Article | Received 15 July 2023; Accepted 2 August 2023; Published 1 December 2023
https://doi.org/10.55092/pcs2023020010

# Imbalanced data handling techniques for classification: a state-of-the-art review

**Asma Basharat[1,2,*], Amna Ali[3], Huma Mughal[3], Mohd Murtadha Bin Mohamad[1]**

[1] Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia
[2] Forman Christian College (A Chartered university), Lahore, Pakistan
[3] Kinnaird College for Women, Lahore, Pakistan

* Correspondence author; E-mail: asmabasharat@fccollege.edu.pk.

**Abstract:** Imbalanced data in one of the major problems faced by Machine learning and deep learning classifiers. The skewness in the data distribution limits the performance of classifiers. This leads to overfitting of the model and misclassification for minority classes. Researchers have been focused on new techniques to balance data by oversampling minority classes, under sampling majority classes or creating a hybrid of oversampling and under sampling. Over the years researchers have also explored algorithmic techniques to adjust weights, create bags of classes and optimally enhance the data. This paper provides a state-of-the-art review of the latest contributions to resolve the imbalance data problem. The major focus of this paper is on the hybrid techniques, ensemble methods and GAN-based data augmentation techniques.

**Keywords:** Imbalance data; ensemble methods; data augmentation; generative adversarial networks

## 1. Introduction

The imbalance problems occur when data has skewed distributions means the data classes have an unequal number of instances. In some cases, imbalance is inherent to the nature of the problem but in others, the constraints during collection of data can result in imbalances. The machine learning approaches when applied to such problems fail to accurately identify instances of minority classes.

Studies suggest that it is a huge challenge to correctly classify the rare events in real world applications due to their low frequency of occurrence and high misclassification cost. Detecting rare events in datasets like intrusion detection and cancer patient involves binary classification of imbalance dataset, where intrusion and cancer is a rare event, and it represents the minority class of a dataset. The skewness of the dataset can be represented by the imbalance rate (IR). It is the ratio of instances of majority class to minority class. Addressing this challenging problem is significantly important for researchers to achieve

accurate classifications in various real-world applications. Generally, the approaches for resolving data imbalance can be broadly categorized into data level and algorithm level methods. The data level methods focus on updating the training dataset to resolve the imbalance issue. The data level techniques can be further divided into three categories: oversampling, under sampling and hybrid. In oversampling, synthetic samples of the minority class are generated from the existing instances to increase the size of minority classes like Synthetic Minority Oversampling Technique (SMOTE). However, to perform under-sampling, instances are removed from the majority class to balance the class sizes. The approaches like Easy Ensemble and Balance Cascade methods are also used for informative under-sampling to improve the performance of classification algorithms. In hybrid methods, both the oversampling and under sampling techniques can be combined to achieve better results.

Algorithm-level approaches handle the data imbalance by modifying the inference algorithm without altering the data distribution of the training dataset. It focuses on increasing the learning capability of algorithms to especially improve classification of minority classes. It includes a cost sensitive method in which higher cost is assigned to the misclassified minority class instances to increase their weight during the learning process. The imbalanced datasets have gained so much attention over the years due to its prevalence in real world scenarios. The conventional approaches discussed above have their own limitations. In oversampling, the dataset original distribution is altered and in case of under sampling there is a risk of losing the useful data. So, studies suggest that the ensemble techniques when combined with sampling techniques give more optimal results. It is important to understand that selection of specific ensemble and sampling techniques depends upon the characteristics of the dataset and the learning algorithm. The Hybrid methods combine data-level and algorithm-level approaches both to provide better classification results. In this case, a combination of methods including up-sampling, down-sampling, data augmentation, bagging and boosting is employed to cater the class imbalance. By combining the data level and algorithm approaches, hybrid techniques can overcome the limitations of individual methods. One of the major revolutionary techniques being explored since its inception is Generative Adversarial Networks (GAN). The GAN architecture adopts two competing Neural Networks to generate realistic data samples. Researchers are exploring new variants of GANs to augment data.

Our paper explores the latest contributions by researchers. This study is organized in three sections: the first section explores the Ensemble techniques: focusing on bagging as well as boosting methods. The second section reviews the latest research articles that propose hybrid techniques, including various ML algorithms. In the last section we will discuss data augmentation using GANs.

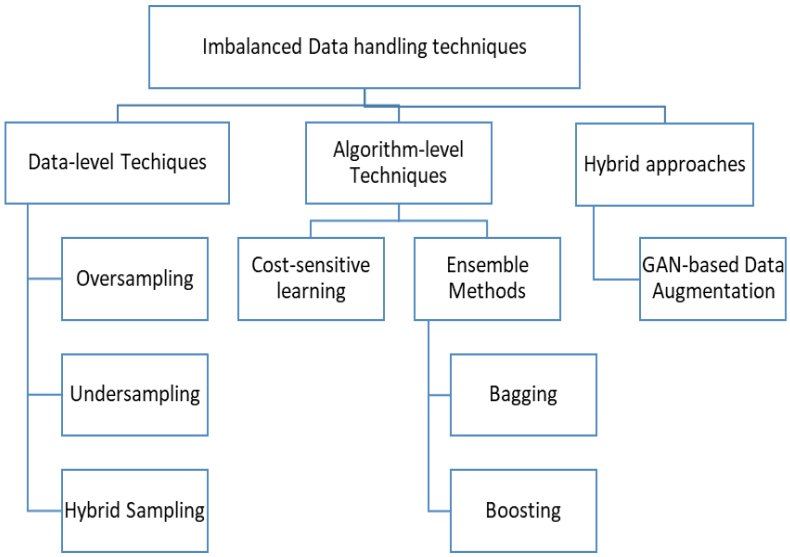


**Figure 1.** Categories of imbalanced data handling techniques.

## 2. Review of state-of-the-art imbalanced data handling techniques

Data imbalance problems in the real-world are unavoidable and persistent. Researchers have been actively finding new and improved solutions to handle dataset imbalance. Among the most recent approaches are the most investigated approaches are: Ensemble techniques, Hybrid techniques and data augmentation using GANs.

### *2.1. Ensemble techniques*

Ensemble classifiers use a combination of different machine learning classifiers to achieve the optimum result. The use of multiple classification techniques enables a reliable outcome by removing the bias of a single learner. The primary concepts behind developing an efficient ensemble learner are bagging, boosting, and stacking.

Bagging is the process of applying the same ML approach to different samples of the same dataset. Stacking on the other hand uses different ML models and tests the same dataset using different techniques. The result of bagging and stacking is the majority or average of all the outcomes. The most rapidly evolving technique of Boosting involves sequentially adding classifiers, where the selective outcome of the weak learner is used to develop a stronger learner to minimize the error rate. Ensemble techniques have been useful in removing the effects of dataset imbalance and have shown great promise to achieve remarkable accuracies.

#### 2.1.1 Bagging

Bagging techniques have shown promising results in combating overfitting and reducing variance in the dataset. Since bagging involves creating samples with replacement, each bootstrap may contain any number of the same instance. Prediction from the weak learners developed using imbalance data is then averaged out. This process continues till the best possible outcome is achieved. Hence the effect of data imbalance becomes insignificant.

Researchers have developed variants of bagging to improve the outcome of the bagging approach used to handle data imbalance. The authors of [10] proposed a model that used the bagging approach with the Xgboost (eXtreme Gradient Boosting) classifier along with the random under sampling technique with replacement to classify 133 records of radio broadcast based on the keyword "Beijing Time" in Mandarin. It was an imbalance dataset with 197 out of the 6906 samples containing the keyword. This approach has shown significant results in handling noisy data.

In this [8] a new bootstrapping approach has been presented to balance the class distribution of the bags using instance hardness. Over bagging is another method which is applied to the minority class where oversampling is used to increase the count of the minority class with respect to the majority class and hence the resulting classifier can be applied to the balanced dataset [15].[17] is research on 7 imbalance datasets obtained from the PROMISE repository. The authors propose an ensemble learning approach called Omni-Ensemble Learning (OEL). It uses the over-bagging technique for class imbalance. The effect of certain different methods on weight allocation to samples has been evaluated. In exactly balanced bagging technique, a random subset is selected from the majority class equal to the size of the minority class. The resulting sample has balanced cardinality of both classes [14]. SMOTE bagging uses the principle of SMOTE that generates synthetic instances for the minority class and combines it with the bagging approach. The new instances in the minority class are generated till they are equal to the majority class. The resulting bootstraps are balanced [12]. The author in [13] has proposed a technique using random undersampling and SMOTE bagging to resolve data imbalance problems in the Customer Churn dataset. it yielded excellent outcomes yielding higher F-scores compared to the non-bagging approach. The authors in [1] have explored the characteristics of the roughly balanced bagging approach for handling imbalance data. They have also analyzed and proved the superiority of performance of this ensemble learner in handling datasets with huge dimensionality and for multiclass imbalance problems.

### 2.1.2 Boosting

Boosting is a very popular approach in ensemble learning which works iteratively by giving more importance to difficult samples. The weights of misclassified samples by the base classifier are increased to improve the learning process. It follows a serial learning approach. In [11] a novel technique IMBoost is proposed which is a variant of AdaBoost. Initially, the weights of minority and majority classes are assigned based on their distribution. Then, weights are adjusted according to classifier performance on the classes individually. The dataset is then resampled based on these updated weights and this process is repeated. The proposed algorithm performs better than other ensemble methods when compared through Geometric mean (G-mean) and Area Under Curve (AUC). Gradient Boosting is a boosting algorithm which trains multiple weak classifiers like decision trees to make a robust classifier. Each new model or a classifier minimizes the loss function of the previous by using gradient

descent. The Gradient Boosting classifier is assembled in stages like other boosting techniques. It can accurately classify the imbalanced datasets [20].

### *2.2. Hybrid techniques*

Researchers have been exploring and modifying hybrid approaches to handle imbalanced data for decades. Hybrid techniques for handling imbalance include combining different sampling approaches, using sampling or weighted methods, utilizing cost sensitive learning, and applying algorithmic modifications according to the dataset. In this study, we focus on state-of-the-art research that demonstrates significant improvement in classification of imbalance problems. In 2020, [2] the authors proposed an improved deep learning technique for handling the imbalance problem in small sized datasets as the conventional machine learning techniques yield suboptimal results in imbalance datasets and classification results are biased towards the majority class. Furthermore, overfitting can occur due to the limited size of available training dataset. To address these issues of misclassification of minority class, the author has proposed an imbalanced multi-layer deep forest technique which is a variant of Deep Forest. Imbalanced Deep Forest (IMDF) is composed of multiple layers and each layer has multiple units. In each unit of IMDF, an improved AdaBoost (Adaptive Boosting) algorithm has been used to pay more attention to the minority classes by increasing their weights in each iteration. The synthetic samples are also generated by SMOTE for the minority class in each iteration to enhance the diversity of base classifiers and improve the learning on misclassified instances. The classification process is of cascaded nature. The tuning of hyper parameters and cascaded nature of the algorithm requires more computational time for higher dimensional and big datasets. To evaluate the performance of IMDF, F-value, Area under the Curve (AUC) and Recall are used. It is shown that IMDF gains the highest ranks against Decision Tree, Adaboost SMOTE, and Deep neural networks (DNN), on all the three metrics.

Robust hybrid data level methods are used in [4] to address several data impurities like noise, data imbalance and class overlapping. The proposed approach works in three steps. Firstly, removal of noise takes place. Then, it identifies majority and minority classes by using kernel-based fuzzy clustering. In the third phase, it reduces the size of the majority class by radial basis kernel fuzzy membership and α-cut. For minority class, firefly-based SMOTE method is used to synthesize the data points of the minority class to balance the class sizes. Now, the balanced classes are merged to form a data set which can then be used by traditional classifiers. According to results, this approach performs well on several conventional datasets, outperforming the other hybrid data level approaches in terms of Area under ROC curve.

**Table 1.** Imbalanced data handling literature review.

| ML technique | Dataset | Dataset description | Imbalance handling technique | Ref. no. |
|---|---|---|---|---|
| Decision tree, KNN, Neural Network, Random Forest, SVM | Open Corrosion dataset (OC), Colorectal cancer surgery (CCS) | Binary | SMOTE (oversampling) , ROSE (oversampling and under sampling) | [25] |
| Random Forest, Balanced RF, Modified Balanced RF | Customer churn data (PT Telekom Indonesia) | Binary | ensemble based learning approach | [24] |
| CART weighted Gini Index | stat log dataset, letter recognition dataset | Multiclass | embedded feature selection | [23] |
| EB-Bagging, RB-Bagging, SMOTE-bagging, RNB- bagging | 36 imbalanced datasets | Binary | ensemble based learning approach | [22] |
| IMDF {varaiant of Deep Forest}usesAdaBoost in each layer | Ecoli sick_euthyroid spectrometer | Binary | SMOTE | [2] |
| SVM, LR, DT, RF | Credit card dataset | Binary | SMOTE | [6] |
| K-means CTGAN | HomeEquity, Company Bankruptcy, Employee Promote and Online Shoppers Intention | Binary | SMOTE, GAN | [3] |
| CNN | Malicious code generation and detection | Multiclass | GAN | [16] |
| LR, RF, XGB and MLP | Credit card fraud dataset Pima diabetes and breast cancer dataset | Binary | SDG-GAN | [19] |
| Tabular data generators | Five datasets: Adult, Intrusion, loan, Cover type, credit | Multi-variable | CTAB-GAN | [21] |

In [5], the author addressed the challenges faced in handling real-world data as most of the data is generated from distributed autonomous sources and it is complex, heterogeneous, and voluminous. He has also discussed the techniques to handle such imbalanced datasets, especially in the context of the IoT framework. To perform experimentations, credit card, cancer patients and environment modeling imbalanced datasets have been used. In this paper, the author has compared the results using data level, algorithm, and hybrid level approaches. According to him, the accuracy of the predictive models for minority classes can be enhanced using hybrid models.

Fraud detection in credit cards is also a challenging domain as it can have high financial consequences. The paper referenced as [6] presents a hybrid approach where oversampling SMOTE technique has been combined with Logistic regression (LR), Decision Tree (DT), Random Forest (RF), and Support vector machines (SVMs). The experimental study has been carried out using an imbalanced credit card fraud dataset. The study has provided substantial evidence to support the claim that classification results drastically improve over imbalanced datasets after using the hybrid techniques and has showcased the superior performance of RF and DT than others. In another paper [7], experiments are carried out on telecom customer churn prediction dataset using RF, K- Neighbor (KN), LR, Multilayer Neural Network (MLP), Linear Support Vector Classifier (SVC) and Naïve Bayes algorithms. Random Forest approach outperforms in Precision, Accuracy, Recall and ROC-AUC score.

### *2.3. GAN-based data augmentation*

Recently Deep learning and its applications in computer vision have seen phenomenal growth. Its applications vary from medical image data, credit card frauds to malware detection. A plethora of data is generated and processed each day. This data has paved the way for extracting deep features and aid in efficient and improved classifications. However, one of the major limitations is the skewness of data distribution. The disparity in classes can vary from 100:1 to 10000:1 for majority and minority classes in binary labeled datasets. It poses an even greater problem for multi-labeled datasets, where multiple classes may have varying numbers of instances/samples. This disparity results in major biases for classifiers where some classes might overfit while some remain inaccurately classified. Generative Adversarial Networks have gained tremendous popularity since its inception. GAN is an unsupervised learning model that learns and discovers patterns from the original s neural networks: Generator (G) and Discriminator (D). The generator learns from the data and generates images (z) which are indistinguishable from the real sample (x) by the discriminator. Both networks are simultaneously trained.

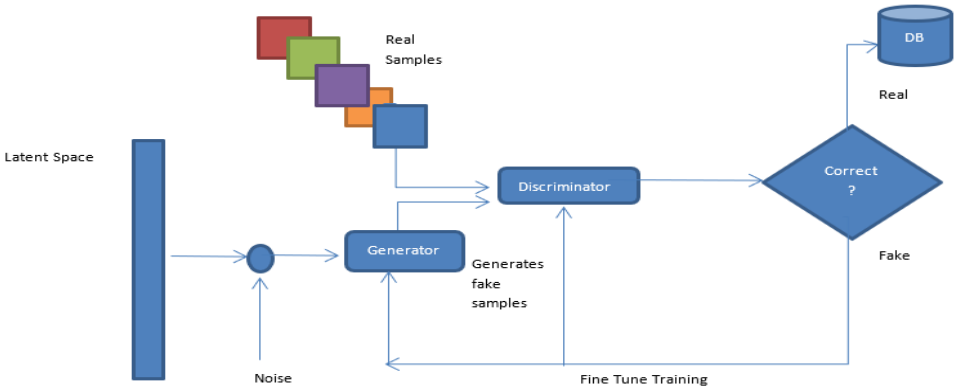


**Figure 2.** Data augmentation using GANs.

GANs have been widely used in recent years for data augmentation. It is an approach to increase the size of the real-world datasets by generation, translation, modification, or alteration of existing real-world samples. Increasing the size of minority data classes helps mitigate the overfitting problem [1]. A balanced dataset results in better accuracy and robustness of the trained model.

GANs can produce high quality synthetically generated data. GAN can learn hidden structures and features of data and reproduce them in such a way that the adversarial "discriminator" neural network is unable to distinguish between the real data sample and the generated data sample. The authors in [21] have used GANs to generate new malware samples to increase the robustness and accuracy of the proposed CNN model. Similarly, in [18], Burks *et al*. performed a comparative study between two generative methods, GANs and Variational AutoEncoders (VAE), in the study it was concluded that GANs outperformed VAEs. The synthetic training data produced by GANs enhanced the performance of Residual Network (ResNet-18) to detect malicious codes. In recent research [19], Charitou *et al*. proposed a Synthetic Data Generation GAN (SDG-GAN) to solve the imbalance issue in online real-world gambling fraud dataset. The synthetically augmented data showed better performance.

Over the years many variants of GANs have been introduced to overcome the issues of vanishing gradient in discriminator. In [3], the authors introduced Conditional Tabular - Generative Adversarial Network (CT-GAN) which can synthetically generate samples of minority classes. It has been shown that K means CTGAN outperforms previous methods that use SMOTE for handling imbalance. The performance of the algorithm has been compared with ADASYN, SMOTE, and K-means SMOTE; the results depict that K-means CT-GAN works on uniform distribution and hence avoids overfitting. The data is first classified into classes using K-means clusters. The minority class is identified, and data is augmented to balance the partitions. This technique demonstrates uniform distribution and overfitting is avoided.

In 2021, Zhao *et al* [21] propose a similar tabular data synthesizer based on GAN to effectively model continuous and categorical variables while preserving privacy. The technique was tested for five datasets with five ML algorithms; it showed promising results. Shamsolmoali *et al*. In [25] has proposed a hybrid technique that implements capsule network

architecture in the discriminator of Convolutional GAN. The proposed capsule-GAN architecture performs effectively in identifying overlapping classes with fewer parameters.

## 3. Conclusion

Researchers have been facing data imbalance issues in real-world problems. Especially now when technological advancements are growing at an unprecedented pace, Machine Learning and Computer Vision have seen tremendous application in this ever-evolving world. Data is being generated in immense volume and velocity. However, this data is not balanced. Some classes have a greater number of samples than the others. This poses the overfitting and misclassification problem. This study will provide potential future researchers with a consolidated research landscape in investigating solutions for handling imbalanced data. This study reviewed three primary techniques: ensemble techniques, hybrid techniques, and data augmentation using GANs.

## Conflicts of interest

The authors declare no conflict of interest.

## References

[1] Strelcenia E, Prakoonwit S. A New GAN-based data augmentation method for Handling Class Imbalance in Credit Card Fraud detection. In *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)* 2023, 627-634.

[2] Gao J, Liu K, Wang B, Wang D, Hong Q. An improved deep forest for alleviating the data imbalance problem. *Soft Comput*. 2021, 25:2085-101.

[3] An C, Sun J, Wang Y, Wei Q. A K-means Improved CTGAN Oversampling Method for Data Imbalance Problem. In *2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS)*, 2021, 883-887.

[4] Kaur P, Gosain A. Robust hybrid data-level sampling approach to handle imbalanced data during classification. *Soft Comput*. 2020, 24(20):15715-32.

[5] Mohindru G, Mondal K, Banka H. Different hybrid machine intelligence techniques for handling IoT-based imbalanced data. *CAAI Trans Intell Technol*. 2021, 6(4):405-16.

[6] Mqadi N, Naicker N, Adeliyi T. A SMOTe based oversampling data-point approach to solving the credit card data imbalance problem in financial fraud detection. *Int J Comput Digit Syst*. 2021, 10(1):277-86.

[7] Yadav S, Bhole GP. Handling imbalanced dataset classification in machine learning. In *2020 IEEE Pune Section International Conference (PuneCon)*, 2020, 38-43.

[8] Chongomweru H, Kasem A. A novel ensemble method for classification in imbalanced datasets using split balancing technique based on instance hardness (sBal_IH). *Neural Comput Appl*. 2021, 33(17):11233-54.

[9] Branco P, Torgo L, Ribeiro RP. Rebagg: Resampled bagging for imbalanced regression. In *Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 2018, 67-81.

[10] Ruisen L, Songyi D, Chen W, Peng C, Zuodong T, YanMei Y, Shixiong W. Bagging of xgboost classifiers with random under-sampling and tomek link for noisy label-imbalanced data. In *IOP Conference series: Materials science and engineering,* 2018, 428(1): 012004.

[11] Roshan S, Hallaji F, Ghanbari MR. Imboost: A New Weighting Factor for Boosting to Handle Imbalanced Problems. 2023.
[12] Iriawan N, Fithriasari K, Ulama BS, Suryaningtyas W, Pangastuti SS, *et al*. On The Comparison: Random Forest, SMOTE-Bagging, and Bernoulli Mixture to Classify Bidikmisi Dataset in East Java. In *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, 2018, 137-141.
[13] Hartati EP, Bijaksana MA. Handling imbalance data in churn prediction using combined SMOTE and RUS with bagging method. In *Journal of Physics: Conference Series*, 2018, 971(1): 012007.
[14] Vijayakumar M, Prabhakar E. A hybrid combined under-over sampling method for class imbalanced datasets. *Int J Res Adv Dev (IJRAD)*. 2018, 2(05):27-33.
[15] Arafat MY, Hoque S, Xu S, Farid DM. Machine learning for mining imbalanced data. 2019.
[16] Wang Z, Wang W, Yang Y, Han Z, Xu D, *et al.* CNN-and GAN-based classification of malicious code families: A code visualization approach. *Int J Intell Syst*. 2022, 37(12):12472-89.
[17] Mousavi R, Eftekhari M, Rahdari F. Omni-ensemble learning (OEL): utilizing over-bagging, static and dynamic ensemble selection approaches for software defect prediction. *Int J Artif Intell Tools*. 2018, 27(06):1850024.
[18] Burks R, Islam KA, Lu Y, Li J. Data augmentation with generative models for improved malware detection: A comparative study. In *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* 2019, 0660-0665.
[19] Charitou C, Dragicevic S, Garcez AD. Synthetic data generation for fraud detection using gans. *arXiv* preprint arXiv:2109.12546. 2021 Sep 26.
[20] Cahyana N, Khomsah S, Aribowo AS. Improving imbalanced dataset classification using oversampling and gradient boosting. In *2019 5th International Conference on Science in Information Technology (ICSITech)* 2019, 217-222.
[21] Zhao Z, Kunar A, Birke R, Chen LY. Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning* 2021, 97-112.
[22] Collell G, Prelec D, Patil KR. A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing*. 2018, 275:330-40.
[23] Liu H, Zhou M, Lu XS, Yao C. Weighted Gini index feature selection method for imbalanced data. In *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)* 2018, 1-6.
[24] Dwiyanti E, Adiwijaya, Ardiyanti A. Handling imbalanced data in churn prediction using rusboost and feature selection (case study: Pt. telekomunikasi indonesia regional 7). In *Recent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016), Bandung, Indonesia, August 18-20, 2016 Proceedings Second* 2017, 376-385.
[25] Chakravarthy AD, Bonthu S, Chen Z, Zhu Q. Predictive models with resampling: A comparative study of machine learning algorithms and their performances on handling imbalanced datasets. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)* 2019, 1492-1495.