International Conference on Cyber Security and Computer Science (ICONCS'18),
Oct 18-20, 2018 Safranbolu, Turkey

A comparison of some soft computing methods on Imbalanced data

Md. Anwar Hossen[1], Fatema Siddika[2], Tonmoy Kumar Chanda[1], Touhid Bhuiyan[1]

[1] Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[2] Department of Computer Science and Engineering, Jagannath University, Dhaka, Bangladesh
anwar.swe@diu.edu.bd, shashi.csejnu@gmail.com, tonmoy616@diu.edu.bd, t.bhuiyan@daffodilvarsity.edu.bd

***Abstract* - Nowadays the computing trend is very large-scale and complex such as the Internet, banking system, online payment system, security, and surveillance system are generating a large amount of data every day. From these data, the percentage of imbalance data is quite high. These imbalanced data is misguiding a machine learning model and data mining technique. Learning from imbalanced data is a new complaint that has created increasing concentration from all over the world. This imbalanced data is creating a problem in learning problem with lots of unevenly distributed class. This paper concentrates on few realistic and appropriate data pre-processing techniques and produces an appropriate class evaluation process for the imbalanced data. An empirical distinction of few well-recognized soft computing methods such as Support Vector Machine (SVM), Decision Tree Classifier (DTC), K-Nearest Neighbor (KNN) and Gaussian Naïve Bayes (GNB) are used to find Accuracy, Precision, Recall and F-Measure from an imbalanced dataset. The imbalanced data were trained after a well-known over-sampling technique named Synthetic Minority Over-sampling Technique (SMOTE), under-sampling using Cluster Centroids (CC) technique and then applied a hybrid technique named SMOTEENN which is the combination of SMOTE and Edited Nearest Neighbor (ENN). Accuracy, Precision, Recall, F-Measure and Confusion matrix are used to evaluate the performance. In this task exhibit an experimental distinction of few well-recognized classification algorithms and performance measure that is authentic for the imbalanced dataset, this results we achieved. The result shows that hybrid method redacts better than Oversampling and under-sampling techniques.**

***Keywords* - Imbalanced data, Over-sampling, Under-sampling, SMOTEENN, SMOTE, Cluster Centroids**

I. INTRODUCTION

MODERN technological devices produce millions of data every day. Among these data, necessary information should be extracted for further used. But, one of the major challenges in machine learning and data mining field is the data imbalance problem. Some data class is dominated as they have the majority number of instance in the data set. On the other hand, some class data are minor in number; these also have some significance in data classification. This problem called class imbalanced problem in classification. Imbalanced class data problem is seen in the different aspect of the data area. Economic, environmental, commercial, software defect prediction, text classification, business risk mining, different medical diagnosis, medial data analysis, bank card fraud detection are the major area of the class imbalance problem. The high percentage of machine learning methods is designed for balanced data. These methods are working with well-balanced data. Class imbalanced data presents a new challenge to these learning methods to classify correctly. But existing methods have not classified these data well as these, not in a class balanced data. The class imbalance data problem can reduce the performance of learning methods. Learning algorithms are learning well for majority class data as they have lots of sample data. So the majority of classes are predicted well. But these results will crease problem in the different application of real life, such as an automatic target detection in an application [1], agricultural insect inspection [2], medical disease diagnosis [3] and others area.

The current research trend in the class imbalanced problem can be differentiated into two sides, one is algorithmic centric methods and sampling methods, as these already discussed in recent at the ICML [4] and AAAI [5]. In the sampling methods, all the class samples are leveling into the same amount of instance so that they are not imbalanced class. These done by two sampling methods, one is under-sampling the major class [6], and another one is over-sampling the minor class [7]. There are also hybrid techniques are available which one is the combination of under and oversampling method. On the other side, in algorithmic methods, adjusting the costs associated to improve the accuracy and performance [8], the bias of a classifier needs to be shifting in respect to the minor class data [9], also need to create boosting schemes [10]. Imbalanced data problem is creating a major problem when the data dimension is high. The number of features is much higher in microarray-based cancer classification [11]. The number of features in text classification is also high. The high dimensional class problem cannot work efficiently with the algorithmic method and sampling methods. Apart from this, feature selection is more important to overcome the over fitting problem than classification methods [12].

The aim of this paper is to study and find out the best methods that will perform best for class-imbalanced data. The dataset needs to preprocess as it contains some noise data. Different hyper parameter tuning, then compare different algorithms before and after re-sampling. Also, apply algorithms after sampling on imbalanced data sets. The tuning results point the best way which one is fast convergence to find best solutions. We keep our focus on obtaining a decision based on imbalanced data which sampling method is suited best among all the sampling

method. The imbalanced data were trained after a well-known Oversampling technique named Synthetic Minority Over-sampling Technique (SMOTE), Under-sampling using Cluster Centroids (CC) technique and then applied a hybrid technique named SMOTEENN which is the combination of SMOTE and Edited Nearest Neighbor (ENN). Accuracy, Precision, Recall, F-Measure and Confusion matrix are used to evaluate the performance. In this task exhibit an experimental distinction of few well-recognized classification algorithms and performance measure that is authentic for the imbalanced dataset, this results we achieved. The result shows that the hybrid method redacts better than over-sampling and under-sampling techniques.

The structure of this paper is as trails. Section I has described the introductory part of the work. Section II reviews the related work that has been studied by different researchers in this area. Section III presents a concise description of all the sampling method and all the algorithms. Section III presents the methodology and experimental process in brief. Also, describe the data processing and sampling step. Section IV shows the empirical results and comparative analysis of all the algorithms and Section V will present concluding remarks and future work of this work.

II. Related works

The challenge of imbalanced data makes it complex to implement experiments in the field of data mining. Although, some research has done to balance the imbalanced data, sampling the data and find key features to predict information to make the useful applications.

Numerous researches have been done to find out best the way to retrieve exact information from imbalanced data. Researchers are continuously trying to keep contribution in this field. In machine learning and data mining technique, learning from class-imbalanced data is a big problem. The best way to learning from imbalanced data is providing more balanced class data to a learning model tends to produce better outcome [13]. Data preprocessing, data resampling and parameter tuning is also a process to implement the learning algorithm method with different existing problems. Key feature selection is not a direct approach to classify the imbalanced data. So, based on the threshold value feature can be assessed and find out key feature from this data. The prediction model is based on a space under the Receiver operating characteristic curve [14]. Some real-world applications data do not have the equal number of the instance and data dimension, as a result, these applications such as fraud detection, diagnosis of disease these applications becomes problematic due to class-imbalanced data [15]. Sampling algorithms such as SMOTE is used for oversampling in the different research project that will balance the imbalanced data [16]. Few researchers pointed on the collapse of the distributive behavior of the class-imbalanced data. These imbalanced data will produce uncertain results for all the classes including major and minor class [17].

III. Methodology

The main contribution of this work is to find the best way to retrieve the information from imbalanced data by critically analyzing some established classification methods. In this work, we have used a car evaluation data set from UCI dataset directory [18]. This dataset has a limited number of attributes but imbalanced in terms of class. It is a multi-class dataset having six class value among them one class have the two-thirds number of instances.

Apart from this, this dataset has the limited number of instances which has created a new challenge. So, based on this information we can say that this dataset is moderately week which is the perfect example of class-imbalanced data. To solve this problem we have maintained some steps. At first, preprocess the data to replace text information with the numeric value, then re-sampling has done to prepare it to fit into the classification model to measure performance evaluation. The imbalanced data were trained after a well-known over-sampling technique named SMOTE, under-sampling using Cluster Centroids (CC) technique and then applied a hybrid technique named SMOTEENN which is the combination of SMOTE and Edited Nearest Neighbor (ENN). Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision tree (DT) and Naive Bayes (NB) classifier are applied after and before the sampling methods. Accuracy, Precision, Recall, F-Measure and Confusion matrix are used to evaluate the performance.

A. Dataset description

This dataset has 1728 instance with five attributes and one class attribute and also a multi-class dataset. The five attribute are Overall buying price, Price of maintenance, Number of door, Capacity in terms of persons to carry and estimated safety of the car. So, class attribute has four different values as ***unaccepted, accepted, good*** and ***very good***. Among them, ***unaccepted*** class has most 1210 instance which is 70.023% of the total instance, ***accepted*** class has 22.222 % instances, ***good*** class as 3.993 % instances and ***very good*** class has only 3.762 % instances. It was clear that ***unaccepted*** class has the highest number of instances and ***very good*** has limited number of instance. So this dataset is a perfect example of imbalanced data. The aim of this paper is to find a suitable sampling method and classifier that will classify these data perfectly and help to improve the performance of the application.

Table 1: Imbalanced data with class distribution

Class	Sample size	Percentage
unaccepted	1210	(70.023 %)
accepted	384	(22.222 %)
good	69	(3.993 %)
very good	65	(3.762 %)

B. Data pre-processing

The first step of our work is to pre-process the data using popular machine learning library named Sci-kit learn. In this dataset, the attribute value in text format. We need to convert it to the numeric value. The numeric value list corresponding to its text value showed in table II. Numeric values ranging from 0 to 5 for separate text attribute value.

Table II: Text attribute to corresponding numeric value

Attributes Value	Represented Value
very high	3
high	2
medium	1
low	0
unaccepted	0
accepted	1
good	2
very good	3
5 or More than 5	5

C. Data sampling

We have used three different data resampling methods. Firstly, SMOTE method is used for over-sampling the major class attribute. Secondly, Cluster Centroids (CC) is used under-sampling the minority class. Finally, SMOTEENN is used for both under and over-sampling purpose.

- *Over-sampling method*
  The task of over-sampling method is to re-sample the minority class instance. Here good and v-good is the minority class. So SMOTE is used as an over-sampling method. SMOTE will add some new instance based on existing instance and that will increase the instance number of a minority class.
- *Under-sampling method*
  The task of under-sampling method is to re-sample the majority class instance. Cluster Centroids is used to under-sample the unaccepted attribute and decrease the instance number. So that this decreased sample instance will close to minority instance number.
- *Hybrid sampling method*
  In hybrid sampling method both over-sampling and under-sampling has been done. Over-sampling is applied to minority class and under-sampling is applied to majority class instance. This method maintains the balance between the majority and minority instance.

D. Model and Classifier

We split the dataset into two parts, one is training data and another one is test data on a random basis. Among all data, 75% data are used as training data and rest 30 % data are used as test data. We have built four different models. The first model is applied without applying any sampling method to the imbalanced data. The second model is applied to over-sampled data. The third model is applied to under-sampled data. Finally, the fourth model is applied to the hybrid sample data. Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes (NB) and Decision Tree (DT) classifier are used to check the accuracy of each model to identify which model predicts better. The output value is checked by K-fold cross-validation for better understanding.

IV. Performance analysis

We have done another check on the trained model, whether our model is suffering from over and under-fitting problem.

If the training score is much higher than the test results and cross-validation score then our used model is affected by the over fitting problem. To overcome this problem we need to add more data which will help us to get the ride from the over-fitting problem. We test the learning curve from SVM as more likely to over-fit over to the more high accuracy. We have done this before sampling and after sampling.

Table III: Accuracy in different sampling method

Algorithm	Before sampling	After Over-sampling	After-Under sampling	After-Hybrid sampling
SVM	89.31	87.91	88.46	97.80
DT	83.53	87.19	71.15	99.53
GNB	75.43	73.14	78.85	77.32
KNN	86.13	86.98	71.15	98.27

We measure the performance-based Accuracy, Precision, Recall and F-measure. In Table III, accuracy value is shown based on different sampling method and classification algorithms. SVM, DT and KNN perform well in after applied Hybrid sampling method. But Gaussian Naïve Bayes perform well after under-sampling method. Hybrid sampling method performs well as it is a combination of both under-sampling and over-sampling.

Table IV: Accuracy, Precision, Recall, F-Measure in different sampling method on imbalanced data

Algorithm		Before sampling	After Over-sampling	After-Under sampling	After-Hybrid sampling
SVM	Accuracy	89.31	87.91	88.46	97.80
	Precision	71.38	82.62	89.87	97.51
	Recall	72.07	87.98	88.46	97.49
	F-Measure	71.09	88.19	87.87	97.49
DT	Accuracy	83.53	87.19	71.15	99.53
	Precision	56.16	87.37	71.76	99.44
	Recall	50.58	87.32	70.60	99.48
	F-Measure	52.78	87.28	70.55	99.46
GNB	Accuracy	75.43	73.14	78.85	77.32
	Precision	52.79	79.39	84.01	80.57
	Recall	63.50	73.29	79.12	77.03
	F-Measure	51.33	73.09	79.01	74.48
KNN	Accuracy	86.13	86.98	71.15	98.27
	Precision	67.54	87.47	73.40	98.02
	Recall	65.01	87.06	70.63	98.25
	F-Measure	65.99	87.16	71.37	98.08

We have also calculated the Precision, Recall, F-measure which are shown in Table IV. Precision is giving best performance in Hybrid sampling method when SVM is used

as classification algorithms. In DT, GNB and KNN hybrid sampling method will perform best for SVM classification algorithms. In terms of calculating F-measure value SVM, DT and KNN algorithm perform, Hybrid sampling method perform best among all the sampling method. But GNB classification algorithm under-sampling method performs best. When we calculate Recall with SVM, DT and KNN classification algorithm, hybrid sampling method perform best. But, in GNB classification algorithms under-sampling method perform best among all the other sampling method.

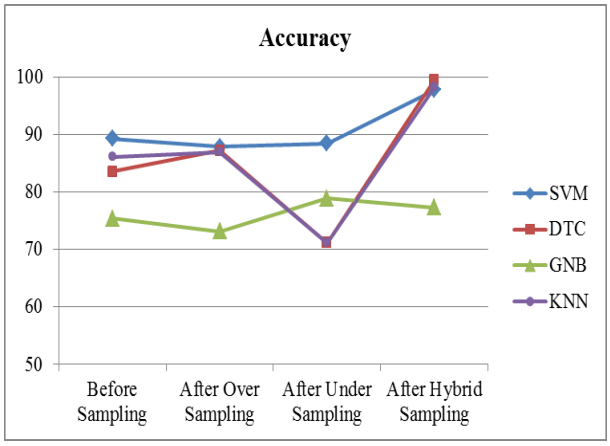

Figure 1: Accuracy curve for different sampling method

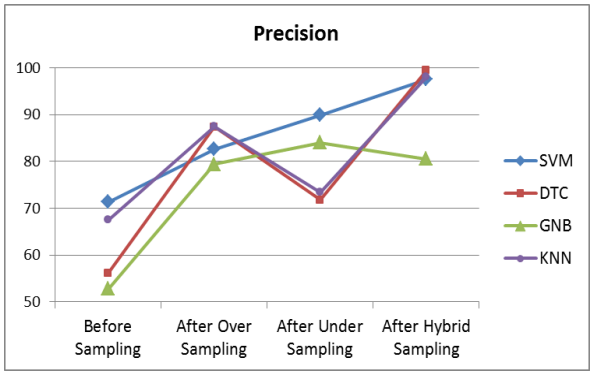

Figure 2: Precision curve for different sampling method

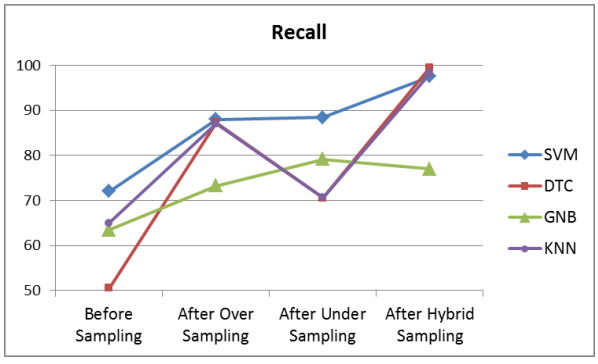

Figure 3: Recall curve for different sampling method

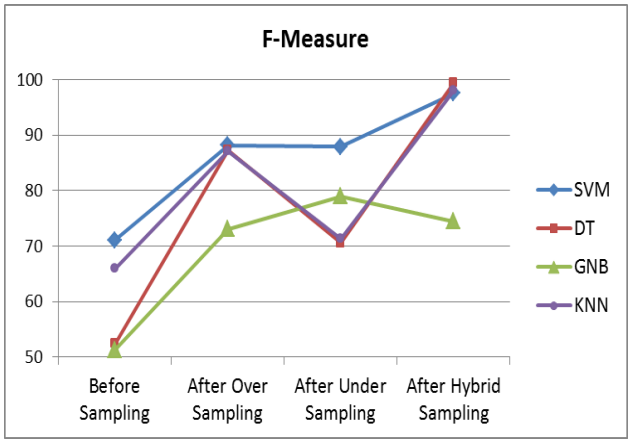

Figure 4: F-Measure curve for different sampling method

The value of accuracy shown in figure 1. We have received highest accuracy using DT classifier with Hybrid sampling method. SVM, DTC, GBN perform well with Hybrid sampling method. DT will perform well with the under-sampling method. Only GNB performing well with the under-sampling method. Figure 2 has described the Precision of all the sampling method. Decision tree classifier has received the highest value among all using Hybrid sampling method. DT and GNB perform well in both oversampling and hybrid sampling.

Figure 3 has shown the recall value of different sample method with different classification algorithm. KNN has performed well using both the over-sampling and hybrid sampling method. GNB perform not good with hybrid sampling method. F-measure value showed in Figure 4. Except for GNB, all the other classification performs well using all the sampling method. Among all DT perform best for F-measure.

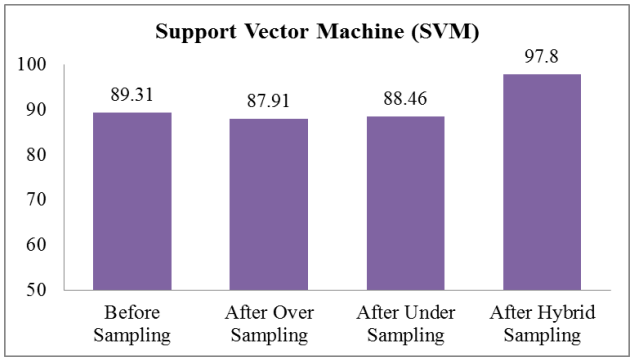

Figure 5: SVM comparison in different sampling method

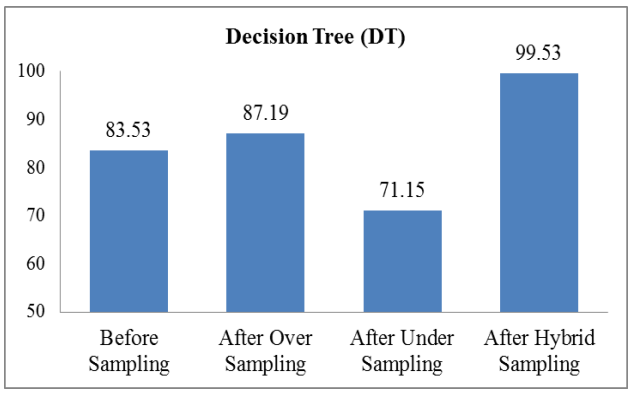

Figure 6: DT comparison in different sampling method

Figure 5 has shown SVM classifier performance comparison among all the sampling method. Without any sampling method, our model performed better than oversampling and under-sampling method. But Hybrid sampling method performs best among all the sampling method.

Figure 6 has shown Decision Tree classifier value comparison among all the sampling methods. The under-sampling method performs less among all. Oversampling and Hybrid sampling method perform best among all.

Gaussian Naïve Bayes classifier algorithm performs best with after under-sampling method in figure 7. Although, Hybrid sampling method performs best than without applying sampling method. Figure 8 shown that, under sampling method perform less among all and hybrid sampling method performs best among all.

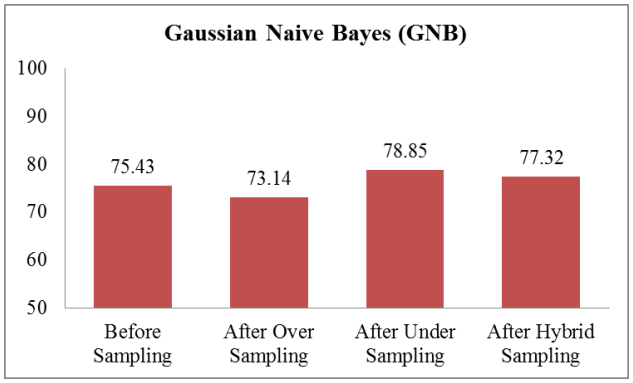

Figure 7: GNB comparison in different sampling method

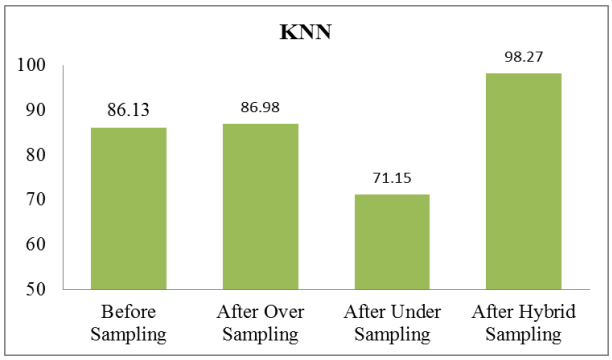

Figure 8: KNN comparison in different sampling method

V. CONCLUSION AND FUTURE WORK

This research paper is based on the imbalanced data and we have illustrated an analytical comparison of different classification algorithm with before sampling and after different re-sampling method. We have used cross-validation process which is playing an important role to find perfect performance evaluation value. Different analyses are required in near future to identify the application of cross-validation in the different application. Some large dataset can be used to figure out more outcomes with sampling method. Though, imbalanced data are more prone to over fitting problem. We can extend our work with some real-time data and also apply another algorithm, sampling method to identify exact measurement. We can also extend our work to extract the key feature from the huge number of features.

REFERENCES

[1] Casasent, D. and Chen, X.-W. 2004. Feature reduction and morphological processing for hyperspectral image data. Applied Optics, 43 (2), 1-10.

[2] Casasent, D. and Chen, X.-W. 2003. New training strategies for RBF neural networks for X-ray agricultural product inspection. Pattern Recognition, 36(2), 535-547.

[3] Nunez, M. 1991. The use of background knowledge in decision tree induction. Machine Learning, 6, 231-250.

[4] Chawla, N., Japkowicz, N., and Kolcz, A. editors 2003. Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets.

[5] Japkowicz, N. editor 2000. Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets. AAAI Tech Report WS-00-05.

[6] Kubat, M. and Matwin, S. 1997. Addressing the curse of imbalanced data set: One sided sampling. In Proc. of the Fourteenth International Conference on Machine Learning, 179-186.

[7] Kubat, M. and Matwin, S. 1997. Learning when negative examples abound. In Proceedings of the Ninth European Conference on Machine Learning ECML97, 146-153.

[8] Domingos, P. 1999. MetaCost: a general method for making classifiers cost-sensitive. Proc. of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 155-164.

[9] Huang, K., Yang, H., King, I., Lyu, M., 2004. Learning classifiers from imbalanced data based on biased minimax probability machine. Proc. of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2(27), II-558 - II-563.

[10] Chawla, N., Lazarevic, A., Hall, L., and Bowyer, K. 2003. SMOTEBoost: Improving prediction of the minority class in boosting. Principles of Knowledge Discovery in Databases, LNAI 2838, 107-119.

[11] Y. Kamei, A. Monden, S. Matsumoto, T. Kakimoto, and K.-i. Matsumoto, "The effects of over and under sampling on fault-prone module detection," in *First International Symposium on Empirical Software Engineering and Measurement, 2007. ESEM 2007.* IEEE, 2007, pp. 196–204.

[12] N. E. Fenton and N. Ohlsson, "Quantitative analysis of faults and failures in a complex software system," *IEEE Transactions on Software Engineering*, vol. 26, no. 8, pp. 797–814, 2000.

[13] Ertekin, Seyda, et al."Learning on the border: active learning in imbalanced data classification." Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM, 2007

[14] Chen, Xue-wen, and Michael Wasikowski. "Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008

[15] Khoshgoftaar, Taghi M., Moiz Golawala, and Jason Van Hulse. "An empirical study of learning from imbalanced data using random forest." 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007). Vol. 2. IEEE, 2007

[16] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321- 357.

[17] He, Haibo, and Edwardo A. Garcia. "Learning from imbalanced data." IEEE Transactions on knowledge and data engineering 21.9 (2009): 1263-1284.

[18] Car Evaluation Dataset, https://archive.ics.uci.edu/ml/datasets/car+evaluation, 17 August 2018.