Effects of Data Augmentation by Replicating Instances: Classification Performance by Ensembles of Decision Trees

Kalaiselvi B[1*], Venkateshan [2]

Associate Professor, Department of Electronics and Communication Engineering, Bharath Institute of Higher Education and Research, Chennai, India [1]

Assistant Professor, Department of Electronics and Communication Engineering, Bharath Institute of Higher Education and Research, Chennai, India [2]

Correspondence: kalaigopal1973@gmail.com 1

Abstract: The classification problem of unbalanced instances is rectified using the resampling technique which makes the prediction easier by modifying the training data. We have the machine learning algorithms to combat the imbalanced classification. Among them, resampling is a useful technique which helps to balance the instance based on the classes' majority and minority by under-sampling and oversampling methods. However, in spite of its circulation, the sampling has issues in the efficient evaluation of small-sized data. This study analyses the sampling with ensembles of decision tree classifiers of different split percentages using a diabetes dataset which waver units of imbalance and produce better accuracy. The evaluation measure for each replication percentage for REPTree and Random Tree classifiers is calculated and the same is interpreted in the Discussion.

Key Words: Resampling data, Diabetes, Bagging, REPTree, Random Tree, Accuracy

1. INTRODUCTION

Massive amount of data has been stored due to the technologies today that have brought extensive signs of progress that lead to the production of data and also led to the huge problem in data organization. The data organization brings insight to the analytics. Data mining is the process of finding information from patterns derived from the data. Many data mining tools are used to build the models to detect the correlation between the data and it is very much helpful to predict the behavior of the data. Mainly in the field of health care, data analytics play a vital role because the industry has large amounts of data and needs mining to get precise information that benefits society.

But still issues are there in models predicting accurately to help to diagnose a disease. Nowadays, Diabetes has become a global threat to people because it's increasing rapidly and it may lead to more health issues. Hence the diabetic data is available in large size and analyst gives more attention to discovering the information from that. Machine learning algorithms enable computers to learn efficiently especially while building prediction models. The data is described with attributes and classes. The classes are the possible outcomes of the prediction. To do this, the data has been classified by the class by considering the features of data already known to be appropriate to the class. Before applying any classification algorithm to data, the data has to be pre-processed to improve the performance of a classifier.

In this paper, we focus on the methods of resampling the data which may help in prediction better than other methods. Here we have used certain machine learning algorithms for classification and predicting the diabetic data without applying any feature selection methods. The classification technique is used to produce a more accurate predictive model by examining the training data and creating a pattern, which can be applied to predict new instances. This paper aims to compare the prediction accuracy and the other metric measures between the original data and the resampling data [1]. Sampling methods modify the imbalanced number of instances between the majority and minority decision classes. Removing and adding majority instances may lead to losing the information of the original data. Hence, dealing with minority instances causes' better performance while in this process the existing minority instances are replicated [2, 17].

WEKA is a data mining tool has data pre-processing tools and machine learning algorithms [16]. Our whole experimental approach is supported effectively in a malleable way. It provides filter to apply according to any kind of data and has classifier that classify both supervised and unsupervised learning data.

This paper is prepared as follows: Section 2 contributes an assessment of the related works of our approach. Section 3 provides the details about the data used with the description and methods used for the experiment. Section 4 analyses the results and interprets the comparative outcomes. The clarification of the results is concluded in the last section 5.

2. RELATED WORKS

The paper aims to analyze the resampling technique in the Prediction of Diabetes. There are many study has been done on the domain related to this paper. Some useful works of research are notified here. [4, 5] These papers analyse many diseases by machine learning algorithms proving better prediction. [6] Explains the types of diabetes and their causes and also brings high prediction accuracy in diabetic diagnosis. The effects of diabetes are described also in [7]. [8] Proves Naïve Bayes gives the best performance in predicting diabetic patients by calculating the evaluation measures and comparing them to the same classifier.

Analysis of effective classification algorithm on diabetes prediction is discussed in [9,6]. Lee et al. [10] emphasizes relating decision tree classifiers to the diabetes dataset which enhances the performances after smearing the resample filter over the data. Comparing the original data set, the resampled data has a notable effect on the performance [1]. However, the imbalanced data is not only limited to medical diagnosis makes more methods like sampling are incoming [11]. Bagging is best in selecting a boost sample of data which helps to improve the accuracy. However, this method is not simple as its structure is not interpretable [12]. Bagging significantly improved the prediction accuracy when it was applied with the base classifier REP tree [13]. Rather than decision tree algorithms, ensemble decision algorithms produce high accuracy in predictions [15]. Bagging with a REP tree outperforms other methods when it is proceeded with gain ratio feature selection [14].

However lots of studies have been led, yet there is no precise set of resampling methods that can be well thought-out the best performer. These methods surely require more replicated studies with different classifiers and different datasets.

3. MATERIALS AND METHODS

This section describes the materials and methods involved in this study. Figure 1 clarifies the framework built in this experimental study.

For our study, we have taken diabetes data which is imbalanced [17]. Then using the Weka tool, we First pre-process the data and apply the filter Resample then classify with the decision tree classifiers (REPTree, Random Tree) and calculate the evaluation measures like Accuracy, precision, Recall, F1 score and ROC. Repeat the same with increasing the replication split percentage i.e., (0%, 10%, 20%,…,100%). Then compare the results to detect the best sampling way to balance the data.

3.1 DATA SET

The dataset is derived from Kaggle https://www.kaggle.com/datasets/mathchi/diabetes-data-set. And the National Institute of Diabetes and Digestive and Kidney Diseases is its source. The aim of this data is to diagnose the patient has disease diabetes or not. This data is about 768 female patients of age 21 and above. There are 7 attributes and 1 decision class. All the values of the attributes are numerical values and there missing values also. The details of the data set attributes are described as follows

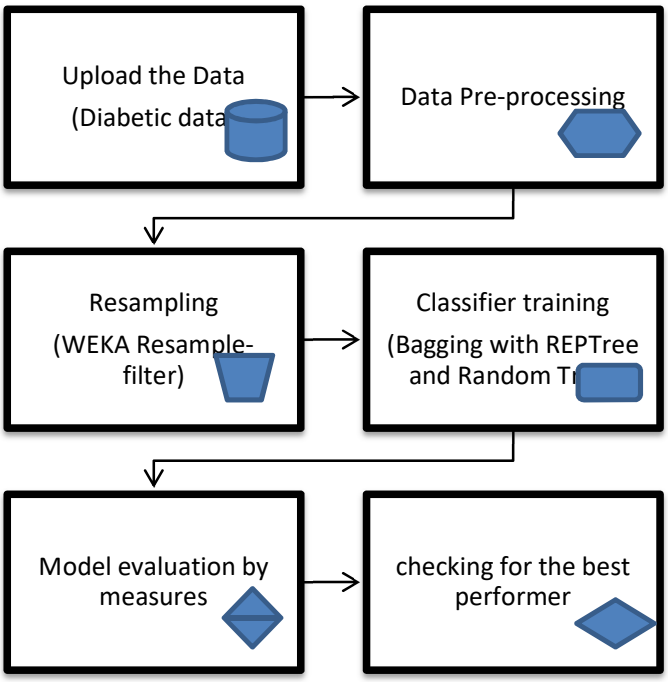

Figure 1: Framework of our methodology

3.2 DATA SET DESCRIPTION

Attribute 1: No. of times female patient gets pregnant

Attribute 2: Glucose tolerance test (level of plasma glucose)

Attribute 3: Blood pressure in mm Hg

Attribute 4: thickness in mm of Triceps skin fold

Attribute 5: insulin in mu U/ml

Attribute 6: BMS (Body mass index in kg/(m)^2)

Attribute 7: Diabetes pedigree function

Attribute 8: Age

Attribute 9: Decision Class (0 or 1)

Class Distribution: Class value 1 is "positive for diabetes" and class value 0 is "negative for diabetes"

3.3 SAMPLING METHOD

When a data has binary classification problem (two classes) and also has less data on class 1 and more on the other class, the prediction may fall on the majority number of negative classes and since the positive classes which are few, become neglected in such a way to get high accuracy of classification. This may lead to loss of the original information of the data. To overcome this issue, two methods have been handled and they are sampling methods: under sampling and over sampling [11].

This study gives interest to the importance of resampling methods to combat unbalanced data to improve classification performance for better prediction. For this sampling method, the data mining tool WEKA is used here. In Weka, under the options of filter select the resample method. In our case, for diabetes data, it is supervised learning and resample filter techniques were applied before the process for classification. The basic decision tree classifier J48 is used to classify the resampled and original data.

3.4 BAGGING

In Machine learning, bagging is a classifier which derives subsets of different combinations of training data from the original data in order to improve the strength of the training data. So it can be called a meta-estimator which reduces the variance between the data by estimating with the decision tree classifier which I act base for it. It randomly constructs each set by different combinations of the original dataset after training them. In this manner, each set will be independent and the repeated sets will be ignored[15]. It works by voting method that helps to reduce the overfitting and at the same time it gives fair variance in reduction. On the whole resampling the data by voting method is bagging. The machine learning tool WEKA does bagging associated with the decision tree classifier REP tree. Among many basic classifiers, the REP tree yields the best results with bagging. After creating the resampled dataset by bagging, the decision tree classifier REPtree does the classification of that data.

3.5 REP TREE

Reduced Error Pruning Tree (REP Tree) is a decision tree classifier which acts rapidly in producing many decision trees using regression tree logic in various iterations [13]. It constructs the tree with the aid of attribute selection information gain ratio and pruning the tree with the assistance of the reduced error pruning method. Among the trees getting from different pruning iterations, it picks the best tree.

3.5 RANDOM TREE

A random tree is a supervised decision tree classifier which has an ensemble learning algorithm to produce decision trees by resampling features. So it is associated with the bagging method to choose the random set of data for constructing the decision tree [15]. While splitting the nodes, this classifier selects the node among the subset of nodes that predicts randomly. On the whole, the training data is resampled like bagging and classified with the random selection of nodes that help to produce decision trees are the random tree. It can deal with both regression and classification problems.

4. RESULTS AND DISCUSSION

This section explains the valuation of our methodology with the diabetic dataset and the evaluation results that follow. To evaluate the proposed method, we have used the metrics which can show the value of correctly predicted instances over incorrect instances. The decision table is generated using the classifiers and metrics like Accuracy (ACC), F1-score, precision, recall, and the ROC Curve were measured. The evaluation metrics used in our methodology are explained below:

A confusion matrix is a value table that is used to review the level of a classification model. Classification models are used to solve problems that have a definite outcome, such as predicting whether an instance is positive or not. From the table, we can get True positive, False positive, True negative and False negative [8].

Accuracy: It is the ratio of True Positive and True Negative over the total number of classifications.

$$Accuracy = \frac{True\ Postive + True\ Negative}{True\ Positive + True\ Negative + False\ Postive + False\ Negative}$$

Precision: It is the ratio of True Positive over the total number of positive classifications.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: It is the ratio of True Positive over the total number of true positive and false negative classifications.

$$Recall = \frac{True\ Positive}{Ture\ Positive + False\ Negative}$$

F measure: it is measured using precision and recall which counts false predictions.

$$F\ measure = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

After processing the data, the resample filter has been applied in the way to get the replicated data of (0-100) percentage in a split of 10. For each split, the concerned data has been classified with ensembles of REPTree classifier by bagging and their evaluation measures like Accuracy, Precision, Recall, F1 score and ROC are calculated using the confusion matrix. The obtained values are tabulated below in Table 1.

Like above the same data is resamples with bagging and classified by random tree classifier. The evaluation measures for each sample are tabulated below in Table 2.

Table 1: Evaluation measures of Bagging of REPTree classifier

S.NO	Classifier/ Base Classifier	Replication in (%)	Accuracy (%)	Precision	Recall	F1-Score	ROC
1.	Bagging/ REP Tree	0	75.2604 %	0.747	0.753	0.748	0.811
2.	Bagging / REP Tree	10	84.3602 %	0.842	0.844	0.841	0.915
3.	Bagging / REP Tree	20	86.2106 %	0.861	0.862	0.861	0.924
4.	Bagging / REP Tree	30	87.976 %	0.879	0.880	0.879	0.924
5.	Bagging / REP Tree	40	88.6512 %	0.886	0.887	0.885	0.939
6.	Bagging / REP Tree	50	87.7604 %	0.877	0.878	0.876	0.939
7.	Bagging / REP Tree	60	88.6808 %	0.887	0.887	0.885	0.946
8.	Bagging / REP Tree	70	89.5019 %	0.894	0.895	0.895	0.945
9.	Bagging / REP Tree	80	89.0014 %	0.890	0.890	0.889	0.949
10.	Bagging / REP Tree	90	89.4448 %	0.894	0.894	0.894	0.957
11.	Bagging / REP Tree	100	90.8203 %	0.908	0.908	0.908	0.960

Table 2: Evaluation measures of Bagging of Random Tree classifier

S.NO	Classifier/ Base Classifier	Replication in (%)	Accuracy (%)	Precision	Recall	F1-Score	ROC
1.	Bagging/ Random Tree	0	74.6094 %	0.746	0.746	0.746	0.799
2.	Bagging/ Random Tree	10	89.455 %	0.895	0.895	0.895	0.945
3.	Bagging/ Random Tree	20	89.7937 %	0.899	0.898	0.898	0.956
4.	Bagging/ Random Tree	30	91.6834 %	0.918	0.917	0.917	0.958
5.	Bagging/ Random Tree	40	92.186 %	0.922	0.922	0.922	0.963
6.	Bagging/ Random Tree	50	92.7951 %	0.928	0.928	0.928	0.970
7.	Bagging/ Random Tree	60	93.6482 %	0.936	0.936	0.936	0.974
8.	Bagging/ Random Tree	70	93.1801 %	0.932	0.932	0.932	0.973
9.	Bagging/ Random Tree	80	93.8495 %	0.939	0.938	0.939	0.976
10.	Bagging/ Random Tree	90	93.2831 %	0.933	0.933	0.933	0.979
11.	Bagging/ Random Tree	100	95.3125 %	0.953	0.953	0.953	0.984

For better understanding, the outcomes of REPtree classifier and Random tree classifier ensembles with bagging are visualized in the graphical plots shown in Figs. 2 to 6. From these plots it is easy to interpret the context of our paper's objective.

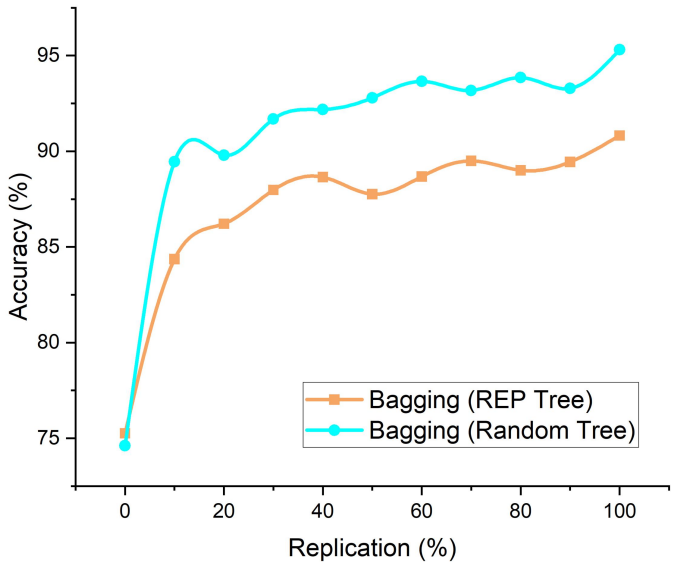

Figure 2: Accuracy of REPTree vs Random Tree

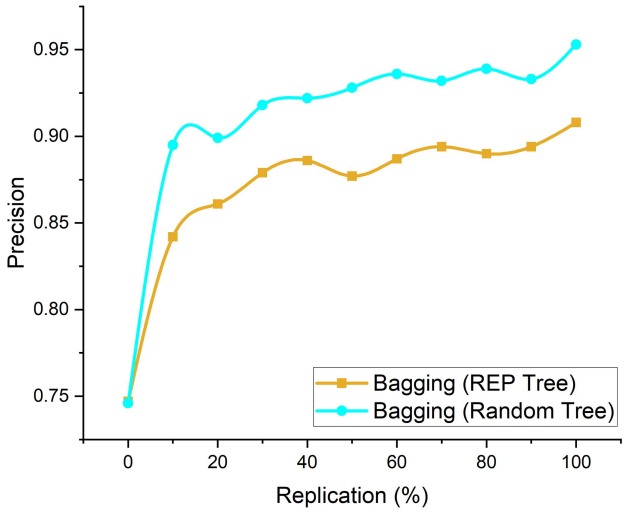

Figure 3: Accuracy of REPTree vs Random Tree

The graphical plots in Fig. 2 show that the accuracy of Bagging-Random tree classifier gives 95.3% whereas the REPTress gives 90.8% of accuracy.

The graphical plots in Fig. 3 show that the precision of Bagging-Random tree classifier gives 0.953 whereas the REPTress gives 0.908 of precision.

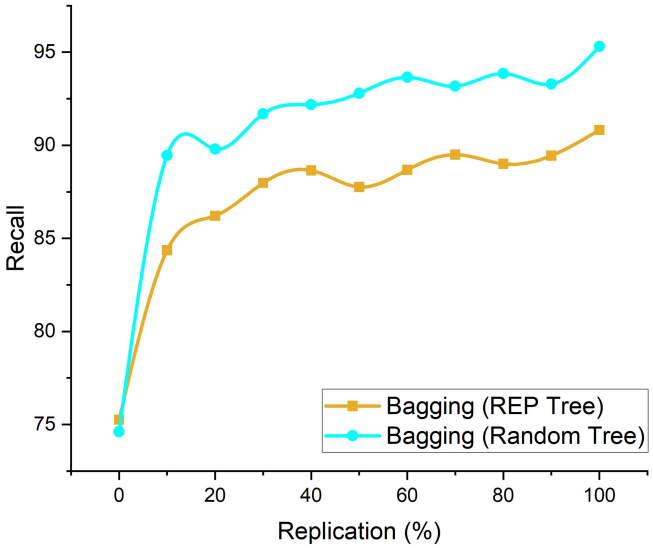

Figure 4: Recall values of ensembles by REPTree Vs Random Tree

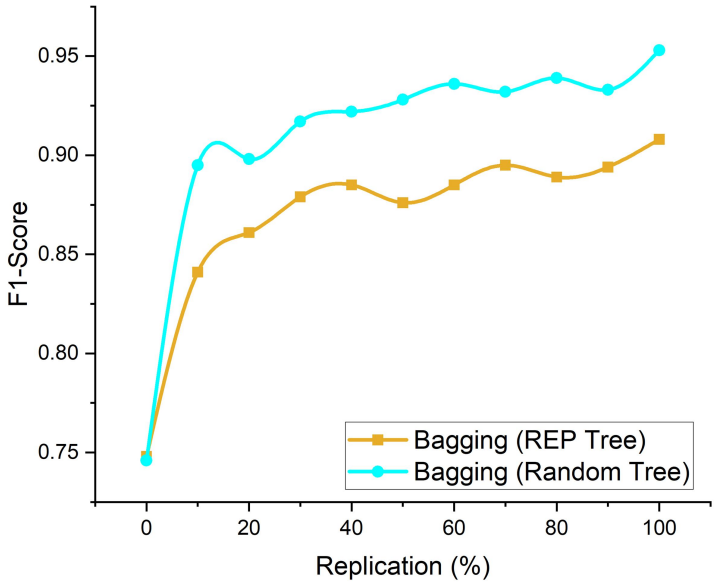

Figure 5: F1 score values of ensembles by REPTree Vs Random Tree

The graphical plots in Fig. 4 show that the Recall values of Bagging-Random tree classifier gives 0.953 whereas the REPTress gives 0.908 of Recall value.

The graphical plots in Fig. 5 show that the F1-score of Bagging-Random tree classifier gives 0.953 whereas the REPTress gives 0.908 of F1-score.

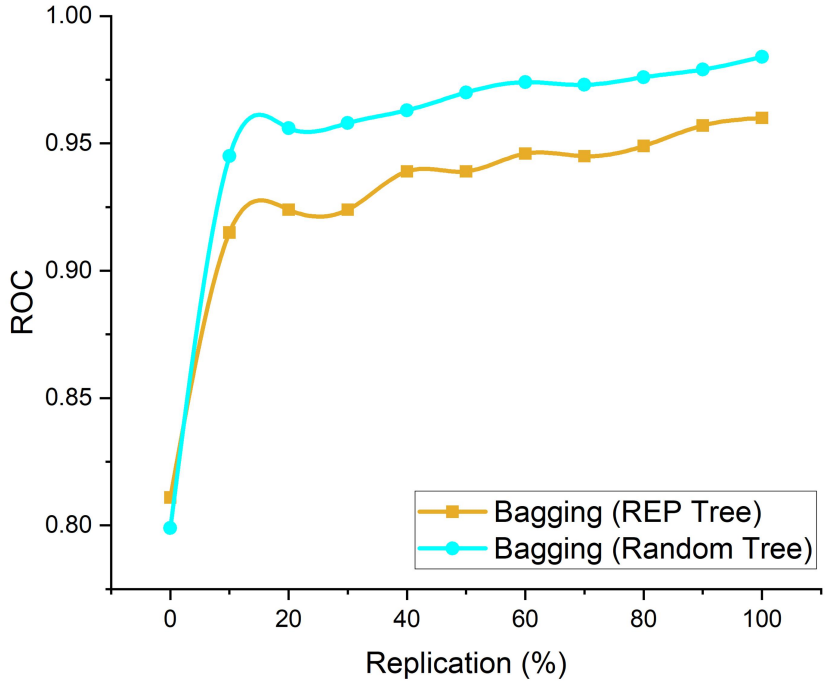

Figure 6: ROC values of ensembles by REPTree Vs Random Tree

The graphical plots in Fig. 6 show that the ROC value of the Bagging-Random tree classifier is 0.984 whereas the REPTress gives 0.96 of ROC value.

All the graphical plots show that Bagging with Random tree decision tree classifier outperforms the REPTree in this case. Though the related works proved that bagging with REPtree improves the prediction accuracy, here the ensemble by Random tree classifier yields high performance in all evaluation measures.

5. CONCLUSION

This study brings the places of interest to know the importance of replication of data based on ensembles of decision trees. This method helps to reduce the imbalance in the decision class of data. We aim to analyse the replication on which percentage the classifier able to bring the better performance. In that case, it was observed that if the evaluation measures are higher than when bagging with Random tree than REPTree. It seems the diabetic data . Thus in the specific data, sampling attains better performance by balancing them. Knowing medical data has more imbalance examples, our future study will be to analyse this method with other classifiers and thereby improve the resampling techniques.

DATA AVAILABILITY: https://www.kaggle.com/datasets/mathchi/diabetes-data-set.

CONFLICTS OF INTEREST: The authors declare no conflict of interest.

REFERENCES

1. Welvaars K, Oosterhoff JHF, van den Bekerom MPJ, Doornberg JN, van Haarst EP; OLVG Urology Consortium, and the Machine Learning Consortium. Implications of resampling data to address the class imbalance problem (IRCIP): an evaluation of the impact on performance between classification algorithms in medical data. JAMIA Open. 2023 May 31;6(2):ooad033. doi: 10.1093/jamiaopen/ooad033. PMID: 37266187; PMCID: PMC10232287.
2. Kim M, Hwang KB. An empirical evaluation of sampling methods for the classification of imbalanced data. *PLoS One* 2022; 17 (7): e0271260.
3. Fujiwara K, Huang Y, Hori K, et al. Over- and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis. *Front Public Health* 2020; 8: 178.
4. Yashoda and M. Kannan, "Analysis of a Population of Diabetic Patients Databases in Waikato", International Journal of Scientific & Engineering Research, vol. 2, no. 5, 2011.
5. A. Ayer, J. S and R. Sumbala, "Diagnosis of Diabetes Using Classification Mining Techniques", IJDKP, vol. 5, no. 1, pp. 01-14, 2015.
6. K. Sharmila and S. Manickam, "Efficient Prediction and Classification of Diabetic Patients from Big Data using R," International Journal of Advanced Engineering Research and Science, vol. 2, Sep 2015.
7. Sassanian and G. Hari Sekaran, "Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients," *International Journal of Science and Research*, vol. 4, April 2015.
8. NIyati Gupta, A. Rawal, and V. Narasimhan, "Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data", *IOSR Journal of Computer Engineering*, vol. 11, no. 5, pp. 70-73, 2013.
9. M. Chicheme. Said, and N. Setout, "Diagnosis of diabetes diseases using an Artificial Immune Recognition System 2 (AIRS2) with fuzzy Knearest neighbour," *Journal of Medical Systems*, vol.36, no.5, pp. 27212729, 2012.
10. P. Lee, "Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets", *International Journal of Environmental Research and Public Health*, vol. 11, no. 9, pp. 9776- 9789, 2014.
11. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Know Data Eng.(2009) 21:1041-4347. Doi:10.1109/TKDE.2008.239.
12. Breiman L (1996) Bagging Predictors.Machine Learning, 24:123–140.
13. Zontul, Metin & Aydin, F. & Doğan, Gaye & Sener, Selcuk & Kaynar, Oguz. (2013). Wind speed forecasting using reptree and bagging methods in Kirklareli-Turkey. Journal of Theoretical and Applied Information Technology. 56. 17-29.
14. Mohmad Badr Al Snousy, Hesham Mohamed El-Deeb, Khaled Badran, Ibrahim Ali Al Khlil,Suite of decision tree-based classification algorithms on cancer gene expression data,
Egyptian Informatics Journal,Volume 12, Issue 2,2011,73-82,https://doi.org/10.1016/j.eij.2011.04.003.
15. Tan, A. C., & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification.
16. Kalmegh, S.R. (2015). Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News.
17. C. Raghavendra, G. Gnana Priya, "Effects of data augmentation by replicating instances: Classification performance by Decision trees.", *Knowledge Transactions on Applied Machine Learning*, Vol. 01, Issue. 04, pp. 11–20, Sep. 2023.DOI:https://doi.org/10.59567/ktAML.V1.04.02