International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

ImbTree: Minority Class Sensitive Weighted Decision Tree for Classification of Unbalanced Data

Pratikkumar A Barot*[1], Harikrishna B Jethva[2]

Submitted: 05/06/2021 Accepted : 08/10/2021

Abstract: A reliable and precise tool for medical machine learning is in demand. The diagnosis datasets are mostly unbalanced. To propose an accurate prediction tool for medical data we need an accurate machine-learning algorithm for unbalanced data classification. In binary class unbalanced medical dataset, accurate prediction of the minority class is important. Traditional classifiers designed to improve accuracy by giving more weight to the majority class. Existing techniques gives good results by accurately classifying the majority class. Despite the fact that they misclassify the minority cases, the total accuracy value does not reflect this. When the misclassification cost of minority class is high, research should focus on reducing the total misclassification cost. This paper presents a new cost-sensitive classification algorithm that classifies unbalanced data accurately without compromising the accuracy of the minority class. Our proposed minority-sensitive decision tree algorithm employs new splitting criteria called MSplit to ensure accurate prediction of the minority class. The proposed splitting criteria MSplit derived from the exclusive causes of the minority class. For our experiment, we mainly focused on the breast cancer dataset by considering its importance in women's health. Our proposed model shows good results as compared to the recent studies of breast cancer detection. It shows 0.074 misclassification cost that is the least among the other comparison methods. Our model improves the performance for other unbalanced medical datasets as well.

Keywords: *Unbalanced Data Learning, Decision Tree, Cost-sensitive learning, Medical Machine Learning Tool, Breast Cancer Classification.*

This is an open access article under the CC BY-SA 4.0 license.
(https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

In the medical domain, early detection of diseases helps doctors to cure with the least harm. In countries with a scarcity of doctors, an accurate medical diagnosis tool can be helpful.

Despite more than two decades of studies performed to improve unbalanced data classification, further study is still required [1]. Unbalanced data have unequal distribution of data among the different categories. It has more number of negative instances as compared to positive instances. In the breast cancer dataset, most of the patients are belonging to the benign (negative) and few are belonging to the malignant (positive) class [2]. This type of dataset is called an unbalanced dataset.

Machine learning tool has huge potential to be very effective in medical diagnosis [1,3–6]. However, the bottleneck for this is the inefficiency of traditional classification algorithms for unbalanced data. Unbalanced data adversely affects the performance of traditional classifiers. Even though they exhibit an accuracy of around 90%, they are not considered as an optimal choice for unbalanced data [2]. Their 90% of accuracy is due to the accurate classification of the majority class [2]. Misclassification of minority class doesn't make a noticeable impact on overall accuracy because they don't have a noticeable representation in the dataset. Minority instances and especially rare instances are remained unnoticed and they are mostly misclassified by traditional algorithms.

Lots of studies have been performed to handle the classification of unbalanced data [6–11]. Many researchers proposed improved classification algorithms using techniques like feature selection and feature weighting for balanced data [12–16]. Recently many authors proposed a similar type of work for unbalanced data [9,11]. Existing studies of unbalanced data are mainly categorized into two groups; one is data level techniques and another is algorithmic techniques [2]. Data level techniques try to balance the dataset. Data level techniques involve data sampling techniques like over-sampling, under-sampling, and mix-sampling [1,2,7]. Over-sampling replicates the minority samples and under-sampling removes the majority samples. In fields like medical diagnosis where accurate identification of minority cases is important, the data level sampling is not considered as an appropriate solution [2]. An algorithmic technique involves cost-sensitive classification [1,2,9,11]. The cost-sensitive approach is about the reduction of misclassification cost or test cost. In the cancer dataset, the misclassification cost of the minority class is huge as compared to the majority class. Thus, to reduce the overall misclassification cost in cancer diagnosis, it is important to achieve an accurate prediction of the minority class.

Fig. 1 shows the impact of the wrong prediction on total misclassification cost. We considered misclassification costs for the minority (positive class) as 10 and majority (negative class) as 1 for the calculation of total misclassification cost. First, we vary the false positive rate (FPR) from 0 to 0.5 and kept the false-

[1] *Computer Engineering Department, Gujarat Technological University, India, ORCID ID: 0000-0002-6771-2470*
2 Computer Engineering Department, Gujarat Technological University, India, ORCID ID: 0000-0003-2954-117X
** Corresponding Author Email: pratikabarot@gmail.com*

negative rate (FNR) fix at 0, and measure the variations in total misclassification cost. Then we reverse the scenario and again calculate the total misclassification cost. Fig. 1 shows less variation in misclassification cost when FPR varies from 0 to 0.5. Misclassification cost increases sharply (after the 6th iteration) with the increase in FNR. This means misclassification of the positive class has a huge influence on total misclassification cost and thus improvement in minority class prediction can reduce the overall misclassification cost.

The motivation of this paper is to propose cost-sensitive classification technique for unbalanced data of breast cancer to minimize misclassification of costly minority class. In this paper, we proposed a decision tree-based cost-sensitive model called ImbTree for the unbalanced data classification. The traditional decision tree algorithm was initially developed for the balanced data and so it is biased towards the majority class [9]. The ImbTree algorithm is the unbiased version of the decision tree algorithm. It treats each class according to its misclassification cost and tries to reduce the overall misclassification cost by accurately classifying costly minority class.

In this paper, section 2 discusses work related to our study. Section 3 explains our proposed ImbTree algorithm. Section 4 gives a detailed data analysis of the breast cancer dataset and highlights the data difficulty and also discusses method implementation with comparison of results.

2. Related Work

Several studies proposed for accurate medical machine learning tools. However, these studies do not consider the unbalanced nature of medical data. Recently, unbalanced data learning attracts the attention of many researchers. Krawczyk et al. [1], stated that unbalanced data learning is still target of intense research. They called cancer malignancy grading one of the recent real-time applications with data imbalanced present. To propose an accurate tool for this kind of application accurate unbalanced data classification is required.

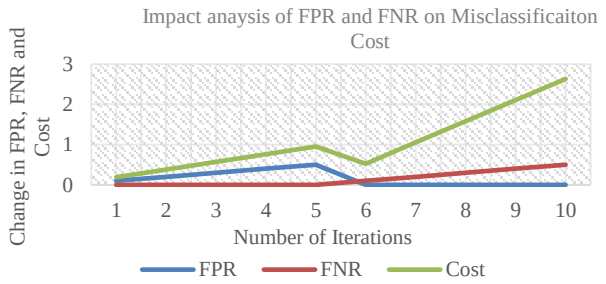


Fig. 1. Behavior of misclassification cost against FPR and FNR

Many researchers proposed data level techniques to handle the data unbalance. However, Due to the drawbacks associated with data level sampling, lots of researches have been performed to eliminate the need for data-level sampling [2,6,7,9,10]. When the dataset has rare sub-concepts with huge misclassification costs, then classification should be handled with extreme care. Data sampling alone does not appropriate for such unbalanced datasets [7].

In 2015 Krawczyk et al. [3] realize the importance of early detection of breast cancer. They performed detailed research on the unbalanced data classification of breast cancer malignancy using a boosting technique. They confirm that the uneven distribution of class is a challenging problem of machine learning [3]. They use under-sampling with boosting technique. However, as per Barot et al. [2], important information like rare sub-concept might be lost during under-sampling.

Coenen et al. [17] proposed an association rule-based classification technique and achieve 90% of accuracy for breast cancer classification. Venkatesan et al. [18] did a performance analysis of the decision tree algorithm for the breast cancer dataset. They had used a different version of decision tree algorithms and found that the J48 shows the best result as compared to the CART, AD TREE, and BF TREE.

In 2016, Shen et al. [19] proposed an evolving support vector machine (SVM) using a fruit fly optimization technique for medical data classification. SVM is tuned using a fruit fly optimization algorithm (FOA). FOA-SVM produces high classification accuracy for the Wisconsin breast cancer dataset, Pima dataset of diabetes, and thyroid dataset.

In 2017, Karar et al. [5] proposed an automated diagnosis of heart sounds using a rule-based classification tree. The diagnosis method uses normalized Lyapunov exponents for rules extraction. Selvi et al. [8] proposed an enhanced grayscale adaptive method for breast cancer prediction. They had proposed an image processing technique for breast cancer prediction. Lee et al. [9] proposed a modified C4.5 algorithm called AUC4.5 as a cost-sensitive approach for unbalanced data. AUC4.5 uses the area under the curve (AUC) as splitting criteria.

Barot et al. [6], proposed an algorithmic approach called minority sensitive decision tree (MiDT) for unbalanced data classification. The MiDT had shown good results as compared to the data level sampling methods.

Devarriya et al. [10] proposed unbalanced data classification for breast cancer using genetic programming. The authors proposed the deletion of the missing values as part of the pre-processing. Deletion of instances with missing values could be costly when such deletion results in a loss of information for the minority class. In the medical domain information belonging to a minority class is important.

Liu et al. [11], proposed information gain directed simulated annealing genetic algorithm wrapper (IGSAGAW) for feature selection and cost-sensitive support vector machine (CSSVM) for breast cancer diagnosis. Authors use information gain (IG) to rank the features and then they use an optimization algorithm to select the best features. They use misclassification cost as evaluation parameters.

Kusuma et al. [20] proposed a back-propagation neural network (BPNN) based study for breast cancer detection. Authors use the Nelder Mead optimization technique to optimize the BPNN. Results show that the decision tree outperforms the proposed NM-BPNN and achieves an accuracy of 96.1%. However, the proposed NM-BPNN outperforms the original BPNN method.

Kaur et al. [21], proposed oversampling technique name General type-2 fuzzy set-based SMOTE (GT2FS-SMOTE) for web spam detection. The GT2FS-SMOTE identifies the minority samples that are close to the center for oversampling. The GT2FS-SMOTE outperform the SMOTE algorithm for unbalanced data classification.

Bej et al. [22], proposed Localized Random Affine Shadowsampling (LoRAS) for oversampling of minority class. The LoRAS generates minority sample without considering majority class distribution.

Kaya et al. [23], proposed differential evolution based oversampling approach named DEBOHID for highly imbalanced datasets. The proposed DEBOHID oversampled the minority class by generating synthetic samples. The DEBOHID generates

minority samples that are similar to existing minority samples and increase the risk of over-fitting. Data over sampling techniques increase the size of dataset that increase learning time and data over-fitting.

3. ImbTree

The proposed ImbTree algorithm is a cost-sensitive classifier for the unbalanced data. The main focus of ImbTree is the minimization of misclassification costs. The objective function defined as in (1).

$$\underset{£p,\ £n}{arg\ min}\ Cpn \times £p + Cnp \times £n \quad (1)$$

Here, Cpn and Cnp are constant and represent the misclassification cost of minority and majority instances respectively. £p is the misclassification factor of the minority class and £n is the misclassification factor of the majority class.

To minimize the objective function we need to minimize the £p and £n. In the medical domain, the misclassification cost of positive instances is generally 10 or more times greater than the misclassification cost of negative instances [9,11]. Thus to minimize the objective function given in (1), focus should be on the reduction of (Cpn × £p). Here Cpn is constant and thus we need to minimize the £p which represents the misclassification factor of minority instances. Fig. 1 shows that the misclassification of minority class has a huge impact on the sharp rise of total misclassification cost. If we can increase the prediction accuracy of the minority class then we can minimize the objective function given in (1).

Fig. 2 shows the model of the proposed minority-sensitive decision tree algorithm. It does not ignore the minority class as the traditional majority-biased decision tree does.

In this approach, first data is pre-processed to remove noise and to fill missing values. After this, ImbTree identifies unique patterns of the target class and extracts them to create a pattern base. Method to identify the responsible causes of targeted minority class was explained by Barot et al. [2]. From these extracted causes feature weights are determined and these weights are used in the calculation of the new splitting method called MSplit. The MSplit is used as the splitting criteria in the ImbTree.

$$MSplit = \begin{cases} fv, & if\ fv \in P \\ fv\ with\ least\ gini\ value, & otherwise \end{cases} \quad (2)$$

$$\forall fv \in\ feature\ value\ pair\ in\ the\ dataset$$

Here, P is a pattern base. The pattern base is build using a causal extraction process as explained by Barot et al. [2] and it is represented as P = {p1, p2, p3,…., pm}. Here p represents the individual pattern and m is the total number of patterns. Each Pi can be defined as D1, D2,.., Dk => Ci. Here, k is the number of dimensions in pattern Pi and Ci is the targeted i^{th} class.

The proposed MSplit method gives special treatment to the important positive class. The MSplit works as the traditional splitting criteria (Gini index) for the balanced datasets. However, for an unbalanced dataset, it uses a pattern base to decide the splitting criteria. When misclassification cost is the same for all the classes then the ImbTree work as traditional decision tree algorithms. But if there is a huge difference in misclassification cost then MSplit introduces special branches in the tree which are responsible to ensure unbiased minority class classification.

4. Experimental Setup

For this study, we used Wisconsin breast cancer, new-thyroid, and Pima Indian diabetes dataset from the UCI machine learning repository. Data distribution has a strong impact on the performance of classifiers. The performance of the classification algorithm deteriorates if the dataset has class overlapping [7].

The breast cancer dataset has 699 instances and 10 attributes. It has 458 instances of class benign and 241 instances of class malignant. Fig. 3 shows a graphical plot with multidimensional scaling (MDS) of the breast cancer dataset. MDS is used to visualize the level of similarity among the multi-dimensional data samples. It transforms multi-dimensional data into two-dimensional data (coordinate 1 and coordinate 2). MDS is used to identify the similarities and dissimilarities between different instances of the dataset.

To ensure that all attributes have the same range of values, we normalized the data using the data normalization technique. Fig. 4 shows MDS on normalized breast cancer data. We use the R language to generate the MDS plot. Both MDS visualizations show that there is a data overlapping. Data and class overlapping are major challenges for the classification [7]. Traditional classification algorithms like SVM and KNN do not perform well if such data overlapping is present in the dataset [7].

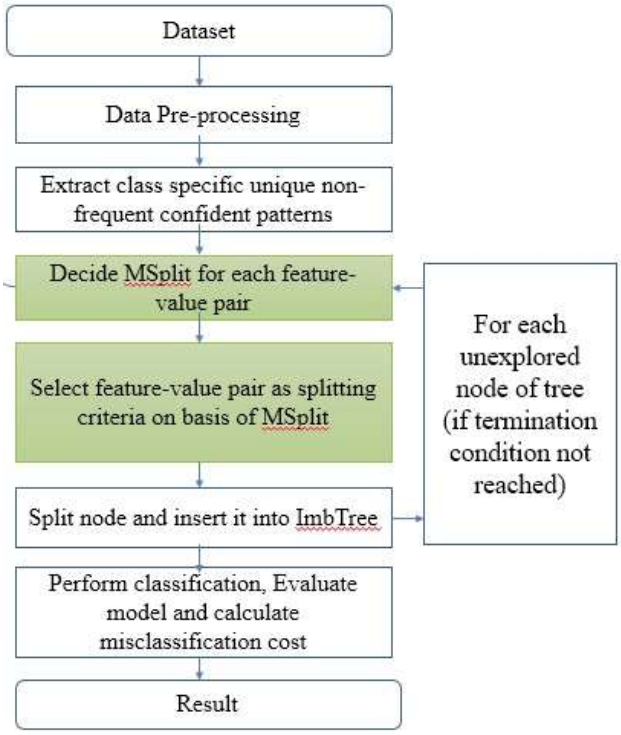


Fig. 2. ImbTree Model

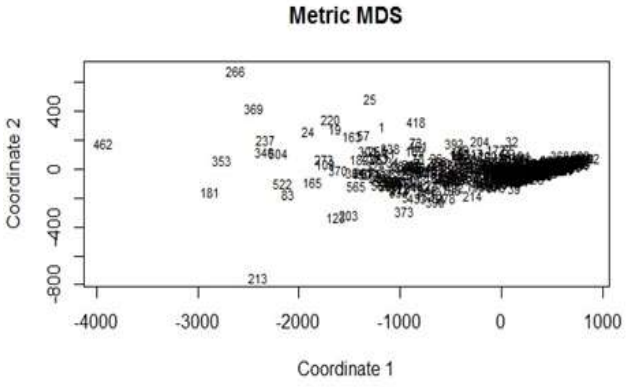


Fig. 3. MDS Visualization of Breast Cancer Dataset.

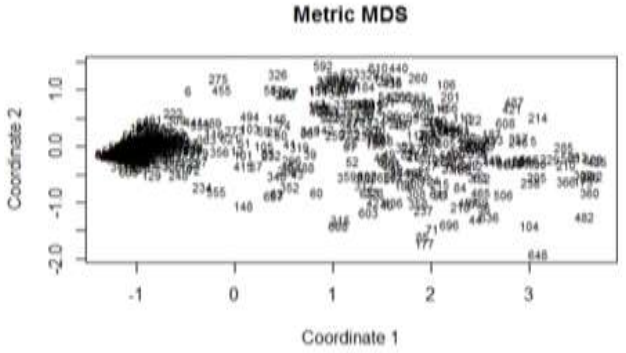


Fig. 4. MDS Visualization of Normalized Breast Cancer Dataset.

Fig. 5 shows the class distribution of the breast cancer dataset. The figure shows class overlapping that makes the traditional classification algorithm of unbalanced data sub-optimal.

Using proximity analysis we did similarity and dissimilarity analysis of the minority Vs minority, minority Vs majority, and majority Vs majority classes. We observed good similarity between some minority and majority class instances as compared to the majority-majority and minority-minority class instances. The minority instances could be misclassified by the traditional majority biased classifiers if they have good similarities with the majority class.

The highlighted portion in Fig. 6 shows that the minority (malignant) instances are surrounded by the majority (benign) instances. Such data distribution characteristics prevent distance-based classifiers from having accurate prediction of minority class [7].

For the implementation of our model, we used python 3.7 on a windows machine with 8GB RAM. We had used three medical datasets for our experiment. Results of our proposed algorithm are compared with the recent studies by Shen et al. [19] and Liu et al. [11].

Our proposed ImbTree algorithm first extracts exclusive patterns of the targeted class. Using extracted unique patterns ImbTree determines splitting criteria called MSplit as per (2). Based on the MSplit value it generates the tree with the special branches known as causal branches. Generation of the minority class-sensitive causal branches is an important part of the ImbTree algorithm to make it unbiased classifiers.

In the medical dataset, the minority class is more important and it has a high misclassification cost as compared to the majority class. Thus the ImbTree only extracts the causes which are responsible for the minority class.

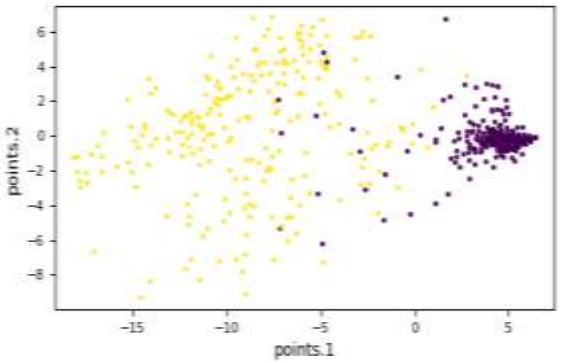


Fig. 5. Class Distribution (Yellow - minority and Purple-Majority) of Brest Cancer Dataset.

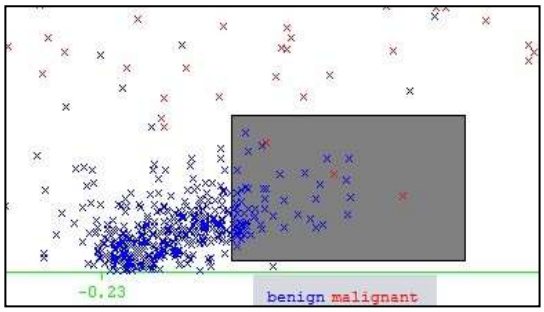


Fig. 6. Proximity Analysis

After the cause extraction phase, we derived interesting patterns from the breast cancer dataset that can play a major role in cancer prediction. Fig. 7 shows extracted major causes and their level of influence in the prediction as malignant.

Fig. 7 shows that if V1 (clump thickness) is more than 8 then there is a sure chance of a diagnosis of the patient as malignant. Among the total malignant cases, more than 50% of cases have clump thickness > 8. From the given dataset, we can say that the pattern related to clump thickness is one of the major and important identifications for the malignant case. Similarly, other important identifications of the malignant patient are uniformity of cell size, marginal adhesion, and normal nucleoli. The ImbTree identifies such unique causes and based on that it build minority sensitive causal branches. The minority sensitive tree alleviates the problem of biasing towards the majority class.

We had used 10-fold cross-validation for training and testing. We had used accuracy and AUC for the performance comparison with FOA-SVM [19]. To test and verify the cost-sensitive capabilities of the ImbTree algorithm we also compared it with the recent cost-sensitive approach called IGSAGAW-CSSVM [11] for breast cancer dataset and performed comprehensive performance analysis in comparison with AUC4.5 [9]. Performance parameters defined as in (3) to (8). TP is true positive, TN is true negative, FP is false positive and FN is false negative value that formulates confusion matrix.

$$\text{Accuracy} = (TP + TN) / (TP+TN+FP+FN) \quad (3)$$

$$\text{Precision} = TP / (TP+FP) \quad (4)$$

$$\text{Recall (Sensitivity or TPR)} = TP / (TP + FN) \quad (5)$$

$$\text{G-mean} = \text{sqrt(Sensitivity*Precision)} \quad (6)$$

$$\text{FPR} = FP / (FP+TN) \quad (7)$$

$$\text{AUC} = (1+TPR - FPR)/2 \quad (8)$$

4.1. Experimental Results and Analysis

The result of ImbTree for the breast cancer dataset is shown in Table-1. The result shows 96.5% accuracy. Particularly for the minority class, it achieved a 98.75% of correct classification rate.
In the cancer dataset, the misclassification cost of the majority class is some medical tests that require some additional expenses while the misclassification cost of the minority class could be a loss of life. Thus by having accurate results for the minority class, our proposed model can reduce the overall misclassification cost. In medical datasets like breast cancer, a reduction in the misclassification cost is considered a huge benefit.
The results of the ImbTree algorithm reveal that out of the total 241 minority instances, 238 instances are correctly classified and three instances are misclassified. Thus, it gives a recall value of 98.75% for the minority class. Recall measure is important to derive misclassification costs for minority class instances. From the total 252 instances predicted as malignant, the true malignant instances are 238 that gives the precision value of 94.8%.

Table 1. Results of ImbTree on Breast Cancer Dataset

Accu.	Recall	Specificity	FPR	FNR	AUC
0.965	0.987	0.953	0.047	0.013	0.970

Table-2 and Table-3 show the performance comparison of the ImbTree with FOA-SVM and PSO-SVM [19]. From the results, we found that the ImbTree based tool is the best performer as compared to the FOA-SVM and PSO-SVM for all three medical datasets.

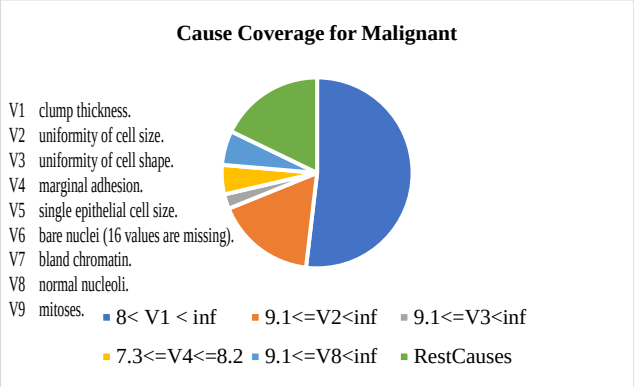


Fig. 7. Analysis of Major Causes of Cancer

Table 2. Performance (Accuracy) Comparison

Dataset	PSO-SVM	FOA-SVM	ImbTree
Breast Cancer	0.962	**0.969**	0.965
New-Thyroid	0.952	0.963	**0.965**
Pima Indian Diabetes	0.765	0.774	**0.856**

Table 3. Performance (AUC) Comparison

Dataset	PSO-SVM	FOA-SVM	ImbTree
Breast Cancer	0.962	0.968	**0.982**
New-Thyroid	-	-	0.978
Pima Indian Diabetes	0.714	0.723	**0.877**

Table 4. Performance Comparison of ImbTree & IGSAGAW-CSSVM

Algorithm	Accuracy	G-mean	AMC
IGSAGAW-CSSVM	0.958	**0.971**	0.121
ImbTree	**0.965**	0.970	**0.074**

Table-4 shows the performance comparison of the ImbTree and the IGSAGAW-CSSVM [11] for the breast cancer dataset. For the performance comparison, we have used accuracy, g-mean, and average misclassification cost (AMC). AMC gives the misclassification cost of the model and it is calculated as per the proposed method in [11].

$$AMC = \frac{\#FP \times MCbm + \#FN \times MCmb}{(\#TP + \#TN + \#FP + \#FN)} \quad (9)$$

#FP gives the number of false positive and #FN gives the number of false negatives. #TP and #TN give the number of true positive and true negative respectively. As per the study of Liu et al. [11], for the breast cancer dataset the cost of misclassification of a minority as a majority (MCmb) is 10 and the cost of misclassification of a majority as a minority (MCbm) is 1.
As shown in Table-4, the ImbTree has better performance as compared to the IGSAGAW-CSSVM. ImbTree shows a more accurate prediction of minority cases. It shows a good prediction rate for both the majority and minority cases and thus it can reduce the total misclassification cost. The ImbTree shows better accuracy and AMC as compared to the IGSAGAW-CSSVM.
Fig. 8 shows the average misclassification cost comparison for PSO-SVM, FOA-SVM, IGSAGAW-CSSVM, and ImbTree. Our proposed ImbTree model achieves an average misclassification cost of 0.074, which is the least among all four algorithms.
Experimental results of NM-BPNN proposed by Kusuma et al. [20] prove that the traditional decision tree outperforms the neural network algorithm for unbalanced data of breast cancer. Table-5 show an accuracy comparison of the traditional decision tree [20], NM-BPNN [20], and ImbTree for the breast cancer dataset.
Our proposed method shows superior accuracy as compared to NM-BPNN and decision tree algorithm. The ImbTree outperforms the original decision tree algorithm and comes out as an optimal alternative for unbalanced data.

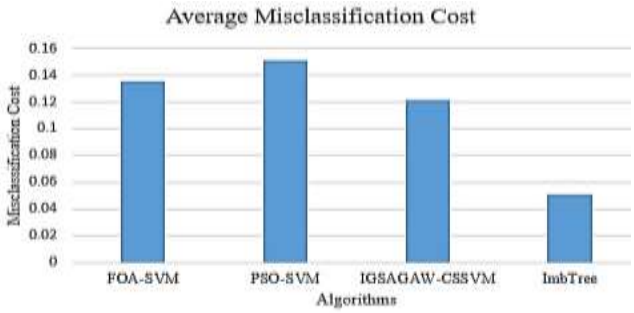


Fig. 8. Misclassification Cost Comparison

Table 5. Accuracy Comparison for Breast Cancer Dataset

Algorithm	Decision Tree [20]	NM-BPNN [20]	ImbTree
Accuracy	0.961	0.898	**0.965**

4.2. Comprehensive Evaluation of ImbTree

In order to obtain a comprehensive evaluation of ImbTree we used a range of datasets as listed in Table 6. We used 13 real life datasets with various class imbalance ratios ranging from 1.15 to 87.83 and results are compared with AUC4.5 [9].
Table 7 shows results of ImbTree and AUC4.5. Performance is evaluated using measures such as Accuracy, Recall, G-mean, Precision and AUC. Except precision measure, other performance measures indicates ImbTree as optimal algorithm as compared to AUC4.5. As ImbTree improves overall accuracy and accuracy of costly minority class, it successfully reduces misclassification cost.

Table 6. Dataset Information

ID	Dataset	#min	#maj	IR
1	Banknote	610	762	1.24
2	Car	65	1663	25.58
3	Diabetes	268	500	1.86
4	Ecoli	20	316	15.8
5	ILPD	167	416	2.49
6	Nursery	330	12630	38.27
7	Wine-quality Red	18	1581	87.83
8	Wine-quality white	175	4723	26.98
9	Yeast	244	1240	5.08
10	Mammographic Mass	445	516	1.15
11	tic-tac-toe	332	626	1.88
12	Credit approval	307	383	1.24
13	Contraceptive Method Choice	333	1140	3.42

5. Conclusion

Lack of in-time diagnosis may result in a loss of life. Machine learning can play a vital role in accurate and in-time disease diagnosis. In developing countries where doctors are less against the total population, the accurate prediction tool could be a blessing. However unbalanced nature of medical data makes the traditional classification less optimal. In the medical dataset, the minority class has a huge misclassification cost. Classification of unbalanced data and especially accurate classification of minority class is always a challenging task. Our proposed ImbTree model is a small effort towards achieving the minority-sensitive machine-learning tool. ImbTree is based on new splitting criteria MSplit, which give equal importance to minority and majority class.

The ImbTree is more accurate for the medical datasets we selected for the experiment. ImbTree shows good results in terms of overall accuracy and average misclassification cost. The result shows that the ImbTree is capable of having a more balanced performance for both the classes and thus successfully reduces the overall misclassification cost. ImbTree shows superior results than other comparison methods. It gives the least total misclassification cost of 0.074 and the best accuracy of 0.965 for the breast cancer dataset.

In the future, studies should be made to reduce misclassification costs for the multi-class unbalanced data.

Acknowledgements

We would like to thank Gujarat Technological University and Government of Gujarat to allow us to do this research study.

References

[1] Krawczyk B. Learning from imbalanced data : open challenges and future directions. Prog Artif Intell 2016;5:221–32. https://doi.org/10.1007/s13748-016-0094-0.

[2] Barot PA, Jethva HB. Statistical Study to Prove Importance of Causal Relationship Extraction in Rare Class Classi fi cation 2018;1. https://doi.org/10.1007/978-3-319-63673-3.

[3] Bartosz Krawczyk, Mikel Galar, Lukasz Jele´n FH. Evolutionary Undersampling Boosting for Imbalanced Classification of Breast Cancer Malignancy. Appl Soft Comput J 2015. https://doi.org/http://dx.doi.org/10.1016/j.asoc.2015.08.060.

[4] Hegde RB, Prasad K, Hebbar H, Singh BMK. Development of a Robust Algorithm for Detection of Nuclei and Classification of White Blood Cells in Peripheral Blood Smear Images. J Med Syst 2018;42. https://doi.org/10.1007/s10916-018-0962-1.

[5] Karar ME, El-Khafif SH, El-Brawany MA. Automated Diagnosis of Heart Sounds Using Rule-Based Classification Tree. J Med Syst 2017;41. https://doi.org/10.1007/s10916-017-0704-9.

[6] Pratikkumar Barot HBJ. Mgini - Minority Class Sensitive Splitting Criterion to Improve Decision Tree for Imbalanced Data of Covid-19. J Inf Syst Eng n.d.

[7] Jerzy Stefanowski. Dealing with Data Difficulty Factors While Learning from Imbalanced Data, Springer; 2016, p. 333.

[8] Selvi C, Suganthi M. A Novel Enhanced Gray Scale Adaptive Method for Prediction of Breast Cancer. J Med Syst 2018;42. https://doi.org/10.1007/s10916-018-1082-7.

[9] Lee JS. AUC4.5: AUC-Based C4.5 Decision Tree Algorithm for Imbalanced Data Classification. IEEE Access 2019;7:106034–42. https://doi.org/10.1109/ACCESS.2019.2931865.

[10] Devarriya D, Gulati C, Mansharamani V, Sakalle A, Bhardwaj A. Unbalanced breast cancer data classification using novel fitness functions in genetic programming. Expert Syst Appl 2020;140. https://doi.org/10.1016/j.eswa.2019.112866.

Table 7. Performance Comparison of ImbTree and AUC4.5

Datasets	ImbTree					AUC4.5				
	Accuracy	**Recall**	**G-mean**	**AUC**	**Precision**	**Accuracy**	**Recall**	**G-mean**	**AUC**	**Precision**
Banknote	**0.988**	**1.000**	**0.989**	**0.990**	**0.974**	0.983	**1.000**	0.984	0.984	0.970
Car	**0.999**	**1.000**	**0.996**	**0.996**	0.844	0.972	**1.000**	0.985	0.986	**0.972**
Diabetes	**0.865**	**0.970**	**0.885**	**0.889**	**0.730**	0.726	0.640	0.703	0.706	**0.737**
Ecoli	0.967	**0.850**	0.910	0.912	0.680	**0.982**	**0.850**	**0.917**	**0.924**	**0.988**
ILPD	**0.797**	0.802	**0.799**	**0.799**	0.612	0.685	**0.856**	0.726	0.735	**0.690**
Nursery	**0.990**	0.873	0.931	0.933	0.754	0.976	**0.936**	**0.956**	**0.956**	**0.976**
Wine quality Red	0.770	0.389	0.548	0.581	0.019	**0.949**	**0.500**	**0.691**	**0.727**	**0.916**
Wine quality white	**0.990**	**0.966**	**0.978**	**0.978**	**0.786**	0.670	0.734	0.700	0.701	0.689
Yeast	**0.917**	**0.791**	**0.863**	**0.866**	0.728	0.792	0.704	0.755	0.756	**0.787**
Mammographic Mass	**0.813**	0.784	0.810	0.811	**0.806**	0.811	**0.918**	**0.812**	**0.818**	0.765
tic-tac-toe	**0.948**	**1.000**	**0.959**	**0.960**	0.869	0.838	0.694	0.796	0.804	**0.890**
Credit approval	**0.835**	0.782	**0.828**	**0.830**	**0.836**	0.820	**0.879**	0.824	0.826	0.795
Contraceptive Method Choice	**0.822**	**0.655**	**0.755**	**0.763**	0.597	0.692	0.604	0.659	0.661	**0.682**

[11] Liu N, Qi ES, Xu M, Gao B, Liu GQ. A novel intelligent classification model for breast cancer diagnosis. Inf Process Manag 2019;56:609–23. https://doi.org/10.1016/j.ipm.2018.10.014.

[12] Kong G, Jiang L, Li C. Beyond accuracy : Learning selective Bayesian classifiers with minimal. Pattern Recognit Lett 2016;80:165–71. https://doi.org/10.1016/j.patrec.2016.06.018.

[13] Dimitrios Gunopulos R. Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection. Proc Work Data Clean Preprocessing (DCAP 2002), IEEE Int Conf Data Min (ICDM 2002) 2002:613–23.

[14] Taheri S, Yearwood J, Mammadov M. Attribute weighted Naive Bayes classifier using a local optimization 2013. https://doi.org/10.1007/s00521-012-1329-z.

[15] Zaidi NA, Carman MJ, Webb GI. Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting 2013;14:1947–88.

[16] Vural MS, Go M. Criminal prediction using Naive Bayes theory ¨ 2016. https://doi.org/10.1007/s00521-016-2205-z.

[17] Coenen F, Leng P. The effect of threshold values on association rule based classification accuracy. Data Knowl Eng 2007;60:345–60. https://doi.org/10.1016/j.datak.2006.02.005.

[18] Venkatesan E. Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification. Indian J Sci Technol 2015;8:1–8. https://doi.org/10.17485/ijst/2015/v8i29/84646.

[19] Shen L, Chen H, Yu Z, Kang W, Zhang B, Li H, et al. Evolving support vector machines using fruit fly optimization for medical data classification. Knowledge-Based Syst 2016;96:61–75. https://doi.org/10.1016/j.knosys.2016.01.002.

[20] Kusuma EJ, Shidik GF, Pramunendar RA. Optimization of Neural Network using Nelder Mead in Breast Cancer Classification. Int J Intell Eng Syst 2020;13:330–7. https://doi.org/10.22266/ijies2020.1231.29.

[21] Kaur P, Gosain A. GT2FS-SMOTE: An Intelligent Oversampling Approach Based Upon General Type-2 Fuzzy Sets to Detect Web Spam. Arab J Sci Eng 2021;46:3033–50. https://doi.org/10.1007/s13369-020-04995-5.

[22] Bej S, Davtyan N, Wolfien M, Nassar M, Wolkenhauer O. LoRAS: an oversampling approach for imbalanced datasets. Mach Learn 2021;110:279–301. https://doi.org/10.1007/s10994-020-05913-4.

[23] Kaya E, Korkmaz S, Sahman MA, Cinar AC. DEBOHID: A differential evolution based oversampling approach for highly imbalanced datasets. Expert Syst Appl 2021;169:114482. https://doi.org/10.1016/j.eswa.2020.114482.