ISSN 2347 - 3983
Volume 9. No.10, October 2021
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter089102021.pdf
https://doi.org/10.30534/ijeter/2021/089102021

# A Survey on Imbalanced Data Handling Techniques for Classification

**Abhisar Sharma[1], Dr. Anuradha Purohit[2], Himani Mishra[3]**
[1]M.E. Comp. Engg., Dept. of Comp. Engg., S.G.S.I.T.S., Indore, India, abhisars09@gmail.com
[2]Associate Professor, Dept. of Comp. Engg., S.G.S.I.T.S., Indore, India, apsgsits@gmail.com
[3]Assistant Professor, Dept. of Comp. Engg., S.G.S.I.T.S., Indore, India, himanimishra.hm21@gmail.com

Received Date : September 02, 2021 Accepted Date : September 23, 2021 Published Date : October 07, 2021

**ABSTRACT**

Classification is a supervised learning task based on categorizing things in groups on the basis of class labels. Algorithms are trained with labeled datasets for accomplishing the task of classification. In the process of classification, datasets plays an important role. If in a dataset, instances of one label/class (majority class) are much more than instances of another label/class (minority class), such that it becomes hard to understand and learn characteristics of minority class for a classifier, such dataset is termed an imbalanced dataset. These types of datasets raise the problem of biased prediction or misclassification in the real world, as models based on such datasets may give very high accuracy during training, but as not familiar with minority class instances, would not be able to predict minority class and thus fails poorly. A survey on various techniques proposed by the researchers for handling imbalanced data has been presented and a comparison of the techniques based on f-measure has been identified and discussed.

**Key words:** Imbalanced dataset, random under-sampling, random over-sampling, SMOTE.

## 1. INTRODUCTION

In classification, the value of a categorical attribute (class) is predicted based on the values of other attributes (predicting attributes) [1]. The task of classification is accomplished by training the classifier. This is done by splitting data into two sets termed as training set and test set. Classification methods assume that class probabilitydistribution is of high entropy [2] in the training dataset. This assumption is not always valid for many real-world applications from medical diagnosis, fraud detection, information retrieval, and so on. In training data, if there is a much lower numberof instances of one class, then the assumed priority distribution for classification will be hindered and this classification paradigm will be termed as imbalance classification [2][24].

Imbalanced classification in terms of the dataset can be understood as when there is the dominance of one or many classes (in a multi-class dataset) in the training dataset in such a manner that it becomes hard for the classifier to understand the characteristics of another class, which results in biased prediction towards dominant class such datasets having a skewed distribution of class instances are termed as imbalanced datasets. This form of imbalance is referred to as between-class imbalance, not uncommon are between class imbalances on the order of 1:100, 1:1000, 1:10,000 are very common wherein each class severely out represents another [3][6]. Between-class imbalance is innately binary. There are two types of between-class imbalances. One is termed as intrinsic, i.e., the imbalance is a direct result of the nature of the dataspace. Variable factors such as time and storage also give rise to imbalance, this type is considered as extrinsic i.e., the imbalance is not directly related to the nature of dataspace [6].

Another type of imbalance occurs due to rare instances, i.e., minority class instances are very limited in number; the target class is rare. Within-class imbalance concerns itself with the distribution of representative data for sub concepts within a class. Figure 1 highlights types of imbalance problems [6].

The primary goal of any classifier is to reduce itsclassification error and maximize its overall accuracy [3]. To accomplish this, classifiers working on imbalanced datasets need to learn equally with minority and majority data instances. Imbalanced datasets lack minority class instances and this

makes classifiers treat minority class instances as outliers or noise.

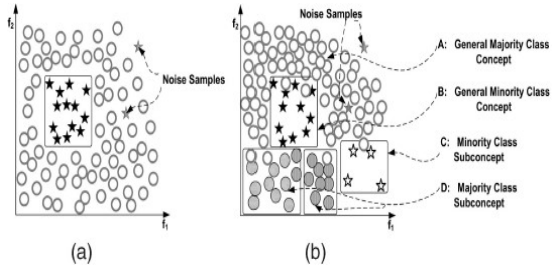


**Figure 1:** (a) A knowledge set with a between-class imbalance. (b) A high-complexity data set with both between-class and within-class imbalances, multiple concepts, overlapping, noise, and lack representative data.

The primary goal of any classifier is to reduce itsclassification error and maximize its overall accuracy [3]. To accomplish this, classifiers working on imbalanced datasets need to learn equally with minority and majority data instances. Imbalanced datasets lack minority class instances and this makes classifiers treat minority class instances as outliers or noise.

Most researchers have strained their attention on synthetic oversampling techniques in which they try to generate minority class instances synthetically so that the imbalance ratio can be normalized or the classifier generalizes well [2][3][6].

The survey presented in this paper is organized into five sections. Section 1 introduces problems that originated because of the imbalanced dataset. Section 2 discusses imbalanced datasets in greater detail and states various approaches used to handle imbalanced dataset problems. Section 3 talks about various approaches for handling imbalanced datasets. Section 4 highlights various research outcomes for f-measure values. Section 5 completes the conclusion of this survey.

## 2. BACKGROUND

Imbalance data possess a major problem of misclassification. As majority class samples are much more in number and minority samples lack behind, classifiers are unable to understand characteristics of minority class and generally treat them as noise or outliers.

The models which are trained with imbalance datasets may possess higher accuracy at the time of training but fail to predict true values and give prediction only for one class, such models fail in real-world performance despite higher accuracy during training.

To understand the implications of the imbalanced learning problem in the real world, let us assume an example from biomedical applications.
Consider the "Mammography Data Set," a collection of images acquired from a series of mammography exams performed on a set of distinct patients, which has been widely used in the analysis of algorithms addressing the imbalanced learning problem [6][19-21].

Positive and Negative are two binary classes for an image representative of a "cancerous" or "healthy" patient, respectively. The data set contains 10,923 Negative (majority class) samples and 260 Positive (minority class) samples. A classifier is required for the 100 percent predictive accuracy of both classes. In reality, we discover that classifiers tend to supply a severely imbalanced degree of accuracy, with the bulk class having on the brink of 100 percent accuracy and therefore the minority class having accuracies of 0-10 percent, for instance [6][19] [21].

If 10 percent of accuracy is achieved on minority class by a classifier, then it would suggest that 234 minority samples are misclassified as majority samples. The consequence of this is equivalent to 234 cancerous patients classified (diagnosed) as non-cancerous. This misclassification can lead to severe ramifications. Therefore, it is evident that for this domain, we require a classifier that will provide high accuracy for the minority class without severely jeopardizing the accuracy of the majority class. Imbalance data handling approaches such as data-level approaches, algorithmic approaches, hybrid approaches, kernel-based approaches, and cost-sensitive approaches are put forwarded [3][6-8].

A survey on imbalanced data handling techniques is presented in this paper along with the comparison of f-measure values for different techniques used for handling imbalanced data.

## 3. APPROACHES FOR HANDLING IMBALANCED DATA SETS

The As per the study presented by Galar et al. [8], and H. He. et al [6], we can categorize imbalanced data handling approaches as data-level approaches, algorithmic approaches, hybrid approaches, Kernel-based approaches, and Cost-based approaches. Hybrid approaches are considered as a

combination of data-level approaches and algorithmic approaches and are termed as ensemble approaches. P. Lim [7] had stated that SMOTEBagging, SMOTEBoosting, UnderBagging, and RUSBoost incorporated with SMOTE and undersampling into respective bagging and boosting framework and worked as ensemble techniques. Data-level approaches, and hybrid approaches are also termed as external approaches [25].

A discussion on various existing approaches for handling imbalanced datasets are discussed in the subsequent section.

### 3.1 DATA LEVEL APPROACHES

Data-level approaches are also termed as sampling-based approaches as the modification is done in data of dataset or we can say that in samples of dataset. In this technique, data samples are manipulated by various means for obtaining desired results. Various sampling-based approaches are depicted in Figure 2.

**3.1.1 Random Undersampling (RUS):** In the undersampling majority of class instances are deleted randomly for making a balanced dataset. Near Miss undersampling is proposed by J. Zhang et al. [9], which refers to a collection of undersampling methods that select examples based on the distance of majority class examples to minority class examples. Distance is determined in feature space usingEuclidean distance. Condensed Nearest Neighbor (CNN) Rule is an undersampling method proposed by P. Hart [10]. CNN technique seeks a subset of a set of samples from the bulk class that end in no loss in model performance, mentioned to as a minimal consistent set.

In random undersampling, potentially important data may get deleted and this can affect the training process.

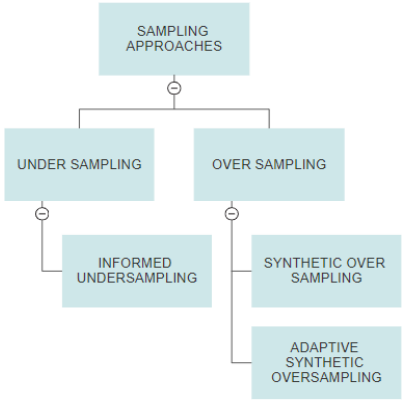


**Figure 2:** Sampling Approaches

**3.1.2 Informed Undersampling:** Informed undersampling is another type of undersampling method which works on no information loss, reversing the case in random undersampling, where there is a risk of loss of potential data. X.Y. Liu et al. [11] had given two examples EasyEnsemble and BalanceCascade algorithms.

**3.1.3 Random Oversampling (ROS):** Oversampling refers to replicating minority class instances forincreasing minority class instances and balancing imbalanced datasets. I. Tomek [12] proposed Tomek Links for increasing the number of minority class data instances.
Oversampling increases the likelihood of occurring overfitting since it makes exact copies of existing instances.

**3.1.4 Synthetic Oversampling:** Synthetic oversampling technique tends to generatesynthetic data for minority classes and is thus frequently referred to as Synthetic Minority Oversampling Technique (SMOTE). Synthetic minority class data is generated by manipulating existing minority class data instances.

This is an effective approach for balancing an imbalanced dataset. While working on this technique some points should be concerned; identification of minority dataspace is most important, if minority dataspace is occurring on boundaries, then generation of new synthetic data will become explicitly difficult because it may be possible that if boundaries are not recognized effectively, newly generated data will fall under majority class. Synthetic data generation is strict to continuous features distribution, there is scope for work that can be performed for discrete feature distributions.
*M. P. Ortiz et al. [2]*, explores the synthetic oversampling within the feature space induced by a kernel function. In this paper convex combination of original points belonging to the same cluster is used for the generation of synthetic data. It states that information available while working on the kernel method is a dot product of images of patterns and this information cannotbe directly used. To overcome this issue, a mechanism is developed, which is based on Euclidean distance, termed as Empirical Feature Space (EFS), which preserves the geometrical structure of the original feature space. The necessary condition for this is that the dot product of images must be original kernel function and the dot product should uniquely determine angles and distance in the feature space.
*S. Barua et al. [3]*, had introduced a new method based on majority weighted minority class instances for generating synthetic data. They had first identified hard-to-learn

informative minority class samples and based on Euclidean distance they had assigned weights accordingly. Synthetic data is then generated from the weighted informative minority class samples using a clustering approach. They had emphasized their approach on strictly identifying the exact point of minority dataspace as if minority class data space falls on boundary condition then oversampling may result in favor of the majority class.

**3.1.5 Adaptive Synthetic Sampling:** SMOTE technique generates the same number of synthetic data samples for each original minority example, this increases the occurrence of overlapping the classes. To overcome this problem, Borderline SMOTE and Adaptive Synthetic Sampling algorithms are introduced.

*S. Ahmed et al. [13]* introduces an ensemble learning approach that uses sampling techniques with bagging or boosting approaches. Paper had proposed ADASYNBagging and RSYNBagging for dealing with imbalanced classification. ADASYN based oversampling technique with bagging is used in ADASYNBagging and random under-sampling and ADASYN based over-sampling technique with bagging algorithm is used in RYSNBagging.

### 3.2 ALGORITHMIC APPROACHES

Algorithmic techniques are also referred to as internal techniques. In this technique classifier learning algorithms are tend for biased learning towards minority data samples. In uneven class distribution, knowledge of the classifier and its application domain is crucial for understanding the failure of the classifier [8].

*Q. Kang et al. [14]*, states that undersampling can be a good option for resolving imbalance dataset problems using a support vector machine. An improved algorithm, the Weighted- Undersampling(WU) scheme is proposed for SVM which is based on geometric distance. In this, some sub-regions are created using majority samples and weights are assigned to these sub-regions based on Euclidean distance to the hyperplane. Samples with higher weight in sub-region have more chances to be sampled and put to use in each iteration, in this way data distribution information of the original dataset would be maintained.

### 3.3 Hybrid Approaches

Hybrid approaches are the combination of data level approaches and algorithmic approaches. They are termed as Ensemble Technique [8]. Ensemble technique is an advanced level technique, which aggregates various classifiers in a single model and constructs a base classifier from training data, and performs classification by taking a vote on the prediction made by each base classifier. Methods used in ensemble techniques are Bagging and Boosting as shown in Figure 3.

For better performance of ensemble classifier than single classifier, following points to be noted;

(a) the base classifier should be independent of each other;
(b) the base classifier should do better than a classifier that performs random guessing.

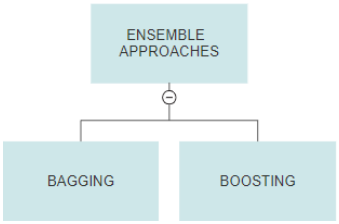


**Figure 3:** Ensemble Approaches

#### 3.3.1 Bagging

Bagging is an approach that focuses onreducing the variance of a high variance low bias base classifier while maintaining the low bias. In this technique, samples with replacement from a dataset according to a uniform probability distribution are repeated in the training process, this approach is also termed as Bootstrap aggregation. The sample size of each bootstrap isthe same as of the original data. Bagging improves generalization errors as it emphasizes reducing variance. The performance of bagging depends on the steadiness of the bottom classifier.

#### 3.3.2 Boosting

Classifiers that are hard to learn and can affect the performance of the classifier are primarily focused on this technique. Adaptive methods are used for the redistribution of training data. Weights are assigned to each training sample and at the end of boosting, weights are changed adaptively. This process goes for various iterations, thus the process is iterative.

*D. Devi et al. [15]* had proposed a boosting-based oversampling technique within an ensemble framework through sensitive error formulation. In this work, the Local Covariance Matrix is used for handling oversampling rate,

and the AdaBoost ensemble model with C4.5 weak learner is implemented as the ensembleframework.

### 3.4 Kernel-Based Approaches

Support Vector Machines are kernel-based learning paradigm representations centered on the principal theories of statistical learning and Vapnik-Chervonenkis (VC) dimensions [6]. SVM focuses on minimizing the total classification error and this learning is facilitated by taking specific examples near concept boundaries to minimize the separation margin between support vectors and the hypothesized concept boundary [18].

In this technique, for achieving separation in higher dimensional space, a kernel function is used to map the linear non-separable space [6].

The kernel function is manipulated in various ways for achieving higher accuracy over minority class during training process using SVM classifier, and sometimes an integrated approach using kernel function and one of the sampling techniques (SMOTE most often) is carried forward.

J. Mathew et al. [22] states that synthetic data generation techniques such as SMOTE distort the performance of an SVM classifier. A kernel-based SMOTE (k-SMOTE) algorithm is proposed which directly generates synthetic data in the feature space. An only kernel function is used to augment the original Gram matrix with newly generated data points.

### 3.5 Cost-Sensitive Technique

The cost-sensitive technique is an intermediate technique between sampling and algorithmic techniques. This technique is also called a Cost-Based Technique. Data level transformations and algorithmic modifications both are incorporated in this technique. This technique assigns cost to instances and develops an effective mechanism for accepting those costs. This technique tries to bias the classifier toward the minority class. A major drawback of this technique is misclassification cost, which has to be defined, as misclassification cost is usually not available in the dataset.

N. Thai-Nghe et al. [23] had combined sampling techniques and compared them with cost-sensitive learning using support vector machines. This approach had reduced the misclassification cost. They had also optimized cost ratio (cost matrix) locally and used cost-sensitive learning for improving classifier performance.

## 4. COMPARISON OF VARIOUS RESEARCH OUTCOMES

In this section, the result outcomes of various researches done in the field of imbalance learning are compared using F-Measure values. Techniques having common datasets and classifiers are considered for comparison. Table1and Table 2 show result comparisons for MWMOTE, SMOTE, and ADASYN techniques are based on a single neural network classifier and kNN classifier [3]. Table 3 contains results for SMOTE, RUS, and Bagging based on the kNN classifier [17]. Table 4 shows the result for HOEC, Rof, Balance Cascade, and Easy Ensemble [17]. Table 5 enlists the result of SVM, SMO, WU-SVM, and U- SVM classifiers [14].

**Table 1:** Result of MWMOTE, SMOTE, and ADASYN for Single Neural Network Classifier [3].

| DATASET | MWMOTE | SMOTE | ADASYN |
|---|---|---|---|
| Abalone | 0.39497 | 0.44167 | 0.29806 |
| Ecoli | 0.73988 | 0.76277 | 0.75505 |
| Glass | 0.87968 | 0.82579 | 0.87148 |
| Page Blocks | 0.93337 | 0.97481 | 0.95289 |
| Pima | 0.67917 | 0.66811 | 0.68335 |
| Vehicle | 0.9208 | 0.91769 | 0.85784 |
| Yeast | 0.68953 | 0.68509 | 0.68609 |

**Table 2:** Result of MWMOTE, SMOTE, and ADASYN for kNN Classifier [3].

| DATASET | MWMOTE | SMOTE | ADASYN |
|---|---|---|---|
| Abalone | 0.44782 | 0.50188 | 0.44095 |
| Ecoli | 0.81196 | 0.78133 | 0.8004 |
| Glass | 0.91928 | 0.88817 | 0.87908 |
| Page Blocks | 0.98068 | 0.9776 | 0.97774 |
| Pima | 0.62194 | 0.63697 | 0.61664 |
| Vehicle | 0.86721 | 0.86446 | 0.85694 |
| Yeast | 0.67901 | 0.65902 | 0.63447 |

**Table 3:** Result of SMOTE, RUS, and BAGGING based on kNN Classifier [17].

| DATASET | SMOTE-kNN | RUS-kNN | BAGGING-kNN |
|---|---|---|---|
| Ecoli | 0.9721 | 0.9175 | 0.9378 |
| Glass | 0.9020 | 0.7012 | 0.9564 |
| Haberman | 0.7283 | 0.6911 | 0.7529 |
| Page Blocks | 0.9721 | 0.9582 | 0.9786 |
| Vehicle | 0.7667 | 0.6798 | 0.8303 |
| Yeast | 0.7665 | 0.7460 | 0.8346 |

**Table 4:** Result of, RoF, BalanceCascade, and EasyEnsemble [17].

| DATASET | HOEC | RoF | Balance Cascade | Easy Ensemble |
|---|---|---|---|---|
| Ecoli | 0.9225 | 0.8219 | 0.9167 | 0.9215 |
| Glass | 0.9587 | 0.9541 | 0.6810 | 0.6785 |
| Haberman | 0.8215 | 0.7283 | 0.6911 | 0.6736 |
| Page Blocks | 0.9804 | 0.9721 | 0.9582 | 0.9589 |
| Vehicle | 0.8565 | 0.8505 | 0.7372 | 0.7380 |
| Yeast | 0.8455 | 0.8219 | 0.7460 | 0.7505 |

**Table 5:** Result of SVM, SMO, WU SVM, AND U-SVM [14].

| DATASET | SVM | SMO | WU SVM | U-SVM |
|---|---|---|---|---|
| Haberman | 0.3391 | 0.4522 | 0.5041 | 0.4142 |
| Page Blocks | 0.6784 | 0.7843 | 0.9324 | 0.8943 |
| Yeast | 0.7843 | 0.8288 | 0.9004 | 0.8372 |

## 5. CONCLUSION

In this paper, various approaches are categorized in data-level approaches, algorithmic approaches, hybrid approaches, kernel-based approaches , and cost-sensitive approaches for handling imbalanced datasets. The approaches are also compared on the basis of f-measure values obtained from researches. As an end note of this work, it can be concluded that SMOTE overcomes limitations of RUS and ROS. Integrating two techniques in hybrid approaches give better results than individual techniques.

## 6. ACKNOWLEDGEMENT

I am grateful to my esteemed guide Dr. Anuradha Purohit, Associate Professor, SGSITS, Indore, and co-guide, Ms. Himani Mishra, Assistant Professor, SGSITS, Indore, for their support and guidance for this work.

## REFERENCES

1. P. G. Espejo, S. Ventura, F. Herrera, **A Survey on the Application of Genetic Programming to Classification.** IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Volume: 40, Issue: 2, March 2010. **DOI:**10.1109/TSMCC.2009.2033566
2. M. P. Ortiz, P. A. Gutierrez, P. Tino, C. H. Martinez, **Oversampling the Minority Class in the Feature Space**, IEEE Transactions on Neural Networks and Learning Systems, Vol. 27, Issue 9, pp. 1947 - 1961, Sept. 2016. **DOI:** 10.1109/TNNLS.2015.2461436
3. S. Barua, M. M. Islam, X. Yao, and K. Murase, **MWMOTE—Majority weighted minority oversampling technique for imbalanced data set learning,** IEEE Trans. Knowl. Data Eng., vol. 26, no. 2, pp. 405–425, Feb. 2014. **DOI:** 10.1109/TKDE.2012.232
4. T.E. Fawcett and F. Provost, **Adaptive Fraud Detection**, Data Mining and Knowledge Discovery, pp 291-316, 1997. **DOI:** 10.1023/A:1009700419189
5. S.H. Clearwater, S.G. Stern, **A Rule learning program in High Energy Physics Event Classification**, Computer Physics Communications, Vol. 62, Issue 2, pp. 159-182, Dec 1991. **DOI:**10.1016/0010-4655(91)90014-C
6. H. He, E.A. Garcia, **Learning from Imbalanced Data,** IEEE Transactions on Knowledge and Data Engineering, Volume: 21, Issue: 9, Sept. 2009**. DOI:** 10.1109/TKDE.2008.239
7. P. Lim, C.K. Goh, K.C. Tan, **Evolutionary Cluster-Based Synthetic Oversampling Ensemble (ECO-Ensemble) for Imbalance Learning.** IEEE Transactions on Cybernetics, Volume: 47, Issue: 9, pp. 2850-2861 Sept. 2017. **DOI:** 10.1109/TCYB.2016.2579658
8. M. Galar, A. Fernandez, E. Barrenechea, H. Bustince. And F. Herrera, **A review on ensembles for the class imbalance problem: Bagging-, Boosting-, and hybrid-based approaches.** IEEE Trans. Syst., Man, Cybern. C, Appl. Rev. vol. 42, no.4, pp. 463-484, Jul. 2012. **DOI:** 10.1109/TSMCC.2011.2161285
9. J. Zhang, I. Mani, **KNN Approach to Unbalanced Data Distribution: A Case Study Involving Information Extraction.** Proceedings of the ICML '2003 Workshop on Learning from Imbalanced Datasets, (2003)
10. P.E. Hart, **The Condensed Nearest Neighbor Rule(corresp).** IEEE transaction on information theory, Volume: 14, issue : 3, pp. 515-516, 1968
11. X.Y. Liu, J. Wu, and Z.H. Zhou, **Exploratory Under Sampling for Class Imbalance Learning,** Proc. Int'l Conf. Data Mining, pp. 965-969, 2006.
12. I. Tomek, **Two Modifications of CNN.** IEEE Transactions on System, Man, and Cybernetics, volume:

SMC-6, issue: 11, pp. 769-772, 1976. **DOI:** 10.1109/TSMC.1976.4309452

13. S. Ahmed, A. Mahbub, F. Rayhan, M. R. Jani, S. Shatabda, D. M. Farid, **Hybrid Methods for Class Imbalance Learning Employing Bagging with Sampling Techniques**, 2nd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2017, 30 August 2018. **DOI:** 10.1109/CSITSS.2017.8447799
14. Q. Kang, L.S. MC. Zhou, X. Wang, QD. Wu, Z. Wei, **A Distance-Based Weighted Undersampling Scheme for Support Vector Machines and its Application to Imbalanced Classification**, IEEE Transactions on Neural Networks and Learning Systems, Volume: 29, Issue: 9, pp. 4152 – 4165, Sept. 2018. **DOI:** 10.1109/TNNLS.2017.2755595
15. D. Devi, S. K. Biswas, B. Purkayastha, **A Boosting based Adaptive Oversampling Technique for Treatment of Class Imbalance**, 2019 International Conference on Computer Communication and Informatics, 02 September 2019. **DOI:** 10.1109/ICCCI.2019.8821947
16. W. Feng, G. Dauphin, W. Huang, Y. Quan, W. Bao, M. Wu, Q. Li, **Dynamic Synthetic Minority Over-Sampling Technique-Based Rotation Forest for the Classification of Imbalanced Hyperspectral Data**, IEEE Journal of Selected Topics In Applied Earth Observations and Remote Sensing, Vol. 12, No. 7, July 2019. **DOI:** 10.1109/JSTARS.2019.2922297
17. K. Yang, Z. Yu, X. Wen, W. Cao, C. L. P. Chen, Hau-San Wong, J. You, **Hybrid Classifier Ensemble for Imbalanced Data**, IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 4, pp.1387-1400, 2020. **DOI:** 10.1109/TNNLS.2019.2920246
18. V.N. Vapnik, **The Nature of Statistical Learning Theory.**, Springer, 1995
19. N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, **SMOTE: Synthetic Minority Over-Sampling Technique,** J. Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.
20. H. Guo and H.L. Viktor, **Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost IM Approach,** ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 30-39, 2004.
21. K. Woods, C. Doss, K. Bowyer, J. Solka, C. Priebe, and W. Kegelmeyer, **Comparative Evaluation of Pattern Recognition Techniques for Detection of Microcalcifications in Mammography,** Int'l J. Pattern Recognition and Artificial Intelligence, vol. 7, no. 6, pp. 1417-1436, 1993.
22. J. Mathew, M. Luo, C. K. Pang, and H. L. Chan, **Kernel-Based SMOTE for Classification of Imbalanced Datasets,** IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society, 2015. **DOI:** 10.1109/IECON.2015.7392251.
23. N. Thai-Nghe, Z. Gantner and L. Schmidt-Thieme, **Cost-sensitive learning methods for imbalanced data,** The 2010 International Joint Conference on Neural Networks (IJCNN), 2010. **DOI:** 10.1109/IJCNN.2010.5596486
24. M.K.R. Mallidi, and Y. Zagabathuni, **Analysis of Credit Card Fraud Detection using Machine Learning models on balanced and imbalanced datasets.** International Journal of Emerging Trends in Engineering Research, vol. 9 no. 7, July 2021. **DOI:** doi.org/10.30534/ijeter/2021/02972021
25. A. Singh, and A. Purohit, **A Survey on Methods for Solving Data Imbalance Problem for Classification.** International Journal of Computer Applications, vol. 127-no. 15, Oct. 2015.