A Comprehensive Study Of Classifying Imbalanced Data Methodologies

Dr. K.B. Jagadish Kumar[1], U. Mohan Srinivas[2], Dr. K. Navaz[3], M Akshitha Laasya[4], A. Ramesh[5]

[1,2,3,5] Department of Computer Science and Engineering
[4] Department of Information of Technology
[1,2,3, 5] QIS College of Engineering and Technology, Ongole, Andhra Pradesh, India
[4] CVR College of Engineering
[1]jagadeesh@qiscet.edu.in,[2]mohan.srinivas@qiscet.edu.in, [3]navaz.k@qiscet.edu.in
[4] 20B81A1206@cvr.ac.in , [5]ramesh.a@qiscet.edu.in
Corresponding Author Mail: qispublications@qiscet.edu.in

,

Article Info
Page Number: 76 - 89
Publication Issue:
Vol 68 No. 1 (2019)

Article History
Article Received:
09 September 2019
Revised: 16 October 2019
Accepted: 21 November 2019
Publication: 28 December 2019

Abstract

In many real-world datasets, class instances are distributed unevenly. In a Class Imbalance Problem (CIP), certain class have a much larger amount of instances than the others. Due to inaccurate predictions of the poor class data samples, the unbalanced data reduces the accuracy of the prediction. Data mining experts from many different fields are familiar with CIP. A major challenge in machine learning (ML) and deep learning is how to classify data that is not evenly distributed (DL). Given its importance, the employment of sample techniques for increasing the classifier performance has attracted considerable attention in the current literature. In this part, we discuss the value of data organisation and the approaches used by numerous researchers to level the playing field in imbalanced classrooms. Various classifiers' accuracy and prediction rates have been analysed, and the criteria used to evaluate them have been examined...

Keywords: Imbalanced data, Strong Class, Weak Class, Class Imbalance Problem, Classification

I INTRODUCTION

It's not simple to handle uneven data throughout the computation process. A powerful class is one that has more examples than any other class in the same dataset. If there are less data instances in one class than in another, for example, that class may be considered inferior. These datasets are called "unbalanced"[1].

Let's talk about how severe the problem of skewed data is: Think of a data set in which 99 percent of the samples come from good categories and 1 % come from bad classes. If the classifier accurately predicts that every piece of data belongs to a strong class, as shown in fig. 1, then the efficiency is at 100%. Since nothing is correctly classified by the low class samples, the result does not accurately reflect the forecast's accuracy.In addition, the categorization will not consider data from classes of low quality. It is necessary to distribute the asymmetrical data in order to improve the classifier's predictive power and get a more accurate result prediction. Real-world data classification may reveal a binary classification with severe class imbalance. Whether making a health prediction for yourself or a loved one, or trying to anticipate things like network penetration or financial crimes, it is important to be able to spot instances of the weak class. Under-representation in data sets is sometimes used to define a low class. This is a major roadblock, though, because most prediction algorithms give more weight to the strong class than the weak, resulting in wrong predictions..It's not simple to handle uneven data throughout the computation process. A powerful class is one that has more examples than any other class in the same dataset. If there are less data instances in one class than in another, for example, that class may be considered inferior. These datasets are called "unbalanced"[1].

Let's talk about how severe the problem of skewed data is: Think of a data set in which 99 percent of the samples come from good categories and 1 % come from bad classes. If the classifier accurately predicts that every piece of data belongs to a strong class, as shown in fig. 1, then the efficiency is at 100%. In addition, the categorization will not consider data from classes of low quality. It is necessary to distribute the asymmetrical data in order to improve the classifier's predictive power and get a more accurate result prediction. Real-world data classification may reveal a binary classification with severe class imbalance. Whether making a health prediction for yourself or a loved one, or trying to anticipate things like network penetration or financial crimes, it is important to be able to spot instances of the weak class. Under-representation in data sets is sometimes used to define a low class. The majority of forecasting models, however, prefer the powerful class over the poor, which causes forecasts to beerroneous...

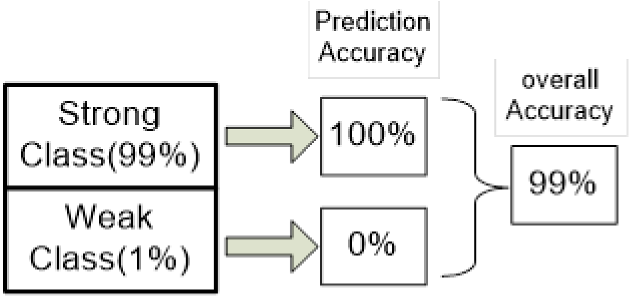

Fig. 1: Illustration of Class Imbalance Problem

Training and testing data shift due to class duplication, inconsistent testing dataset, small differences, inconsistent data, and the relevance of outlier samples [3]. In statistics, the term "class distribution" describes the breakdown of samples into distinct groups. When a model is trained using a dataset with instances that are not uniformly distributed across all class labels, a phenomenon known as class imbalance problem (CIP) occurs.

When the class distributions of the instances are uneven, a problem known as unbalanced classification arises in predictive analytics. Possible root reasons of the class disparity might be traced back to the way in which data was collected or to specific features of the domain itself. It's possible that the disparity across class instances resulted from the way the data was collected or sampled from the original issue. As a result, there is a chance of skewed results and other mistakes in the data gathering process. Human error is inevitable in data collecting.

One common error is to create numerous instances with incorrectly labelled classes. Errors or deficiencies in the methods used to collect samples might lead to unbalance.

In many cases, the discrepancy may be resolved by either improving data collection procedures or rectifying measurement errors. The uneven class distribution cannot be fixed by simply collecting more data from the domain. There has to be a model used to examine the differences across class instances. Problems with class inequality may also arise from measurement inconsistencies and from biassed sampling.

In a classification problem, an inequality between groups may be negligible or severe. In classification problems with mild imbalance, the distribution of examples in the dataset used to train a model may be somewhat unequal. In highly unbalanced classification problems, the distribution of examples in the dataset used to train a model may be extremely uneven. A minor discrepancy is not cause for alarm and is typically handled as a matter of predictive labelling. To simulate a significant socioeconomic gap often necessitates the use of sophisticated methods. It's possible that the strong classes contain a disproportionately high number of instances compared to the poor classes.

This allows the model deployment procedure to focus entirely on identifying the properties of the data while ignoring examples from the weak class and those that are more valuable for predictions. Many machine learning classification techniques are constructed on the assumption that each class has an equal number of examples, making imbalanced classification a barrier to accurate predictive modelling. That's why it's so hard to get accurate results from models involving the under-represented group. This is problematic because the weaker (minority) group is often more ignorant about the contributions of the more powerful (underrepresented) group.

When the distribution of the classes in the dataset used to build a model is unequal, a categorization technique called "imbalanced classification" is applied. However, it can be difficult to assess a significant imbalance without the aid of specialised techniques [6]. If the percentage of cases in each class is not consistent, the data is skewed in many classification tasks. If the distribution of the class data in the dataset is not nearly even, it is considered to be unbalanced. It's possible, for instance, for one class to be weak and have few instances, while another class has many. This imbalanced distribution of training data causes the classifier to favour robust classes. The prediction model's accuracy rate is impacted by the underrepresentation of weak classes, as demonstrated by numerous assessment techniques. Many methods are currently in use to deal with class imbalance issues, and this research will discuss many of them, such as ensemble methods, bagging, boosting, fuzzy classifiers, fuzzy sets, etc. Extensive studies are performed on how to improve learning from imbalanced data. Features of the CIP, advancements in technology, and current evaluation metrics used to gauge the success of learning via an unbalanced learning paradigm will all be subject to scrutiny. We also highlight key challenges, opportunities, and potentially relevant topic areas for learning from unbalanced data in the hopes that this will inspire future researchers to tackle these problems. Here we present the final parts of the project. The paper's second section analyses the literature concerning the Imbalance Classification of various authors' works, weighing their merits and demerits. Section III wraps up the discussion.Training and testing data shift due to class duplication, inconsistent testing dataset, small differences, inconsistent data, and the relevance of outlier samples [3]. In statistics, the term "class distribution" describes the breakdown of samples into distinct groups. When a model is trained using a dataset with instances that are not uniformly distributed across all class labels, a phenomenon known as class imbalance problem (CIP) occurs.

When the class distributions of the instances are uneven, a problem known as unbalanced classification arises in predictive analytics. Possible root reasons of the class disparity might be traced back to the way in which data was collected or to specific features of the domain itself. It's possible that the disparity across class instances resulted from the way the data was collected or sampled from the original issue. As a result, there is a chance of skewed results and other mistakes in the data gathering process. Human error is inevitable in data collecting. One common error is to create numerous instances with incorrectly labelled classes. Errors or deficiencies in the methods used to collect samples might lead to unbalance.

In many cases, the discrepancy may be resolved by either improving data collection procedures or rectifying measurement errors. The uneven class distribution cannot be fixed by simply collecting more data from the domain. There has to be a model used to examine the differences across class instances. Problems with class inequality may also arise from measurement inconsistencies and from biassed sampling.

In a classification problem, an inequality between groups may be negligible or severe. An extreme class imbalance in a classification problem may involve tens of billions of instances in one class and millions of instances in another class for the same training dataset. In classification problems with mild imbalance, the distribution of examples in the dataset used to train a model may be somewhat unequal. In highly unbalanced classification problems, the distribution of examples in the dataset used to train a model may be extremely uneven. A minor discrepancy is not cause for alarm and is typically handled as a matter of predictive labelling. To simulate a significant socioeconomic gap often necessitates the use of sophisticated methods. It's possible that the strong classes contain a disproportionately high number of instances compared to the poor classes.

The classes are assumed to be evenly distributed by the majority of machine learning (ML) classification prediction algorithms (5). This implies that samples from the weak class and those that are better for prediction may be excluded during model deployment so that the process may concentrate solely on identifying the attributes of the data.Due to the fact that many machine learning classification approaches are built on the assumption that each class contains an equal number of samples, unbalanced classification is a hindrance to precise predictive modelling. As a consequence, it is challenging to get reliable findings from models that include the under-represented population. This is a concern since members of the marginalised group (the minority) are frequently less familiar with the achievements of the dominant group (the underrepresented group

When the classes in the dataset used to build a model are unevenly distributed, a classification method known as "imbalanced classification" is applied. However, it can be difficult to assess a significant imbalance without the aid of specialised techniques [6In many classification, the data is skewed if the percentage of cases in each class is not constant. The dataset is deemed imbalanced if the distribution of the studied groups is not nearly even.. It's possible, for instance, for one class to be weak and have few instances, while another class has many. This imbalanced distribution of training data causes the classifier to favour robust classes. The prediction model's accuracy rate is impacted by the underrepresentation of weak classes, as demonstrated by numerous assessment techniques. Many methods are currently in use to deal with class imbalance issues, and this research will discuss many of them, such as ensemble methods, bagging, boosting, fuzzy classifiers, fuzzy sets, etc. Extensive studies are performed on how to improve learning from imbalanced data. Features of the CIP, advancements in technology, and current evaluation metrics used to gauge the success of learning via an unbalanced learning paradigm will all be subject to scrutiny. We also highlight key challenges, opportunities, and potentially relevant topic areas for learning from unbalanced data in the hopes that this will inspire future researchers to tackle these problems. Here we present the final parts of the project. The paper's second section analyses the literature concerning the ImbalanceClassification of various authors' works, weighing their merits and demerits. Section III wraps up the discussion...

II.**RELATED WORK**

There has been some discussion of using modelling techniques and sampling strategies to eradicate CIP. Applying sampling methods will get rid of the asymmetrical nature of the datasets. To establish a more even distribution of classes, over-sampling involves adding more data instances or modified data instances to the weak class (OS). Under-sampling is when information is deliberately left out of a powerful group in order to make the weaker groups seem more numerous (US). Many scholars, professionals, and researchers have conducted exploratory studies, evaluations, and investigations on the topics of class imbalance.

According to [7], In order to get representative results from extremely imbalanced data sets, SMOTE (Synthetic Minority Over-sampling Technique) used boosting and bootstrap aggregating methods. Their primary focus was on problems of binary classification, in which information was divided into two groups based on the presence or absence of a single label. SMOTE has seen widespread application for its ability to rapidly produce new instances, learn from imbalanced datasets, and build new instances. Initial investigation revealed that when compared to current ensemble technologies, implementing the classification tree, SMOTE, Bootstrap aggregating, and boosting approaches may lead to improved performance. S. V. Spelmen, et al (2018). [9] Clarified why addressing CIP is very important. Correctness and prediction rates of classifiers were analysed, along with the strategies for defining the kind of unbalanced data and assessment methodology offered by the different experts.

By F. Shakeel, et al (2017). When evaluating the skewed data sets, [10] they looked at every technique that had been used. The computational procedures utilised in the first phases of processing are also examined. Compound and ensemble methods have also been studied.

Additionally, two different types of skewed data sets are explored for binary and multi-class data distributions.

As an illustration, Wang, Le, and others (2021). Cost-sensitivity, feature depth, deep learning methods such as Convolution Neural Network (CNN), Neural Learning, DL techniques, sampling techniques such as SMOTE, SVM, and k-nearest neighbour (KNN), learning algorithms such as single class or EnsembleLearning (EL), and evaluation criteria such as Receiver Operating Characteristic Curve (ROC), F-Measure, and recall were all taken into account in [11].

Predicted forest fires employing a variety of mining techniques, including RF, BPNN, and SVM, are shown in [8]. They used three different methods of data balancing to ensure that everything was balanced. One method is called "random under-sampling," and it entails arbitrarily excluding or replacing samples from the "strong class" in the training dataset. Samples will be added and subtracted at random until a uniform distribution of values is achieved. Two further methods exist for achieving this balance: (2) SMOTE, which involves duplicating instances in the weak class, and (3) Easy Ensemble (EE), which involves deliberately undersampling the strong class in order to get a more equitable distribution of data. The results of the exploratory study suggest that EE is superior than the other two methods of data balancing in terms of effectiveness. Predictions of forest fire occurrences employing mining techniques like RF, BPNN, and SVM achieved an excellent 0.97 ROC Curve Area thanks to the EE model's standardisation. They used three different methods for achieving data parity. The first is a technique known as "random under-sampling," which involves selecting and excluding samples at random from the strong class of the training dataset. Samples will be added and discarded at random until a uniform distribution of values is achieved. (3) Easy Ensemble (EE): balancing data subsets by arbitrarily undersampling the strong class; (2) SMOTE: duplicating instances in the weak class. The results of the exploratory investigation suggest that EE outperforms the other two methods of data balancing. The 0.97 ROC Curve Area of the EE model shows high standardisation.

Collectively, X. Y. Jing and coworkers (2021). [12] It was determined that UCML, or uncorrelated cost-sensitive multi-set learning, was the most effective strategy for resolving CIP. Experiments conducted on eight expert-level and two large-scale imbalanced datasets revealed that UCML outperformed state-of-the-art methods and could maintain high imbalance Ratios. Figure 2 shows the many possible causes of social stratification. Classification approaches for asymmetric data sets are summarised in Table 1.

Research on CIP classification algorithms that is now available often begins with data-level techniques, which are often split into two camps: US and OS. From an EL point of view, there are additional options for classifying data that is not evenly distributed. It is now possible to use a variety of complex CIP techniques, from the simplest US and OS procedures to real-valued unfavorable-favored-operating-systems. To further enhance the outcomes of the DL approach, researchers are now combining ensemble methods with Neural Networks (NN). In recent years, the unsupervised learning method based on GANs has been widely used for the categorization of imbalanced data sets. Classification methods that rely on attribute level and sensitive cost functions are widely employed to overcome issues with imbalanced data sets, such as CIP. Data-level techniques, which may be further categorised into US and OS categories, provide the basis for many of the currently available research on CIP classification algorithms. From an EL point of view, there are additional options for classifying data that is not evenly distributed. A wide range of advanced CIP techniques, from the simplest US and OS procedures to real-valued unfavourable favoured OS, are now available. To further enhance the outcomes of the DL approach, researchers are now combining ensemble methods with Neural Networks (NN). GAN-based unsupervised learning is becoming more popular for the categorization of imbalanced data sets. Classification methods that rely on attribute level and sensitive cost functions are widely employed to overcome issues with imbalanced data sets like CIP.

It has never been more crucial to have a robust foundation in original data interpretation, analysis, and exploration to support key choices due to the expansion of digital access to more complex and time-sensitive systems including surveillance, healthcare, the Internet, and banking. Despite the fact that current information retrieval and predictive analytics systems have demonstrated remarkable success in a number of practical applications, the class imbalance issue (the unbalanced learning problem) is a new demanding endeavour for researchers and academics. A fundamental interpretation of original data, a thorough examination of original data, and exploration from original data are crucial for supporting decision-making in light of the ongoing improvements in digital accessibility in many large, sophisticated, and real-time systems like surveillance, health, the Internet, and banking. Addressing the class imbalance issue (the imbalanced learning problem) is a new challenging challenge for academics and researchers, despite the fact that modern information retrieval and predictive analytics systems have shown great achievement in several practical fields..

The imbalanced learning issue has many factors: the presence of the weak class, the degree of class imbalance, and the classifier's slant toward the strong class. Learning from biassed and distorted data sets requires cutting-edge ideas, concepts, approaches, and procedures in order to effectively turn massive volumes of underlying data into usable data, knowledge, intelligence, and data analytics. This is due to the fact that skewed and skewed data sets contain intricate qualities that make them challenging to use.

Table 1:study on the classification techniques for unbalanced datasets

Study	Advantages	Disadvantages
[13]	For unbalanced Phosphonic databases, unbalanced information may be identified and detected with great efficiency and consistency.	Investigating this issue in depth is crucial for finding solutions to similar problems in biochemical studies, biology,and BDA
[14]	Enhanced F1-minus scores	The Lemon Shark

	for poor categories in the unbalanced ADL data set.	datasets we were able to get did not have any helpful time or Wavelet transform.
[15]	Top-1 reliability was improved by 4.39 percentage points on the EMNIST and 6.51% on the CINIC-10 datasets, respectively, while overall reliability was reduced by 52.0% and 46.9%, respectively..	Results may be misleading if reliability is the sole criterion used for assessment..
[16]	By minimising redundant instances in the top four credit card datasets from the UCI ML Repository, sparse extracted features was shown to improve classifier effectiveness using the F-measure and G-mean assessment metrics. The feature-selection algorithm might be improved to handle CIP and could be applied in other fascinating fields like medical, business, economics, and the hard sciences..	In the long term, additional unbalanced statistics will require verification.
[17]	New, high-quality examples in low-confidence classes may be provided for fixing problems with coils and gearboxes.	There is no consideration given to the construction and use of GAN into detection methods for defect diagnosis through the adversarial learning experience.
[18]	By creating near-identical copies of low-quality class data, this method restores parity across classes and, in	Additionally, even with an unbalanced data, the trained model may be

	the end, increases the likelihood of overfitting.	properly trained for sampling from multiple classes..
[19]	Finding the ideal sample rate for each categorization using exhaustive search is fast and accurate.	As the number of characteristics, the procedure becomes more laborious.
[20]	Overall accuracy increased by 2% to 10%, and the distribution was fairly even in the categorization..	It is constructed using just a small learning set and no stringent adjusting parameters..
[21]	The classifier's precision on the CHB MIT [22] and iNeuro sets has been enhanced.	Data sets that have the potential to develop issues.
[23]	Adaptable to collections with varying degrees of mismatch.	The research focuses only on problems that can be sorted into two categories. In software development, however, there are many situations involving over than 2 class labels.
[24]	The effects of income inequality were shown using actual information consisting of over 75 million card transactions collected from the online purchases of European cardholders.	For feedback- and delayed-supervised sampling-based systems, there was no discussion of flexible and, at times, unexpected collecting strategies..
[25]	Their research showed that when data is highly skewed, unbalanced classification systems perform poorly. Their findings demonstrate that the present methods produce a high rate of	There has been no research on the CIP that determines whether or not sensitivity or accuracy must be sacrificed..

	inaccurate reports, which is costly for enterprises and leads to an inaccurate forecast.	
[26]	It was found via testing on European card payment datasets that the integrated technique is superior for CIP.	This tactic cannot be used in an unattended setting.
[27]	Studies show that GAN may improve ML accuracy by 27%.	Research that focuses on only one metric for asymmetrical knowledge is misleading.
[28]	The results of the experiments show that RCT achieves the highest efficiency and improves the program's predicting quality by up to as 15.2 percentage points and 29.9 percentage points, respectively, when compared to standard class overlapping cleaning procedures..	The cost-learning method would make this more efficient..
[29]	By using a graph-based technique that simultaneously determines quantified labels and trains the induction model, they were able to solve the CIP.	The classifier's effectiveness may suffer when presented with brand-new, false data..

III. CONCLUSION

CIP has a wide range of real-world uses, including detecting pollution, detecting biomedical fraud, detecting credit card fraud, detecting phone fraud, detecting clinical therapeutics, detecting network infiltration, detecting computational biology, detecting oil spills using satellite radar, identifying fake phone calls, predicting software defects, and articulating words. An organization's attrition rate may be predicted using fraud detection. Today's world is rife with unbalanced data sets, making this study all the more important. The purpose of this article is to discuss the best practises for identifying and labelling biassed data sets. There are a variety of methods used in order to categorise data sets that are not evenly distributed. CIP has many real-world uses, including but not limited to: oil spill detection via satellite radar; biomedical fraud detection; credit card fraud detection; phone call fraud detection; clinical therapeutics; network infiltration detection; computational biology; text categorization; identification of fake phone calls; prediction of software defects; word-learn articulation; and so on. Predictive categorization model errors might harm CIP in a number of different ways. The issues that firms encounter include predicting insurance claims, default, attrition rates, and conversions, to name a few. Because incomplete and inconsistent data sets are so common in today's world, this study is extremely important. This article discusses many approaches of classifying imbalanced data sets as a solution to this issue.

REFERENCES

[1] H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

[2] Fan, Yang, Zheng Kai, and Li Qiang. "A revisit to the class imbalance learning with linear support vector machine." 2014 9th International Conference on Computer Science & Education. IEEE, 2014.

[3] Jabbari, Fattaneh, Mohammad H. Falakmasir, and Kevin D. Ashley. "Identifying thesis statements in student essays: The class imbalance challenge and resolution." The Twenty-Ninth International Flairs Conference. 2016.

[4] Al-Roby, Marwa F., and Alaa M. El-Halees. "Classifying Muti-Class Imbalance Data." Egyptian Computer Science Journal 37.5 (2013): 74-81.

[5] Kumar, P. Ganesh, and J. Briso Becky Bell. "Using Continuous Feature Selection Metrics to Suppress the Class Imbalance Problem." Int. J. Sci. Eng. Res. 3 (2012): 1-9.

[6] Sikora, Riyaz, and Sahil Raina. "Controlled under-sampling with majority voting ensemble learning for class imbalance problem." Science and Information Conference. Springer, Cham, 2018.

[7] K. UlagaPriya and S. Pushpa, "A Comprehensive Study on Ensemble-Based Imbalanced Data Classification Methods for Bankruptcy Data," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 800-804, doi: 10.1109/ICICT50816.2021.9358744.

[8] W. Zhou, W. Chen, E. Zhou, Y. Huang, R. Wei and Y. Zhou, "Prediction of Wildfire-induced Trips of Overhead Transmission Line based on data mining," 2020 IEEE International Conference on High Voltage Engineering and Application (ICHVE), 2020, pp. 1-4, doi: 10.1109/ICHVE49031.2020.9279835.

[9] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018, pp. 1-11, doi: 10.1109/ICCTCT.2018.8551020.

[10] F. Shakeel, A. S. Sabhitha and S. Sharma, "Exploratory review on class imbalance problem: An overview," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, pp. 1-8, doi: 10.1109/ICCCNT.2017.8204150.

[11] Wang, Le, et al. "Review of Classification Methods on Unbalanced Data Sets." IEEE Access 9 (2021): 64606-64628.

[12] X. -Y. Jing et al., "Multiset Feature Learning for Highly Imbalanced Data Classification," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 139-156, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2929166.

[13] Hussin, Sahar K., et al. "Handling Imbalance Classification Virtual Screening Big Data Using Machine Learning Algorithms." Complexity 2021 (2021).

[14] Y. Yang, H. -G. Yeh, W. Zhang, C. J. Lee, E. N. Meese and C. G. Lowe, "Feature Extraction, Selection, and K-Nearest Neighbors Algorithm for Shark Behavior Classification Based on Imbalanced Dataset," in IEEE Sensors Journal, vol. 21, no. 5, pp. 6429-6439, 1 March1, 2021, doi: 10.1109/JSEN.2020.3038660.

[15] M. Duan, D. Liu, X. Chen, R. Liu, Y. Tan and L. Liang, "Self-Balancing Federated Learning With Global Imbalanced Data in Mobile Systems," in IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 1, pp. 59-71, 1 Jan. 2021, doi: 10.1109/TPDS.2020.3009406.

[16] E. B. Fatima, B. Omar, E. M. Abdelmajid, F. Rustam, A. Mehmood and G. S. Choi, "Minimizing the Overlapping Degree to Improve Class-Imbalanced Learning Under Sparse Feature Selection: Application to Fraud Detection," in IEEE Access, vol. 9, pp. 28101-28110, 2021, doi: 10.1109/ACCESS.2021.3056285.

[17] Z. Li, T. Zheng, Y. Wang, Z. Cao, Z. Guo and H. Fu, "A Novel Method for Imbalanced Fault Diagnosis of Rotating Machinery Based on Generative Adversarial Networks," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-17, 2021, Art no. 3500417, doi: 10.1109/TIM.2020.3009343.

[18] R. Low, L. Cheah and L. You, "Commercial Vehicle Activity Prediction With Imbalanced Class Distribution Using a Hybrid Sampling and Gradient Boosting Approach," in IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 3, pp. 1401-1410, March 2021, doi: 10.1109/TITS.2020.2970229.

[19] J. Yao, Y. Zheng and H. Jiang, "An Ensemble Model for Fake Online Review Detection Based on Data Resampling, Feature Pruning, and Parameter Optimization," in IEEE Access, vol. 9, pp. 16914-16927, 2021, doi: 10.1109/ACCESS.2021.3051174.

[20] Z. Lv, G. Li, Z. Jin, J. A. Benediktsson and G. M. Foody, "Iterative Training Sample Expansion to Increase and Balance the Accuracy of Land Classification From VHR Imagery," in IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 1, pp. 139-150, Jan. 2021, doi: 10.1109/TGRS.2020.2996064.

[21] D. Hu, J. Cao, X. Lai, J. Liu, S. Wang and Y. Ding, "Epileptic Signal Classification Based on Synthetic Minority Oversampling and Blending Algorithm," in IEEE Transactions on Cognitive and Developmental Systems, vol. 13, no. 2, pp. 368-382, June 2021, doi: 10.1109/TCDS.2020.3009020.

[22] Ali Shoeb. Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment. PhD Thesis, Massachusetts Institute of Technology, September 2009.

[23] J. Zheng, X. Wang, D. Wei, B. Chen and Y. Shao, "A Novel Imbalanced Ensemble Learning in Software Defect Predication," in IEEE Access, vol. 9, pp. 86855-86868, 2021, doi: 10.1109/ACCESS.2021.3072682.

[24] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE transactions on neural networks and learning systems, 29(8), 3784-3797.

[25] Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., & Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access, 7, 93010-93022.

[26] Tingfei, H., Guangquan, C., & Kuihua, H. (2020). Using variational auto encoding in credit card fraud detection. IEEE Access, 8, 149841-149853.

[27] M. A. Aydin, "Using Generative Adversarial Networks for Handling Class Imbalance Problem," 2021 29th Signal Processing and Communications Applications Conference (SIU), 2021, pp. 1-4, doi: 10.1109/SIU53274.2021.9477939.

[28] S. Feng, J. Keung, J. Liu, Y. Xiao, X. Yu and M. Zhang, "ROCT: Radius-based Class Overlap Cleaning Technique to Alleviate the Class Overlap Problem in Software Defect Prediction," 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), 2021, pp. 228-237, doi: 10.1109/COMPSAC51774.2021.00041.

[29] G. Du et al., "Graph-Based Class-Imbalance Learning With Label Enhancement," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2021.3133262.

[30] P Ramprakash, M Sakthivadivel, N Krishnaraj, J Ramprasath. ”Host-based Intrusion Detection System using Sequence of System Calls” International Journal of Engineering and Management Research, Vandana Publications, Volume 4, Issue 2, 241-247, 2014

[31] N Krishnaraj, S Smys.”A multihoming ACO-MDV routing for maximum power efficiency in an IoT environment” Wireless Personal Communications 109 (1), 243-256, 2019.

[32] N Krishnaraj, R Bhuvanesh Kumar, D Rajeshwar, T Sanjay Kumar, Implementation of energy aware modified distance vector routing protocol for energy efficiency in wireless sensor networks, 2020 International Conference on Inventive Computation Technologies (ICICT),201-204

[33] Ibrahim, S. Jafar Ali, and M. Thangamani. "Enhanced singular value decomposition for prediction of drugs and diseases with hepatocellular carcinoma based on multi-source bat algorithm based random walk." Measurement 141 (2019): 176-183. https://doi.org/10.1016/j.measurement.2019.02.056

[34] Ibrahim, Jafar Ali S., S. Rajasekar, Varsha, M. Karunakaran, K. Kasirajan, Kalyan NS Chakravarthy, V. Kumar, and K. J. Kaur. "Recent advances in performance and effect of Zr doping with ZnO thin film sensor in ammonia vapour sensing." GLOBAL NEST JOURNAL 23, no. 4 (2021): 526-531. https://doi.org/10.30955/gnj.004020 , https://journal.gnest.org/publication/gnest_04020

[35] N.S. Kalyan Chakravarthy, B. Karthikeyan, K. Alhaf Malik, D.Bujji Babbu,. K. Nithya S.Jafar Ali Ibrahim , Survey of Cooperative Routing Algorithms in Wireless Sensor Networks, Journal of Annals of the Romanian Society for Cell Biology ,5316-5320, 2021

[36] Rajmohan, G, Chinnappan, CV, John William, AD, Chandrakrishan Balakrishnan, S, Anand Muthu, B, Manogaran, G. Revamping land coverage analysis using aerial satellite image mapping. Trans Emerging Tel Tech. 2021; 32:e3927. https://doi.org/10.1002/ett.3927

[37] Vignesh, C.C., Sivaparthipan, C.B., Daniel, J.A. et al. Adjacent Node based Energetic Association Factor Routing Protocol in Wireless Sensor Networks. Wireless Pers Commun 119, 3255–3270 (2021). https://doi.org/10.1007/s11277-021-08397-0.

[38] C Chandru Vignesh, S Karthik, Predicting the position of adjacent nodes with QoS in mobile ad hoc networks, Journal of Multimedia Tools and Applications, Springer US,Vol 79, 8445-8457,2020