plot

Table 2: Performance of ML algorithms with data sampling techniques

	Model	No Changes	Smote	ADASYN	Bborderline Smote	Random over sampler	AllKNN
Precision	**Decision Tree**	0.44	0.85	0.84	0.86	0.9	0.6
	Random Forest	0.76	0.91	0.86	0.9	0.9	0.93
	KNN	0.47	0.87	0.84	0.88	0.92	0.95
	SVM	**0.8**	**0.97**	**0.97**	**0.99**	**1**	**0**
Area Under Curve(AUC)	**Decision Tree**	0.683	0.862	0.844	0.886	0.9398	0.817
	Random Forest	0.859	0.932	0.912	0.95	0.941	0.906
	KNN	0.665	0.91	0.896	0.918	0.951	0.873
	SVM	**0.712**	**0.972**	**0.973**	**0.978**	**0.9935**	**0.923**

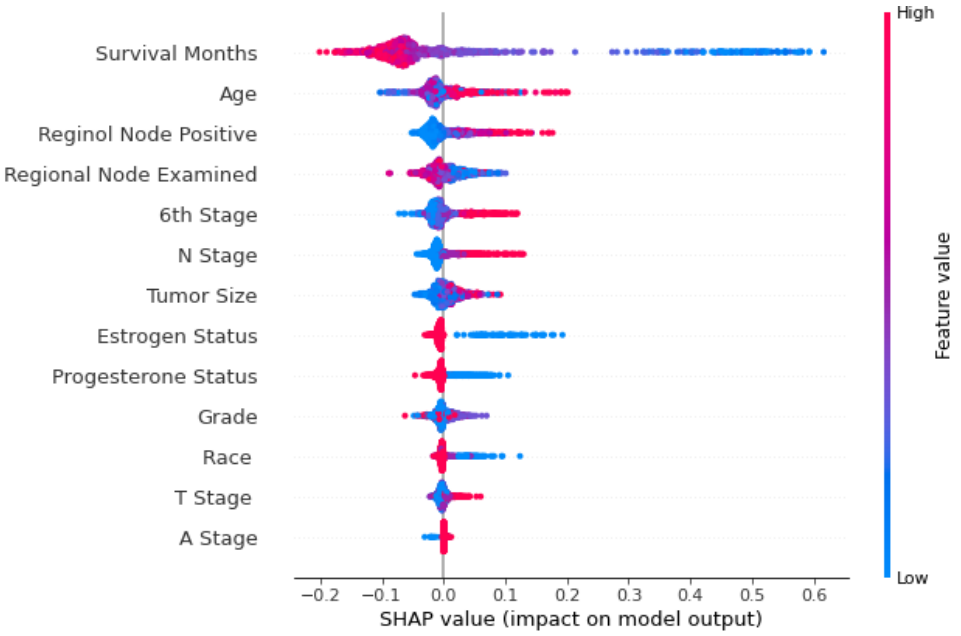

Figure3: Summary plot of SHAP values

5. Conclusion and future Scope

The explainability of machine learning algorithms prediction can be calculated in terms of numeric values and plotted by various graphs. The prediction of Random Forest classifier algorithm is explained by plotting partial plots graphs, summary plots of shap values of attributes for prediction .With the help of these values and graphs, we can interpret the prediction to understand the role of attributes in the survival prediction of a breast cancer patient for five years. In the study SMOTE, ADASYN, Borderline SMOTE, Random oversampler and AllKNN techniques are applied to solve the imbalance of SEER dataset. The performance of four machine learning algorithms namely Decision Tree, Random Forest, KNN, and SVM are compared. The performance of these models was evaluated by computing the Precision and AUC score with the used dataset. The result showed that SVM model with Randomoversampler performed better among all models and achieved the precision and AUC score of 1 and 0.9935 respectively.

References

[1] American Cancer Society, “Breast Cancer What is breast cancer ?,” Am. Cancer Soc. Cancer Facts Fig. Atlanta, Ga Am. Cancer Soc., pp. 1–19, 2017, [Online]. Available: http://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html.

[2] M. T. Phung, S. Tin Tin, and J. M. Elwood, “Prognostic models for breast cancer: A systematic review,” BMC Cancer, vol. 19, no. 1. BioMed Central Ltd., pp. 1–18, Mar. 14, 2019, doi: 10.1186/s12885-019-5442-6.

[3] A. Masarwah et al., “Prognostic contribution of mammographic breast density and HER2 overexpression to the Nottingham Prognostic Index in patients with invasive breast cancer,” BMC Cancer, vol. 16, no. 1, p. 833, Nov. 2016, doi: 10.1186/s12885-016-2892-y.

[4] K. G. M. Moons, P. Royston, Y. Vergouwe, D. E. Grobbee, and D. G. Altman, “Prognosis and prognostic research: What, why, and how?,” BMJ, vol. 338, no. 7706, pp. 1317–1320, 2009, doi: 10.1136/bmj.b375.

[5] D. G. Altman and P. Royston, “What do we mean by validating a prognostic model?,” Stat. Med., vol. 19, no. 4, pp. 453–473, 2000, doi: 10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5.

[6] G. M. Clark, “Do we really need prognostic factors for breast cancer?,” Breast Cancer Res. Treat., vol. 30, no. 2, pp. 117–126, 1994, doi: 10.1007/BF00666054.

[7] A. C. Justice, K. E. Covinsky, and J. A. Berlin, “Assessing the Generalizability of Prognostic Information,” Ann. Intern. Med., vol. 130, no. 6, pp. 515–524, Mar. 1999, doi: 10.7326/0003-4819-130-6-199903160-00016.

[8] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,” Feb. 2017, Accessed: Aug. 06, 2021. [Online]. Available: http://arxiv.org/abs/1702.08608.

[9] M. Oladunjoye, “A Comprehensive Analysis of Handling Imbalanced Dataset,” Int. J. Adv. Trends Comput. Sci. Eng., vol. 10, no. 2, pp. 454–463, 2021, doi: 10.30534/ijatcse/2021/031022021.

[10] C. H. Yang, S. H. Moi, L. Y. Chuang, and J. B. Chen, “Higher-order clinical risk factor interaction analysis for overall mortality in maintenance hemodialysis patients,” Ther. Adv. Chronic Dis., vol. 11, 2020, doi: 10.1177/2040622320949060.

[11] M. Darshini Ganggayah, N. Aishah Taib, Y. Cheng Har, P. Lio, and S. K. Dhillon, “Predicting factors for survival of breast cancer patients using machine learning techniques,” doi: 10.1186/s12911-019-0801-4.

[12] J. Li et al., “Predicting breast cancer 5-year survival using machine learning: A systematic review,” PLoS One, vol. 16, no. 4 April, Apr. 2021, doi: 10.1371/JOURNAL.PONE.0250370.

[13] Y. Han et al., “Metastasis patterns and prognosis in breast cancer patients aged ≥ 80 years: a SEER database analysis,” J. Cancer, vol. 12, no. 21, p. 6445, 2021, doi: 10.7150/JCA.63813.

[14] A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, and G. Geleijnse, “Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival,” Sci. Reports 2021 111, vol. 11, no. 1, pp. 1–13, Mar. 2021, doi: 10.1038/s41598-021-86327-7.

[15] “seer-breast-cancer-data @ ieee-dataport.org.” [Online]. Available: https://ieee-dataport.org/open-access/seer-breast-cancer-data#.

[16] G. Almahadin, · Ahmad Lotfi, M. Mc Carthy, and · Philip Breedon, “Enhanced Parkinson’s Disease Tremor Severity Classification by Combining Signal Processing with Resampling Techniques,” vol. 3, p. 63, 2022, doi: 10.1007/s42979-021-00953-6.

[17] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, “Handling imbalanced datasets : A review,” Science (80-.)., vol. 30, no. 1, pp. 25–36, 2006, [Online]. Available:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.9248& rep=rep1 &type=pdf.