Diagnosis of Breast Cancer on Imbalanced Dataset Using Various Sampling Techniques and Machine Learning Models

Conference Paper · December 2021

DOI: 10.1109/DeSE54285.2021.9719398

Manoj Jayabalan
Liverpool John Moores University
65 PUBLICATIONS **568** CITATIONS

2021 14th International Conference on Developments in eSystems Engineering (DeSE)

Diagnosis of Breast Cancer on Imbalanced Dataset Using Various Sampling Techniques and Machine Learning Models

Ruchita Gupta
Liverpool John Moores University
Liverpool, L3 3AF, UK
ruchitagup@gmail.com

Rupal Bhargava
Upgrad Education Pvt. Ltd,
Mumbai, India
rupal.bhargava@upgrad.com

Manoj Jayabalan
Liverpool John Moores University
Liverpool, L3 3AF, UK
m.jayabalan@ljmu.ac.uk

***Abstract*— Breast Cancer is the second most leading cause of death among women. The early detection of the disease increases the chances of survival of the patient. Therefore, there is always a need for techniques that can accurately predict the presence of cancer. Data Mining is one such powerful technique that can assist clinicians to effectively use the data for timely prediction of the disease. In the medical domain, data is usually imbalanced with unequal distribution of the positive and negative classes. Imbalanced datasets introduce a bias in the model and can thus reduce the accuracy of the minority class predictions. In the case of cancer detection, the mammographic data is highly imbalanced, and predicting the positive (minority) class is of the utmost importance. To achieve this, different models using various class balancing techniques are built and evaluated. The experiments show that the performance of the weighted approach and the undersampling technique is better than oversampling and hybrid techniques. The best performing classifiers are the weighted XGBoost model and Stacking ensemble with the average AUC of 0.78 and 0.76 respectively.**

Keywords*— *Breast Cancer, Class-weighted models, Imbalanced Data, Machine Learning, Ensemble, Voting, Stacking

I. INTRODUCTION

According to the World Health Organization (WHO), cancer is one of the leading causes of death before the age of 70 years [1]. Breast cancer is the most commonly diagnosed cancer and the major cause of death among females. The survivability of a breast cancer patient depends upon the stage at which the cancer is detected. According to the American Cancer Society [2], the earlier the detection, the better are the chances of survival, refer Table I. Hence the clinicians need to diagnose the presence of breast cancer accurately as early as possible by reducing the false-negative cases and thus provide timely treatment to the patients.

Several studies have used classification techniques like Decision Trees, Logistic Regression, Random Forest, Naïve Bayes, Support Vector Machine, and K-Nearest Neighbour for the diagnosis of breast cancer on both mammography and Fine Needle Aspiration Cytology (FNAC) data [3]–[6]. As these medical datasets are highly imbalanced, many researchers have used various under-resampling, over-sampling, and hybrid methods and improved the performance of the classification of the minority class [7]–[12]. These data sampling techniques have several disadvantages. The under-sampling techniques cause a loss of information, whereas the oversampling techniques may lead to an increase in the time complexity of the model and also overfitting in many cases [13]. To handle these issues, there are algorithm-driven data balancing techniques that take care of class imbalance without affecting the distribution of classes.

TABLE I. 5-YEAR BREAST CANCER SURVIVABILITY RATE

Stage	5-year survival rate
Stage-I	98%
Stage - II	92%
Stage - III	75%
Stage - IV	27%

Numerous studies [14]–[16], have used class-weighted ML algorithms on imbalanced medical data and have shown better performance than the conventional data sampling techniques. The ensemble-based technique is another way of improving classification accuracy. These ensemble techniques when combined with data resampling methods, result in better prediction of the minority class [13], [17]–[21].

The purpose of this study is to explore the impact of various class balancing techniques on models built on an imbalanced mammographic Breast Cancer Surveillance Consortium (BCSC) dataset. The various data-driven sampling techniques – Random under-sampling (RUS), Adaptive Synthetic (ADASYN) as an over-sampling technique, and a hybrid technique using RUS and ADASYN, are applied. An algorithm-driven technique – a class weighted approach that handles the class imbalance without the need for data resampling is also explored. Various ensemble models using Voting and Stacking mechanisms are built for improvement in performance in predicting the presence of breast cancer. These models are evaluated on sensitivity, specificity, their average, and AUC. The best performing classifier is compared with the models proposed earlier [9], [22], [23]. Better models and hence better predictability will enable the early initiation of the treatment and thus reduce the complications of treatment and mortality.

II. METHODOLOGY

The various steps required to be performed for the correct diagnosis of breast cancer in women undergoing screening mammography refer to the pre-processing, resampling of the data, data mining techniques, and finally, the evaluation of the models built as depicted in Fig. 1.

978-1-6654-0888-2/21/$31.00 ©2021 IEEE

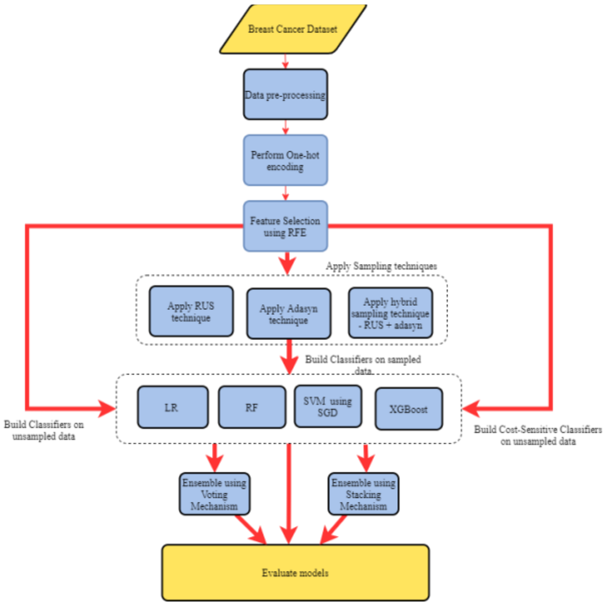
Breast Cancer Dataset
Data pre-processing
Perform One-hot encoding
Feature Selection using RFE
Apply Sampling techniques
Apply RUS technique
Apply Adasyn technique
Apply hybrid sampling technique - RUS + adasyn
Build Classifiers on sampled data
LR
RF
SVM using SGD
XGBoost
Build Classifiers on unsampled data
Build Cost-Sensitive Classifiers on unsampled data
Ensemble using Voting Mechanism
Ensemble using Stacking Mechanism
Evaluate models

Fig. 1. Research Approach

A. *Dataset description and preprocessing*

The dataset includes 1,007,660 screening mammograms (called the "index mammogram") from women included in the Breast Cancer Surveillance Consortium [22]. This second version of the risk estimation dataset that limits observations to one per woman, as opposed to multiple observations was added in August 2012. All women included in the dataset did not have a previous diagnosis of breast cancer and did not have any breast imaging in the nine months preceding the index screening mammogram. However, all women had undergone previous breast mammography in the prior five years (though not in the last nine months). Cancer registry and pathology data were linked to the mammography data and incident breast cancer (invasive or ductal carcinoma in situ) within one year following the index screening mammogram was assessed. To reduce the size of the dataset, the data have been aggregated by the cross-classification of risk factors and outcomes with a count indicating the frequency of each combination. This reduces the dataset to 181,903 records. The dataset consists of 16 attributes including the output variable. The attributes in the dataset are described in Table II. It contains an attribute ‘invasive’ which indicates the diagnosis of invasive breast cancer within one year of the index screening mammogram. There is another attribute ‘cancer’ which indicates the diagnosis of ductal carcinoma in situ breast cancer within one year of the index screening mammogram. Since predicting the presence of any type of cancer is the aim of this research, the attribute ‘invasive’ is dropped and ‘cancer’ is taken as the target variable.

The attribute ‘count’ represents the frequency count of the records having the same combination of covariates in the row. In [9], every record is replicated ‘count’ number of times which had made the dataset even more imbalanced. In this study, the ‘count’ variable is dropped without replicating the records. There are a lot of unknowns present in the dataset which are denoted by 9 in the dataset. Since the model should be able to predict even for those women which have some of the attributes unknown, the records with only one unknown are not deleted. The rows with more than 6 attributes (out of the 12 predictor variables) as unknowns are deleted. Finally, the dataset consists of 173,330 records and 14 attributes. Another attribute ‘training’ is used later for splitting the dataset into training and testing datasets.

TABLE II. Dataset Description

S No	Variable	Short Name
1	Menopausal status	menopaus
2	Age Group	agegrp
3	Breast Density	density
4	Race	race
5	Hispanic	hispanic
6	Body Mass Index	bmi
7	Age at first birth	Agefirst
8	Number of first degree relatives with breast cancer	nrelbc
9	Previous breast procedure	brstproc
10	Result of the last mammogram before the index mammogram	lastmm
11	Surgical menopause	surgmeno
12	Current hormone therapy	hrt
13	Diagnosis of invasive breast cancer within one year of the index screening mammogram	invasive
14	Diagnosis of invasive or ductal carcinoma in situ breast cancer within one year of the index screening mammogram	cancer
15	Training data	training
16	Frequency count of this combination of covariates and outcomes (all variables 1 to 14)	count

B. *Data transformation*

No ordinal relationship exists between the predictor attributes, though they have integer values, Hence it is important to transform them into categorical values using the one-hot encoding method. The number of predictor features after one-hot encoding becomes 40.

C. *Feature selection*

A large number of features lead to high computational costs in training the models. The irrelevant features can also lead to overfitting and reduce the performance of the models. Hence, the Recursive Feature Elimination (RFE) method along with the Backward Elimination method is used and features having a p-value<0.05 are chosen. Finally, 25 features are selected for training the models.

D. *Train-Test splitting*

The dataset is split into training and validation sets based on ‘training’ and then this feature is dropped. The distribution of the positive and negative classes before and after the split is given in Table III.

TABLE III. DISTRIBUTION OF CLASSES

Details	Class = Yes (1)	Class = No (0)	Total
Before splitting	5,864 (3.4%)	167,466 (96.6%)	173,330
Training (65%)	4,293	109,536	113,829
Testing (35%)	1,571	57,930	59,501

E. *Data Sampling techniques*

In this study, different sampling techniques were applied to the training dataset which are described below.

1) Random Under Sampling (RUS): RUS is the simplest and fastest under-sampling technique. It randomly selects samples from the majority class and deletes them from the training set until a balanced distribution is achieved. Other under-sampling techniques like CNN remove only the redundant samples while ENN removes noisy or ambiguous examples. The Neighbourhood Cleaning Rule (NCL) technique combines both the CNN and ENN techniques [24]. The One-Sided Selection (OSS) technique removes the ambiguous samples on the class boundary and redundant samples of the majority class that is far away from the boundary. All these under-sampling techniques, remove only a few samples from the majority class and are not able to bring a balanced distribution. Moreover, they are quite complex and slow for a large dataset. Hence RUS is chosen for the BCSC dataset.

2) *Adaptive Synthetic Sampling (ADASYN):* ADASYN is an oversampling method [25] that generates minority class data synthetically and it is a modification of the Synthetic minority oversampling technique (SMOTE). It generates more samples of those minority class examples which are difficult to learn, that is, it generates more samples in the region where the density of the minority samples is low as compared to the region where the density is large. Thus, it helps in improving the classification in two ways. Firstly, the data distribution improves by reducing the class imbalance and secondly allows the model to be trained on the harder (boundary) samples. Therefore, this technique is explored on the BCSC dataset to study its efficacy.

3) Hybrid Sampling Technique: RUS + ADASYN is a hybrid technique that combines both the over-sampling and under-sampling techniques. It carries out over-sampling using ADASYN and under-sampling using RUS. This hybrid technique improves the balance ratio by removing instances of the majority class and adding instances of the minority class. Hybrid techniques combine the advantages of both under-sampling and over-sampling techniques and are known to produce reliable results. RUS+ADASYN hybrid technique has not been explored on the BCSC dataset. Hence it is evaluated in this study.

The distribution of the classes in the training data after applying the various sampling techniques is given in Table IV.

TABLE IV. DISTRIBUTION OF CLASSES AFTER SAMPLING TECHNIQUES

Sampling technique	Positive Class (cancer=1)	Negative Class (cancer=0)	Total
Unsampled training data	4,293	109,536	113,829
RUS	4,293	4,293	8,586
ADASYN	109,330	109,536	218,866
RUS (sampling_strategy = 0.5) with ADASYN	7,991	8,586	16.577

F. Classifiers

After applying the data sampling techniques, the classifiers are built using the LR, XGBoost, Support Vector Machine with Stochastic Gradient Descent optimization (SVM-SGD), and Random Forest (RF) techniques. Cost-sensitive (weighted) classifiers are also built on unsampled data using these four algorithms. The weighted classifiers take care of class imbalance issues by penalizing the misclassification of the minority class. The hyper-parameters of the classifiers are tuned by doing 5-fold cross-validation on the training set.

1) Logistic Regression: Logistic regression is the most basic and popular supervised binary classification technique. The models built using LR are highly interpretable as they provide the coefficients of the predictors along with the signs that help to know the importance of each feature and whether they affect the target variable in a positive or negative direction. This technique is well suited for linearly separable datasets.

2) Support Vector Machine (SVM): SVM is a powerful classification technique that can be used to build a linear classifier or a non-linear classifier with the help of kernels. SVM is also known as the Maximum Margin classifier as it minimizes the classification error and simultaneously maximizes the margins. Stochastic Gradient Descent (SGD) is one of the optimization techniques which when used with the SVM model not only optimizes the accuracy of the model but also reduces the execution time [26]. SVM has been known to perform better than many other ML techniques in the medical domain [4], [27]–[29]. Hence SVM is chosen as one of the classifiers.

3) Random Forest: Random Forest is another supervised learning technique. It is an ensemble of Decision Trees that are trained using the bagging method. Bagging stands for Bootstrap Aggregation. Each tree can be constructed in parallel since they are constructed independently. RF also reduces the bias in the final model as each tree is constructed on different samples and different features. It also does not suffer from the curse of dimensionality as only a subset of features is considered for building a tree. Many studies [4], [16], [30] have found that RF has a good performance in classifying a breast tumour. Hence, this study explores the RF classifier.

4) Extreme Gradient Boosting (XGBoost): XGBoost is a powerful ensemble of Decision Trees used for classification. In boosting, the weak learners are combined sequentially to boost the overall performance. Each subsequent model assigns a higher weight to the samples that are misclassified in the previous model and hence obtains a strong model that has reduced bias. XGBoost is the most popular algorithm due to its scalability, speed, and capability to handle sparse data [31]. A class-weighted XGBoost handles the class imbalance issues by assigning more weight to the minority class [32]. XGBoost has shown good performances in previous medical-related studies [9], [33]. Thus, this study evaluates the model on the XGBoost with all the data sampling techniques as well as the weighted approach.

5) Ensemble Model using Majority Voting mechanism: Majority Voting is an ensemble technique in which odd number of classifiers are chosen as base classifiers. The classification results of each of these models are used to predict the outcome using majority voting. In this study, the best performing models - weighted XGBoost, weighted LR, and weighted SVM classifiers are chosen to build the ensemble as shown in Fig 2.

6) Ensemble Model using Stacking mechanism: Stacking is another ensemble method that can be used for classification. It consists of two parts – Base Classifiers and Meta-Classifiers. In this work, the base classifiers used are weighted LR, weighted XGBoost, and weighted RF. SVM-SGD is not chosen as it only gives the predicted label and does not give the probability. In one ensemble (passthrough=False), the predicted probability outputs of these three classifiers are combined with the true class to form a new dataset. Another classifier called the Meta-Classifier is then used for predicting the class on the new dataset. In this work, classifier LR is used as the Meta-classifier as LR is known to be a simple yet powerful classifier. In another variation of this ensemble (passthrough=True), the predicted probabilities of the base classifiers along with the true class and also the original data are given as input to the meta-classifier. Both these variations are depicted in Fig 3.

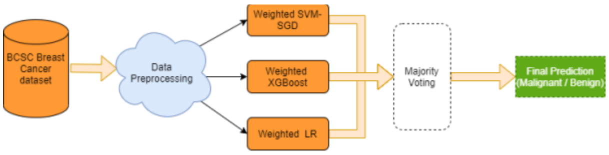

Fig. 2. AN ENSEMBLE USING VOTING MECHANISM

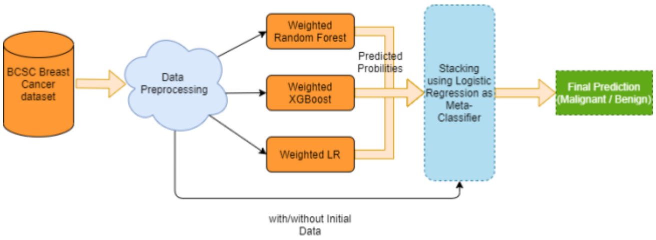
BCSC Breast Cancer dataset
Data Preprocessing
Weighted Random Forest
Weighted XGBoost
Weighted LR
Predicted Probilities
Stacking using Logistic Regression as Meta-Classifier
Final Prediction (Malignant / Benign)
with/without Initial Data

Fig. 3. AN ENSEMBLE USING STACKING MECHANISM

G. *Evaluation Metrics*

The choice of metric is quite important in the evaluation of the model as well as during the tuning of hyper-parameters. An incorrect metric may lead to a sub-optimal model and incorrect conclusions. The dataset used in this study is highly imbalanced. The minority (positive) class is only about 3.4%. If the classifier predicts all samples as a negative class, the accuracy of the model will be more than 99% but the model will not be of any use. So, the accuracy of the model is of no significance in this case. In predicting life-threatening diseases, a False Negative is more disastrous than a False Positive in a preliminary diagnosis. Hence, precision is not important in this study. The area under the ROC curve is generally used to compare various models but a basic weakness with AUC is that it treats both False positives and False negatives equally. It has no consideration for the relative severity between these two types of errors. Hence evaluating a model solely on AUC would be misleading. A high Recall would mean low False Negatives. Therefore, it is an important metric. If the model predicts all samples as a positive class, 100% recall can be achieved, but that model will not be useful as it will have very high False Positives. The purpose of the model is to reduce the number of False Negatives without much increase in False positives. The sensitivity and specificity can be tuned by changing the probability cut-off. On lowering the cut-off, the sensitivity increases while the specificity decreases and vice-versa, but the average remains the same. Therefore, the average is considered for comparing the various models. Therefore, in this study, the various classifiers would be compared based on the sensitivity (recall of positive class), specificity (recall of negative class), an average of sensitivity and specificity, and AUC.

III. EXPERIMENTS AND RESULTS

The results of the various experiments and performance evaluation of various models are discussed here.

A. *Comparison of proposed classifiers*

Various classifiers using LR, XGBoost, Random Forest, and SVM-SGD algorithms are built on the sampled data. Normal and weighted versions of the above algorithms are also used to build classifiers on the unsampled data. The comparison of these classifiers is given in Table VI.

The AUCs of the models on the unsampled data are good but the TPR (sensitivity) is 0 as the model is predicting all instances as negative. Therefore, this proves that evaluating the model alone on AUC is highly misleading. The under-sampling technique RUS gives better results than the other sampling techniques on this dataset. The weighted models that intrinsically handle the class imbalance issue and the ensemble models seem to be the best among the classifiers.

B. *Comparison between proposed classifier and previous studies*

The previous study [22], has built two separate risk prediction models using LR on premenopausal and postmenopausal women and achieved an AUC of 0.631 (pre-menopausal) and 0.624 (post-menopausal). When the same approach was implemented with the version of the dataset used in this study, the metrics obtained are listed in Table V. Reference [23] computed the risk score using k-NN on only four parameters - age, breast density, number of affected first-degree relatives, and prone to breast biopsy and achieved an AUC of 0.637. With this dataset, the metrics obtained are given in Table V. These two studies have not employed any mechanism to handle the class imbalance issues. When the LR/k-NN model is built on the unsampled data, it predicts all instances as the majority (negative) class. So these models cannot be used for the diagnosis of breast cancer. Moreover, the evaluation metric - AUC used in these two studies, does not correctly depict the goodness of the model.

TABLE V. COMPARISON OF PREVIOUS WORKS WITH THIS STUDY

Previous Literatu-re	Predictive Model	Sensiti-vity	Specif-icity	Avera-ge	AUC
[22]	LR (pre-menopausal)	0	1	0.5	0.75
	LR (post-menopausal)	0	1	0.5	0.72
[23]	k-NN	0	1	0.5	0.6
[9]	XGBoost with SMOTE+ENN sampling technique	0.406	0.899	0.653	0.74
This study (2021)	**Weighted XGBoost**	**0.708**	**0.613**	**0.661**	**0.72**

Another study [9], applied various undersampling, oversampling techniques, and hybrid class balancing techniques. It found out that the best Recall (0.83) and F1 score (0.43) on the minority class were obtained when SMOTE + ENN was used for data resampling with the XGBoost classifier. The experiments were conducted on a

TABLE VI. PERFORMANCE EVALUATION METRICS OF VARIOUS PROPOSED CLASSIFIERS

Classifier	**Sensitivity**		**Specificity**		**Average of Sensitivity and Specificity**		**AUC**	
	Train	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***
LR on unsampled data	0	0	1	1	0.5	0.5	0.66	0.66
LR with RUS	0.707	0.719	0.666	0.587	0.687	0.653	0.75	0.71
LR with ADASYN	0.657	0.621	0.612	0.582	0.635	0.602	0.67	0.62
LR with RUS and ADASYN	0.631	0.577	0.628	0.602	0.629	0.589	0.66	0.61
Weighted LR	0.71	0.722	0.659	0.588	0.684	0.655	0.75	0.71
XGBoost on unsampled data	0	0	1	1	0.5	0.5	0.78	0.71
XGBoost with RUS	0.7	0.701	0.704	0.610	0.702	0.656	0.77	0.71
XGBoost with ADASYN	0.846	0.689	0.685	0.615	0.765	0.652	0.85	0.71
XGBoost with RUS and ADASYN	0.722	0.660	0.729	0.652	0.726	0.656	0.81	0.71
Weighted XGBoost	0.713	0.708	0.687	0.613	**0.7**	**0.661**	**0.78**	**0.72**
RF on unsampled data	0.317	0.09	0.995	0.99	0.656	0.54	0.99	0.71
Random Forest with RUS	0.685	0.697	0.673	0.597	0.679	0.647	0.75	0.7
Random Forest with ADASYN	0.74	0.686	0.646	0.594	0.693	0.64	0.79	0.69
Random Forest with RUS and ADASYN	0.679	0.659	0.688	0.629	0.683	0.644	0.76	0.7
Weighted Random Forest	0.662	0.674	0.672	0.614	0.667	0.644	0.74	0.7
SVM-SGD on unsampled data	0	0	1	1	0.5	0.5	0.64	0.62
SVM-SGD with RUS	0.765	0.766	0.594	0.51	0.68	0.638	0.74	0.7
SVM-SGD with ADASYN	0.744	0.727	0.509	0.478	0.626	0.602	0.65	0.6
SVM-SGD with RUS and ADASYN	0.705	0.676	0.519	0.49	0.612	0.583	0.6	0.57
Weighted SVM-SGD	0.716	0.731	0.644	0.576	0.68	0.654	0.74	0.71
Voting Ensemble	0.714	0.724	0.665	0.592	0.689	0.658	-	-
Stacking Ensemble with passthrough=False	0.71	0.714	0.68	0.605	0.695	0.659	0.76	0.72
Stacking Ensemble with passthrough=True	0.713	0.719	0.678	0.605	**0.696**	**0.662**	**0.76**	**0.72**

different BCSC dataset after removing all the unknowns from the data. But as it is important to diagnose the disease even when some of the factors are unknown, this study did not remove all the unknowns. And when the same method (SMOTE+ENN with XGBoost) was applied to the dataset used in this study, the model is highly overfitted. There is a huge gap in the metrics on the training and test dataset. On the training set, the sensitivity is 0.90 and AUC is 0.98 and on the test set, sensitivity is 0.41 and AUC is 0.74.

IV. Conclusion and Future Work

Predicting the risk of breast cancer at the time of screening mammography is quite challenging. An early diagnosis of this disease can reduce the mortality rate to a large extend. Hence, there is an effort to build models that can assist clinical oncologists in forming their opinion. In this research, the BCSC data contains personal information (age, race, menopausal status, etc.), medical history (number of first-degree relatives with breast cancer, the result of the last mammography, etc.), and the breast density (recorded using BI-RADS). The main aim of this study is to improve the classification of the minority class since this dataset is highly imbalanced. After one-hot encoding, the number of features became 40 with some having p-values >0.05. Hence, feature selection was done using the Recursive Feature Elimination (RFE) and Backward Elimination method, and 25 features were selected. Various data sampling methods were applied and classifiers using the machine learning algorithms – LR, SVM with SGD optimization, RF, and XGBoost were built on the sampled data. Classifiers were built on unsampled data using the cost-sensitive version of these algorithms. Ensemble models were also explored. All these models were compared using the metrics - AUC, sensitivity, specificity, and their average.

It was experimentally found out that the undersampling technique performed better than oversampling and hybrid techniques. But the best performance was found by using the cost-sensitive algorithms on the imbalanced data. Out of all the models, the weighted XGBoost model and Stacking ensemble (with passthrough=True) provided the best classification results with the average score (average of sensitivity and specificity) of 0.661 and 0.662 respectively.

The dataset used in this study consists of clinical factors and BIRADS breast density. The chosen model built on these features has been able to produce modest results. For improving the prediction capability of the model, some more information can be captured at the time of screening mammography. For instance, a more detailed family medical history including the occurrence of any type of cancer in the first-degree relatives can be added to the feature set. More studies can be conducted on the ML algorithms that handle class balancing issues intrinsically to improve classification accuracy. Deep Learning techniques can be applied for model building that may improve the results.

Acknowledgment

We are grateful to the Breast Cancer Surveillance Consortium for making the dataset accessible for the research. Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer

Surveillance Consortium (HHSN261201100031C). A list of the BCSC investigators and procedures for requesting BCSC data for research purposes are provided at: http://breastscreening.cancer.gov/.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA. Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018, doi: 10.3322/caac.21492.

[2] C. E. DeSantis *et al.*, “Breast cancer statistics, 2019,” *CA. Cancer J. Clin.*, vol. 69, no. 6, pp. 438–451, 2019, doi: 10.3322/caac.21583.

[3] M. Gupta and B. Gupta, “A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques,” *Proc. 2nd Int. Conf. Comput. Methodol. Commun. ICCMC 2018*, no. Iccmc, pp. 997–1002, 2018, doi: 10.1109/ICCMC.2018.8487537.

[4] D. Bazazeh and R. Shubair, “Comparative study of machine learning algorithms for breast cancer detection and diagnosis,” in *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, Dec. 2016, pp. 1–4, doi: 10.1109/ICEDSA.2016.7818560.

[5] A. Basu, R. Roy, and N. Savitha, “Performance Analysis of Regression and Classification Models in the Prediction of Breast Cancer,” *Indian J. Sci. Technol.*, vol. 11, no. 3, pp. 1–6, Jan. 2018, doi: 10.17485/ijst/2018/v11i3/119179.

[6] V. Kumar, B. K. Mishra, M. Mazzara, D. N. H. Thanh, and A. Verma, “Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications BT - Advances in Data Science and Management,” 2020, pp. 435–442.

[7] S. M. Rostami and M. Ahmadzadeh, “Extracting Predictor Variables to Construct Breast Cancer Survivability Model with Class Imbalance Problem,” *J. AI Data Min.*, vol. 6, no. 2, pp. 263–276, 2018, doi: 10.22044/JADM.2017.5061.1609.

[8] S. Fotouhi, S. Asadi, and M. W. Kattan, “A comprehensive data level analysis for cancer diagnosis on imbalanced data,” *J. Biomed. Inform.*, vol. 90, no. January, p. 103089, 2019, doi: 10.1016/j.jbi.2018.12.003.

[9] M. F. Kabir and S. Ludwig, “Classification of Breast Cancer Risk Factors Using Several Resampling Approaches,” in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2018, pp. 1243–1248, doi: 10.1109/ICMLA.2018.00202.

[10] M. Naseriparsa, A. Al-Shammari, M. Sheng, Y. Zhang, and R. Zhou, “RSMOTE: improving classification performance over imbalanced medical datasets,” *Heal. Inf. Sci. Syst.*, vol. 8, no. 1, pp. 1–13, 2020, doi: 10.1007/s13755-020-00112-w.

[11] Z. Rustam, D. A. Utami, R. Hidayat, J. Pandelaki, and W. A. Nugroho, “Hybrid preprocessing method for support vector machine for classification of imbalanced cerebral infarction datasets,” *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 9, no. 2, pp. 685–691, 2019, doi: 10.18517/ijaseit.9.2.8615.

[12] Z. Xu, D. Shen, T. Nie, and Y. Kou, “A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data,” *J. Biomed. Inform.*, vol. 107, no. May 2019, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.

[13] T. Chakraborty and A. K. Chakraborty, “Superensemble classifier for improving predictions in imbalanced datasets,” *Commun. Stat. Case Stud. Data Anal. Appl.*, vol. 6, no. 2, pp. 123–141, 2020, doi: 10.1080/23737484.2020.1740065.

[14] C. Wang, C. Deng, and S. Wang, “Imbalance-XGBoost: Leveraging Weighted and Focal Losses for Binary Label-Imbalanced Classification with XGBoost,” *Pattern Recognit. Lett.*, 2019, doi: 10.1016/j.patrec.2020.05.035.

[15] A. Iranmehr, H. Masnadi-Shirazi, and N. Vasconcelos, “Cost-sensitive support vector machines,” *Neurocomputing*, vol. 343, pp. 50–64, 2019, doi: 10.1016/j.neucom.2018.11.099.

[16] M. Zhu *et al.*, “Class weights random forest algorithm for processing class imbalanced medical data,” in *IEEE Access*, 2018, vol. 6, pp. 4641–4652, doi: 10.1109/ACCESS.2018.2789428.

[17] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, “MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble,” *Arab. J. Sci. Eng.*, vol. 39, no. 11, pp. 7771–7783, 2014, doi: 10.1007/s13369-014-1315-0.

[18] N. El-Rashidy, S. El-Sappagh, T. Abuhmed, S. Abdelrazek, and H. M. El-Bakry, “Intensive Care Unit Mortality Prediction: An Improved Patient-Specific Stacking Ensemble Model,” *IEEE Access*, vol. 8, pp. 133541–133564, 2020, doi: 10.1109/ACCESS.2020.3010556.

[19] M. Abdar and V. Makarenkov, “CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer,” *Meas. J. Int. Meas. Confed.*, vol. 146, pp. 557–570, 2019, doi: 10.1016/j.measurement.2019.05.022.

[20] M. Abdar *et al.*, “A new nested ensemble technique for automated diagnosis of breast cancer,” *Pattern Recognit. Lett.*, vol. 132, pp. 123–131, 2020, doi: 10.1016/j.patrec.2018.11.004.

[21] A. S. Assiri, S. Nazir, and S. A. Velastin, “Breast Tumor Classification Using an Ensemble Machine Learning Method,” *J. Imaging*, vol. 6, no. 6, p. 39, May 2020, doi: 10.3390/jimaging6060039.

[22] W. E. Barlow *et al.*, “Prospective Breast Cancer Risk Prediction Model for Women Undergoing Screening Mammography,” *JNCI J. Natl. Cancer Inst.*, vol. 98, no. 17, pp. 1204–1214, Sep. 2006, doi: 10.1093/jnci/djj331.

[23] E. Gauthier, L. Brisson, P. Lenka, and S. Ragusa, “Breast cancer risk score: a data mining approach to improve readability,” *Int. Conf. Data Min.*, pp. 15–21, 2011.

[24] J. Laurikkala, “Improving identification of difficult small classes by balancing class distribution,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2101, pp. 63–66, 2001, doi: 10.1007/3-540-48229-6_9.

[25] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Jun. 2008, no. 3, pp. 1322–1328, doi: 10.1109/IJCNN.2008.4633969.

[26] S. Diab, “Optimizing Stochastic Gradient Descent in Text Classification Based on Fine-Tuning Hyper-Parameters Approach. A Case Study on Automatic Classification of Global Terrorist Attacks,” vol. 16, no. 12, pp. 1–6, 2019, [Online]. Available: http://arxiv.org/abs/1902.06542.

[27] S. Jatav and V. Sharma, “An Algorithm for Predictive Data Mining Approach in Medical Diagnosis,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 10, no. 1, pp. 11–20, 2018, doi: 10.5121/ijcsit.2018.10102.

[28] K. Rajendran, M. Jayabalan, V. Thiruchelvam, and V. Sivakumar, “Feasibility study on data mining techniques in diagnosis of breast cancer,” *Int. J. Mach. Learn. Comput.*, vol. 9, no. 3, pp. 328–333, 2019, doi: 10.18178/ijmlc.2019.9.3.806.

[29] M. Tahmooresi, A. Afshar, B. Bashari Rad, K. B. Nowshath, and M. A. Bamiah, “Early detection of breast cancer using machine learning techniques,” *J. Telecommun. Electron. Comput. Eng.*, vol. 10, no. 3–2, pp. 21–27, 2018.

[30] B. Prabadevi, N. Deepa, L. B. Krithika, and V. Vinod, “Analysis of Machine Learning Algorithms on Cancer Dataset,” in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Feb. 2020, pp. 1–10, doi: 10.1109/ic-ETITE47903.2020.36.

[31] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[32] R. Sharma, N. K. Valivati, G. K. Sharma, and M. Pattanaik, “A New Hardware Trojan Detection Technique using Class Weighted XGBoost Classifier,” in *2020 24th International Symposium on VLSI Design and Test (VDAT)*, Jul. 2020, pp. 1–6, doi: 10.1109/VDAT50263.2020.9190603.

[33] A. Tahmassebi *et al.*, “Impact of Machine Learning With Multiparametric Magnetic Resonance Imaging of the Breast for Early Prediction of Response to Neoadjuvant Chemotherapy and Survival Outcomes in Breast Cancer Patients,” *Invest. Radiol.*, vol. 54, no. 2, pp. 110–117, Feb. 2019, doi: 10.1097/RLI.0000000000000518.