ISSN: 2366-1313

Deep Learning basedSMOTE Algorithmfor License Plate Recognition

[1]Kovvurisarveswara Rao, [2]M. Rama Bhadra Rao,

[1]MCA Student, Dept. Of MCA, Swarnandhra College of Engineering and Technology, Seetharampuram, Narsapur, Andhra Pradesh 534280,

Sarveswararaokovvuri123@gmail.com

[2],Assistant Professor, Dept. Of MCA,Swarnandhra College of Engineering and Technology, Seetharampuram, Narsapur, Andhra Pradesh 534280,

Abstract:*Severe class imbalance is one of the main conditionsthat make machine learning in cyber security difficult. A varietyofdatasetpre-processingmethodshavebeenintroducedoverthe years. These methods modify the training dataset by over-sampling,under samplingoracombinationofbothtoimprovethe predictive performance of classifiers trained on this dataset.Althoughthesemethodsareusedincyber securityoccasionally,a comprehensive, unbiased benchmark comparing their performance over a variety of cyber security problems is missing. Thispaperpresentsabenchmarkof16pre-processingmethodsonsix cyber security datasets together with 17 public imbalanceddatasetsfromotherdomains.Wetestthemethodsundermultiplehyper parameter configurations and use an AutoML system totrainclassifiersonthepre-processeddatasets,whichreducespotential bias from specific hyper parameter or classifier choices.Specialconsiderationisalsogiventoevaluatingthemethodsusing appropriate performance measures that are good proxiesfor practical performance in real-world cyber security systems.The main findings of our study are: 1) Most of the time, a datapre-processing method that improves classification performanceexists. 2) Baseline approach of doing nothing outperformed alarge portion of methods in the benchmark. 3) Oversamplingmethods generally outperform under sampling methods. 4) themostsignificantperformancegainsarebroughtbythestandard SMOTE algorithm and more complicated methods providemainlyincrementalimprovementsatthecostofoftenworsecomputationalperformance.*

Keywords-component; machine learning, cyber security, classification,imbalancedclassification

I INTRODUCTION

Aclassificationproblemissaidtobeclass-imbalancedwhentheclasspriorprobabilityof atleastoneclass,usuallytheclassofinterest,issignificantlysmallerthanthepriorprobabilityofsomeotherclass.Class-imbalancedproblemsoccuracrossa variety of machine learning application domains such asmedicine [48], finance [47],[58], astronomy [32] and manyothers.

Specifically, in cyber security, virtually all of the frequentlystudied classification problems are class-imbalanced (e.g. intrusion detection [13], malware detection [18], phishing detection [21]). Furthermore, the class imbalance is frequentlysevere, with prior probabilities of the classes of interest being10^{-5} andlower[13]becauseseveremaliciousbehaviourandattacksare(thankfully)extremelyrare.Forexample, innetworktelemetry,themajorityoflogsarerelatedtoordinary (benign) traffic, and only a tiny portion is related tomaliciousactivities.Interestingly,aclassimbalanceexistsevenin the already small portion of telemetry related to maliciousactivities,astheprevalenceoflow-riskactivitiessuchasmalicious advertising and tracking is much greater than theprevalence of the most exciting threats with high severity (e.g.remote access Trojans, ransom ware, APTs). The difficultiesandtheimportanceoftheseveresclassimbalanceproblemin cyber security were, to our knowledge, first mentioned byAxel's son [7] in 2000. Now, more than two decades later, aclass imbalance is still among the most critical factors thatmakemachinelearningincyber securitydifficult[5],[27].

Whileslightclassimbalancedoesnotusuallyposeaproblem, once it reaches a certain degree, machine learningclassifiers without appropriate countermeasures cannot learnreliablyfromthedata[31].Insuchcases,classifierstendto become biased toward the majority class and neglect theunderrepresented one, resulting in a situation in which overallaccuracy is high due to the classifier predicting the majorityclass all of the time. However, other, more relevant performance measures that reflect performance on all classes arepoor.

Over the years, there has been a great deal of interest in theimbalanced classification problem. Many different approacheswereproposedspanningallthemajorstagesofmachinelearning model development. These stages are [6]: 1) datamanagement,2)modellearningand3)modelverification.Approaches applied in the first stage are sometimes called *data-level methods*, while approaches applied in the second stagearecalled*algorithm-levelmethods*[34].Multipleliteraturereviews [15], [35], [54], [31], [34] summarising the conceptsandpopularapproachesineachstage havebeenpublishedovertime.

In this paper, we focus on data-level methods suitable forclass-imbalancedlearning.Theideabehindthesemethodsis centred on modifying the distribution of the trainingdatasettomakeitlessimbalanced.Thisis,inprinciple,achieved via either oversampling the minority class or un-dersamplingthemajorityclass.Manysuchmethodshavebeen published over the years, and sometimes the rationalebehind them is contradicting. The current situation concerningwhichmethodsareworthusingwhenandwhichareperhaps unnecessarilycomplexforlittletonobenefitis unclear.Inthe worst case, this may lead to a promising, high-performingmethod being ignored by the field in favour of a simpler ormore traditional one. Our goal in this paper is to improveunderstanding of strengths, weaknesses and various trade-offs(both predictive and computational) between a range of themostwell-knowndata-levelmethods.

To achieve this, we perform an extensive empirical bench-markofdata-levelmethodsonvariousdatasetsspanningdifferent application domains with special attention dedicatedto the cyber security domain. We aim to compare the methodsobjectivelyonasequalgroundaspossible,whichishelpedbyusnothavinganyhorsesintherace.Tothebestofour knowledge, there does not exist a more comprehensivebenchmark of data oversampling and under sampling methods.The results help better navigate the problem landscape andselectappropriatemethods,hopefullyleadingtoimprovedpredictive performance on various tasks in cyber security andotherdomains.

II RELATEDWORK

Over the years, many statistics pre-processing techniques suitable for sophistication-imbalanced mastering had been published, but in comparison, only a surprisingly small wide variety of benchmarks encompassing an extensive range of each methods and datasets exist. Typically, each guide introducing a

ISSN: 2366-1313

brand new approach includes experimental assessment, but the scope of those experiments tends to be small. For instance, a paper introducing ADASYN [30] carries experiments on 5 datasets and compares the method most effective against SMOTE [16] and undeniable decision tree baseline.

With that said, there exist already guides that attention particularly on evaluating pre-processing methods, however normally, they have a tendency to focus handiest on oversampling strategies. Most of this research [26], [3], [10] also is carried out on a exceptionally small wide variety of datasets. An exception is a take a look at by Kava's [36], which may be very great both in terms of strategies compared and datasets used. However, it focuses handiest on oversampling techniques and additionally does not include experiments inside the cy- be security domain. Additionally, not one of the researches above performs as huge a search in hyper parameters and successive classifier fashions as we do.

In the cyber security domain, Wheel us et al. [59] com- pared several pre-processing techniques at the UNSW-NB15 [45] dataset. Bagui and Li [9] in comparison 5 pre-processing methods on six network intrusion detection datasets and used a feed- ahead neural network with one hidden layer for classification. Furthermore, the most famous records pre-processing

Strategies are acknowledged and used in cyber security [1], [43], [2], [53], [8], however to our know-how, a broader comparative observe is lacking.

Lastly, previous studies also summarise the consequences of in- dividable methods right into a single variety. Usually, that is the common rank or scores the approach done across all datasets. In this paper, we provide rank distribution density plots instead of single numbers. These plots display a extra entire image as the ranks tend to have a massive variance and overlap across the datasets.

III METHOD

This phase contains an outline of pre-processing strategies used within the benchmark. For the sake of area, we refrain from thorough causes and seek advice from unique publications.

Oversampling Methods

Oversampling techniques represent one feasible technique to fixing the imbalanced category trouble. The most important goal of oversampling strategies is to alter the empirical distribution with the aid of growing the range of samples belonging to the minority magnificence. The empirical distribution is modified both by using duplicating the prevailing samples or generating new synthetic samples till the desired imbalance ratio is reached. The maximum honest approach is called Random Over- sampling, which, as its call shows, randomly duplicates

Already present samples in the dataset.

One of the primary and maximum widely used oversampling strategies which produce artificial facts samples is SMOTE [16]. It creates new artificial examples on the line segments between current examples from the minority magnificence. SMOTE, but, considers all of the minority samples to have the identical importance. It does no longer don't forget prior sample density and does now not care approximately the neighbourhood of a minority pattern. Various enhancements had been proposed to resource those shortcomings of the unique SMOTE set of rules. We include 4 of those upgrades in our benchmark, particularly BorderlineSMOTE [28], SVM SMOTE [46], KMeansSMOTE [38]

And ADASYN [30].

BorderlineSMOTE, in place of SMOTE, selects only minority samples with at least half of their neighbours belonging to most people elegance. The idea behind this method is that minority samples surrounded via greater majority samples are close to the so-called selection boundary and are, consequently, vital in category.

SVM SMOTE builds on the equal idea however uses the SVM algorithm rather than the kNN set of rules to locate minority samples near the decision boundary.

KMeansSMOTE tries to generate new synthetic samples in regions in which minority samples are sparse and for that reason avoids further inflation of dense areas. It uses the KMeans clustering algorithm to stumble on clusters containing more minority samples than majority samples. This avoids interpolation among noisy minority samples. Subsequently, new samples are generated in every decided on cluster primarily based on its density, i.e. Extra samples are generated in sparse clusters.

ADASYN differs from SMOTE by way of assigning weights to minority samples based totally on their problem in learning. Difficulty in getting to know, in this situation, way the portion of ok-nearest neighbours that belong to the opposite class. More synthetic information is generated in regions in which it's miles hard to learn minority

ISSN: 2366-1313

samples, and much less information is generated in other, less difficult-to-study areas.

Under sampling Methods

Under sampling strategies awareness on the majority magnificence, rather than the oversampling strategies, to address the problem of imbalanced classification. These techniques reduce the wide variety of samples in most people elegance to create a extra balanced distribution of samples between classes. Most of the beneath- sampling techniques mentioned are so-known as prototype selection strategies. Prototype selection methods reduce the variety of samples by getting rid of useless samples from the dataset and using most effective a subset of the authentic statistics. The Cluster Centurions method is the most effective example of a prototype era approach used in the benchmark. Prototype era strategies reduce the variety of samples by producing new samples,

E.g.Centurions of clusters acquired through the KMeans set of rules, as opposed to the usage of a subset of the authentic ones.

Again, the most effective approach based totally on random selection and removal of most people samples is called Random Under- sampling. The following several strategies construct on the kNN algorithm and adjust it to obtain barely one of kind consequences.

Condensed Nearest Neighbours - CNN [29] reduces a probably massive dataset into a constant dataset which, whilst used inside the 1-NN rule, efficaciously classifies all the examples from the original dataset.

Edited Nearest Neighbours - ENN [60] classifies all samples in the elegance to under sample with the aid of computing k-nearest neighbours for every on the whole unique set. It then proceeds to do away with all such samples under consideration whose actual label does no longer match the label of maximum of their neighbours.

Repeated Edited Nearest Neighbours [55] includes repeating the previous set of rules more than one instances to lessen the wide variety of majority samples even further.

All KNN [55] uses the equal concept as the 2 preceding pre-processing methods to eliminate samples from the majority magnificence while there may be a label war of words among a sample below consideration and its ok-nearest neighbours. However, in preference to using a set quantity of neighbours to check a settlement, it begins via searching on the unmarried nearest neighbour, then two nearest neighbours and so on, till it reaches okay- nearest neighbours. A pattern is saved in the general public magnificence best if its label is of the same opinion in all cases.

Near Miss [41] is a group of 3 algorithms that use kNN to select majority elegance samples to maintain. Near Miss 1 selects majority samples that show off the smallest common distance to N closest minority samples. In contrast, Near Miss 2 selects the ones samples that exhibit the smallest average distance to N furthest minority samples. Near Miss 3 selects a given range of closest majority samples for every minority sample.

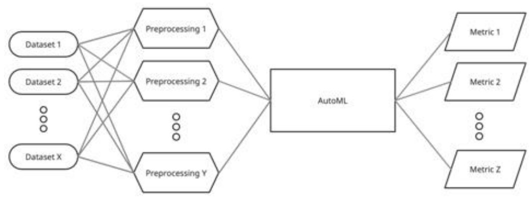

Fig. 1. High-level architecture of the benchmarking framework.

IVEXPERIMENTSETUP

We constructed a benchmark framework too efficiently and robustly behaviour experiments with many exclusive pre-processing methods over many datasets reporting as many evaluation metrics as provided. The centre concept of the framework is depicted in Figure 1. Each run combines a dataset, a pre-processing technique, and an instantiation of its hyper parameters observed the use of a grid seek. In every run, a pre-processing method is applied to the education a part of a dataset, yielding a new resample schooling set, which is then exceeded to the AutoML factor of the framework. We use a ultra-modern AutoML framework Auto-Sklearn [23] for deciding on, schooling and tuning a classifier suitable for a given dataset. We offer greater details about Auto-Sklearn in Section IV-A1. Once a classifier has been educated, we carry out predictions the use of unseen examples from the take a look at set and report evaluation rankings achieved.

A. **Benchmark Setup**

We ran a benchmark masking 16 pre-processing techniques discussed in Section III and one no-op baseline technique. We included several possible hyper parameter configurations for each technique proven in Table I. All implementations of pre-processing methods used inside the benchmark originate from the Imbalanced Learn library [40].

Every pre-processing method turned into run on 23 public and proprietary datasets proven in Table II. Non-cyber security public datasets have been downloaded from OpenML [57]. We selected datasets carefully based totally on multiple criteria such as dataset size, quantity of lacking values and imbalance ratio. We required every dataset from OpenML to be binary and to have as a minimum 5000 samples; at maximum 20% of samples ought to have lacking values, and the minimal imbalance ratio had to be 1:10. Although we awareness only on binary class, imbalanced datasets arise inside the multi-elegance environment as properly. However, for the sake of simplicity and consistency with different authors and publications, we cognizance only at the binary case. The generalisation to the multi-magnificence placing may be without problems achieved by way of employing one-vs.-one or one-vs.-rest schemes to pre-processing techniques and micro and macro averaging to evaluation metrics. We used seventy five% of records samples from every dataset as a training set and the last 25% as a trying out set. The cut up become performed to maintain the original imbalance in both units.

We utilised Auto-Sklearn IV-A1 to discover, educate and tune the first-class-performing classifier on the education set the usage of five-fold pass-validation because the validation technique. Auto-Sklearn changed into set to optimise the ROC AUC score IV-B2. Each run was given a total of 30 minutes for education on public datasets; a single system studying model had 10 minutes to finish education. Unsuccessful runs have been now not repeated. Due to their sizes, it was sufficient to devote handiest five mines to Auto-Sklearn on proprietary datasets, and no repetitions have been wished. We did no longer restrict the time for pre-processing step in any way to attain data about the overall performance of pre-processing techniques on datasets of various sizes.

1) AutoSklearn: Auto-Sklearn [23] is a library for automated model selection and hyper parameter tuning. Auto- Sklearn permits us to explore many fashions without introducing our very own bias into the technique. We chose Auto-Sklearn for its appreciably better overall performance than other competing AutoML systems [23]. Although the second one model of Auto-Sklearn, bringing enormous advances [22], has been available on the grounds that 2020, we selected now not to apply it because it turned into nonetheless

ISSN: 2366-1313

in an experimental phase at the time of the experiments.

Auto-Sklearn extends current Auto ML architectures utilising the Bayesian optimiser via using meta-mastering and ensemble building to in addition improve the gadget's performance. We briefly explain how each of the additives works and provide remarks in instances wherein we have had to adjust the behaviour of Auto-Sklearn to allow whole manage over the experiment.

Method	Hyper parameter Configurations
Baseline	1
Random Oversampling	2
SMOTE	4
Borderline SMOTE	16
SVM SMOTE	8
K Means SMOTE	4
ADASYN	4
Random Under sampling	2
CNN	2
ENN	4
Repeated ENN	4
All KNN	4
Near Miss	12
Tomek Links	1
One-Sided Selection	2
NCL	8
Cluster Cancroids	4
Σ	82

TABLE I
HYPERPARAMETER CONFIGURATIONS FOR PREPROCESSING METHODS. THE TABLE SHOWS THE NUMBER OF AVAILABLE HYPERPARAMETER CONFIGURATIONS IN THE BENCHMARK.

Name	Maj. Size	Min. Size	Imbalance
Asteroid	125,975	156	807.532
Credit Card Subset [17]	14,217	23	618.130
Credit Card [17]	284,315	492	577.876
PC2 [50]	5,566	23	242.000
MC1 [50]	9,398	68	138.206
Employee Turnover	33,958	494	68.741
Satellite [25]	5,025	75	67.000
BNG - Solar Flare	648,320	15,232	42.563
Mammography	10,923	260	42.012
Letter [24]	19,187	813	23.600
Relevant Images	129,149	5,582	23.137
Click Prediction V1	1,429,610	66,781	21.407
Click Prediction V2	142,949	6,690	21.368
Amazon Employee	30,872	1,897	16.274
BNG - Sick	938,761	61,239	15.329
Sylva Prior	13,509	886	15.247
BNG - Spect	915,437	84,563	10.826
CIC-IDS-2017 [51]	227,132	5,565	40.814
UNSW-NB15 [45]	164,673	9,300	17.707
CIC-Evasive-PDF [33]	4,468	555	8.050
Ember [4]	200,000	26,666	7.500
Graph - Embedding [20]	394	154	2.558
Graph - Raw [20]	394	154	2.558

TABLE II
DATASETS. THE TABLE SHOWS BASIC INFORMATION ABOUT THE DATASETS USED IN THE BENCHMARK. THE UPPER PART OF THE TABLE SHOWS PUBLICLY AVAILABLE NON-CYBERSECURITY DATASETS; THE LOWER PART SHOWS CYBERSECURITY DATASETS AND TWO PROPRIETARY DATASETS CONCERNING THE CLASSIFICATION OF NODES IN NETWORK GRAPHS.

suitable for finding the extreme of objective functions expensive to evaluate, such as tuning hyper parameters in a machinelearning model, in as few sampling steps as possible [14].Bayesian optimisation fits a probabilistic model capturing arelationshipbetweenhyper parametersandmodelperformance.Theprobabilisticmodelsuggestsapromisingconfigurationofhyper parametersbasedonitscurrentbeliefs.

V DISCUSSION

In this section, we take a deeper look at the results andsummarise the most important findings and recommendations.Firstly, we analyse the summary results over all datasets. Secondly, we take a look specifically at the results on cyber security datasets to see whether the findings and

ISSN: 2366-1313

recommendationsdiffer.Lastly,wediscussth ecomputationalperformanceofthestudiedm ethods.

Tostartwith,letusconsidertheperformanceof thebaselinemethod,wherenopre-processingisappliedtothetraining dataset. The baseline method achieved a reasonablerankamongallmethodsandacross allthemeasures.InthePRAUCandROCAUC measuresdisplayedinFigures2and 3, the baseline consistently ranked in the top half of thestudiedmethods.IntheP-ROCAUCmeasureinFigure4,thebaselineus uallyendsupinthesecondhalfofmethods,buti tisrarelytheworstmethod.Thebaseline'sperf ormance is slightly surprising because all the methods generally claimto bring performance gains in these types of problems. Weoffer several hypotheses to explain this result. First, we arelookingatsummarystatisticsacrossavarie tyofdatasets.Some methods are not meant to be used in every scenario butare tailored for datasets with specific properties. For example,NearMiss[42]aimstoremovesampl esattheboundaryof the majority class. This may work if these samples aremainlypresentduetonoise,butiftheyareva lidsamples;such removal may significantly increase the false positive rateoftheclassifier.Second,weperformhyper parametertuningoftheclassificationlayervia AutoML,whichisamuchstrongerbaselinetha nusual.

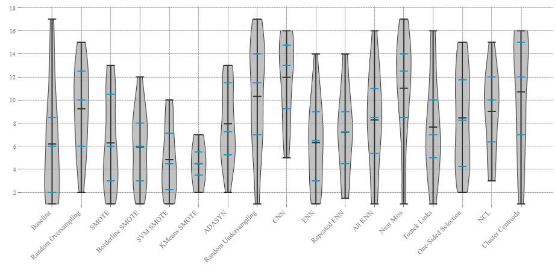

Fig. 2. **Area under PR Curve (PR AUC).** Ranks for each method were measured across all datasets in the benchmark.

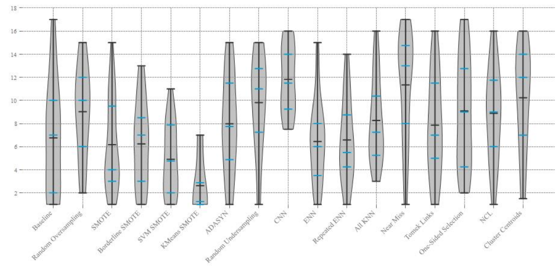

Fig. 3. **Area under ROC Curve (ROC AUC).** Ranks for each method were measured across all datasets in the benchmark.

A major takeaway is that, in general, oversampling methodsoutperformunder samplingmethods.Thispatternisvisibleacros s all performance measures and is most evident in P-ROCAUC,whichweconsidertobethemostpr acticallyrelevantmeasure.Beforetheexperi ment,ourintuitionwasthat under sampling of the majority class is one of the leastpreferable ways to address class imbalance because it providestheclassifierwithlessinformationto extract.Theexperiment'sresultssupportthisi ntuition.Onrareoccasions,under samplingmay perform well. However, unless we have a good reason tobelieve that it may improve a particular dataset or we havecomputation power to spare, we

ISSN: 2366-1313

should prefer rebalancing thedataset via oversampling.

VI CONCLUSION

We have carried out a novel observe of 16 pre-processing methods on 23 datasets, of which six are from the cyber security area. We studied both predictive and computational performance. To that give up, we applied a big-scale test which employs AutoML to don't forget a wide range of classifiers and includes a hyper parameter search to take away ability sources of bias found in beyond benchmarks.

Our fundamental findings are that the usage of dataset pre-processing while managing class-imbalanced classification is frequently useful. However, on the identical time, a big part of the techniques fails to always outperform the baseline solution of doing not anything. Most of the time, oversampling strategies out per shape under sampling techniques, however exceptions exist. Amongthe oversampling methods, the conventional SMOTE algorithm achieves the maximum sizeable performance gains, while its more state-of-the-art versions possibly cause upgrades of simplest incremental nature.

When we isolated our evaluation most effective to the cyber security datasets which span a couple of cyber security domain names, we reached the identical conclusions as above.

Finally, it's miles important to be aware that the approach ranking is prompted by means of the performance degree selected. We encompass a couple of performance measures which are comprehensive and suitable in realistic category situations while dealing with magnificence imbalance. Even though the specifics of the rankings range via degree, the primary takeaways mentioned above are consistent.

REFERENCES

1. MostofaAhsan,RahulGomes,andAnneDenton.Smoteimplementationon phishing data to enhance cyber security. In *2018 IEEE InternationalConferenceonElectro/InformationTechnology(EIT)*,pages0531–0536.IEEE,2018.

2. BathiniSaiAkash,PavanKumarReddyYannam,BokkasamVenkataSaiRuthvik, Lov Kumar, Lalita Bhanu Murthy, and Aneesh Krishna.Pre-dicting cyber-attacks on iot networks using deep-learning and differentvariants of smote. In *International Conference on Advanced InformationNetworkingandApplications*,pages243–255.Springer,2022.

3. Adnan Amin, Sajid Anwar, Awais Adnan, Muhammad Nawaz, NewtonHoward, Junaid Qadir, Ahmad Hawalah, and Amir Hussain. Comparingoversamplingtechniquestohandletheclassimbalanceproblem:Acustomerchurnpredictioncasestudy.*IEEEAccess*,4:7940–7957,2016.

4. HyrumSAndersonandPhilRoth.Ember:anopendatasetfortraining static pe malware machine learning models.*arXiv preprintarXiv:1804.04637*,2018.

5. Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke,Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and KonradRieck. Dos and don’ts of machine learning in computer security. In *31stUSENIX Security Symposium (USENIX Security 22)*, pages 3971–3988,Boston,MA,August2022.USENIXAssociation.

6. RobAshmore,RaduCalinescu,andColinPaterson.Assuringthemachine learning lifecycle: Desiderata, methods, and challenges. 54(5),May2021.

7. Stefan Axelsson.The base-rate fallacy and the difficulty of intrusiondetection.*ACMTransactionsonInformationandSystemSecurity(TISSEC)*,3(3):186–205,2000.

8. Salahuddin Azad, Syeda Salma Naqvi, Fariza Sabrina, Shaleeza Sohail,and Sweta Thakur.Iot cyber security: On the use of machine learningapproaches for unbalanced datasets. In *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6.IEEE,2021.

9. Sikha Bagui and Kunqi Li.Resampling imbalanced data for networkintrusiondetectiondatasets.*Journal ofBigData*,8(1):1–41,2021.

10. RicardoBarandela,RosaMValdovinos,JSalvadorSánchez,and Francesc J Ferri.The imbalanced training sample problem: Under orover sampling?In *Joint IAPR international workshops on statisticaltechniques in pattern recognition (SPR) and structural and syntacticpatternrecognition(SSPR)*,pages806–814.Springer,2004.

11. JanBrabec,TomášKomárek,VojtěchFranc,andLukášMachlica.On model evaluation under non-constant class imbalance. In *InternationalConferenceonComputational Science*,pages74–87.Springer,2020.

12. Prasadu Peddi (2015) "A review of the academic achievement of students utilising large-scale data analysis", ISSN: 2057-5688, Vol 7, Issue 1, pp: 28-35

13. Prasadu Peddi (2015) "A machine

ISSN: 2366-1313

learning method intended to predict a student's academic achievement", ISSN: 2366-1313, Vol 1, issue 2, pp:23-37.