Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 9, 4584-4595

# Frequent Relational Smote For Class Imbalance An Experimental Approach

**Ratul Dey[1]*, Rajeev Mathur[2]**

*[1] *Research Scholar, School of Engineering & Technology, Jaipur National University, Jaipur , E-mail:- ratulcs92@gmail.com*

[2] *Director, School of Engineering & Technology, Jaipur National University, Jaipur, E-mail:- mathur.rajeev96@gmail.com*

*[1] *Research Scholar, School of Engineering & Technology, Jaipur National University, Jaipur E-mail:- ratulcs92@gmail.com*

**Corresponding Author: - Ratul Dey*

**Abstract:-**
Over the year, our society has been trying to implement such kind of computer which is intelligent, efficient, and perfect. Computer intelligence is based upon previous data analysis. Class imbalance is one of the significant difficulties in the machine learning field, which degrade the performance of analysis and decision-making ability. In this paper we propose FR-SMOTE to mitigate the class imbalance problem. It aids in the improvement of learning performance in the context of an imbalanced learning environment. Our result analysis shows a significant improvement over the existing state-of-art, where we can see the achievement of 86% true positive rate over the imbalanced classes in machine learning algorithm. In this paper we apply several machine learning algorithm like SVM, KNN, Random Forest in highly unbalanced data(WOS), after applying RUS, OS-SMOTE and implemented FR-SMOTE in the base of accuracy, Precision, Recall and F1 Score.

**Index Terms:-** Class Imbalance, Random over sampling, Random Under sampling Technique, Synthetic Minority Oversampling Technique, Support Vector Machine, K-Nearest Neighbor.

## I. INTRODUCTION

In the recent development of machine learning and data science methodology, they are working with a different type of dataset. Typically, data is an essential part of data science and machine learning. Now raw data that comes from observing any other source, not in proper format or data is highly imbalanced. In recent decades class imbalance problems are a significant amount of interest from the industry. The main reason behind this is that whatever algorithm we apply to highly imbalanced data will not give substantial accuracy. In this paper, we describe the different types of class imbalance problems and how to mitigate this. The class imbalance problem has been recognized in many application domains such as telecommunication customer management, medical diagnosis, text classification, and credit card fraud detection. There has been a lot of research on the class imbalance problem paper [1], [2] which review different type of sampling method for imbalanced learning like under-sampling and random oversampling. Sampling techniques are used for handling the class imbalance problem; a few of are Synthetic sampling, adaptive synthetic sampling, sampling with data cleaning technique, cluster-based sampling method, and cost sensitive method. Different types of training data resampling techniques like Gaussian process model, synthetic binary classification dataset, uterine contraction datasets [3], fuzzy rule-based system [4]. There are different scores to identify accuracy like classifier performance matrix, F-measure, Geometric mean, ROC Curve, logarithmic Score, Brier inaccuracy Kolmogorov- Smirnov statistic [5]. Feature selection is one of the parts of machine learning, and data mining techniques [6]. In the class imbalance problem, different types of algorithms are present to handle the dataset. Balanced distribution can be achieved either using under sampling for the majority class or oversampling for the minority class [7].

**Table I:** Comparison Of Different Types Of Learning Algorithms

| | | | | |
|---|---|---|---|---|
| ANN(Artificial Neural Network) | Deep Learning | Implement human intelli-gence as artificially | Using mutation implement | Computational model based on the structure and function of biological neural network |
| KNN (k- near-est neighbor) | Data mining | Working with deep neural network | Find the nearest node | Non pandemic method used for classification and Regression |
| Dimensionality Reduction Algorithms | Unsupervised Learning | Linear dimensionally | Reduce the number of a random variable | n-dimensional converted to two dimensional, still maintain the structure of data |
| Gradient Boosting algorithms | Supervised Learning | Decision tree | Bosting classification | It divides a space between two class using an ensemble of decision tree |
| Protocol | Category | Dataset | Working Principal | Study /Remarks |
| Linear Regression | Supervised Learning | Linear dataset | It defines the relationship between two continuous variables | Many real-world situations are simplified by linear regression. There is no linear link between covariates and response variables. |
| Logistic Regression | Supervised Learning | Linear dataset | It is a binary dependable variable | can't solve non-linear problems. only give the categorical outcome |
| SupportVector Machine | Supervised Learning | Linear dataset | a model that assigns new values to one category or the other | give the boundary |
| Decision Trees | Supervised Learning | nonlinear dataset | the result in the binary format and it's working in a nested system | One kind of binary, not give breakage information. Only graphical representation |
| Random Forest | Supervised Learning (Classification/Regression) | nonlinear dataset | Discover more complex dependencies at the cost of more time for fitting | Ensemble learning method for classification |
| Na¨ıveBayes Classification | Supervised Learning | Linear and non-linear | Probabilistic classifier | Based on Baye's theorem |
| OrdinaryLeast Square Regression | Supervised Learning | Linear data set | Classifier | Minimize the sum of the squre of the difference between the observed dependent variable |
| K-Means | Unsupervised Learning | nonlinear dataset | Create cluster to identify the different cluster | Finding groups of object that are more similar to each other and create a cluster. |
| Ensemble Method | Supervised learning | Linear and non-linear data set | Marge different type of algorithm and get the best one | Compare and identify the accurate result |
| Apriori Algorithm | Unsupervised Learning and data mining technique | Linear data set | Using association rule | Frequent itemset mining |
| Principal Component Analysis | Unsupervised Learning | Non-linear Data | A different view for dataset | Find the best fitting |
| Nearest Neigh bor. | Supervised Learning | Linear Data | It is used for classification and regression. | Finding the point in a given set that is close to a given point. |
| Neural Networks | Reinforcement Learning | Used all type of dataset | Work as a biological neural network | It is a network or circuit of neuron |
| CNN (Convolution Neural Network) | Deep Learning | Image classification | Identify images | From image to identify data |
| RNN(recurrent Neural Network) | Deep Learning | Language or pat-terned classification | Identify pattern | Used in different language and NLP |

The above table defines the different types of learning algorithms and how they work with varying data sets. There are a few types of machine learning Algorithm that one determine with various kind of data set.

## A. Imbalanced Learning

Imbalance data sets corrupt the presentation of data mining and machine learning strategies as the general accuracy and decision making be one-sided to the greater part class, which prompts misclassifying the minority class tests or besides regarding them as commotion. We choose RUS, ROS and SMOTE for imbalanced problem and apply boosting algorithm to get cost sensitivity. There are few technique to reduce the imbalance problem, SMOTE is one of them that create synthetic node for the minority class.

### 1) Random Oversampling Technique (ROS):

Oversampling looks to increase the quantity of minority class individuals in the training set. Random oversampling is a simple approach to cope with re-sampling in which individuals from the minority class are chosen at random. These individuals are then replicated and assigned to the dedicated training set. [8] Sample data set DB is formed in several samples randomly selected from the minority class $X_{min}$, by copying and then deleting the selected sample $X_{min}$. A new minority class is derived from the raw dataset $X_{new} = X_{min} + DB$.

### 2) Random Under sampling Technique (RUS):

Undersampling is a cycle that looks to lessen the quantity of majority class individuals in the training set. Randomly oversampling is a famous method for resampling. The majority class documents are randomly eliminated in the training set until the minority, and majority class ratio is at the desired level. [8] In some samples randomly picked from the majority class $X_{max}$, sample data set DB is generated by copying the selected sample and then deleting it to $X_{max}$. The raw dataset is deleted to establish a fresh minority class X new= X max-DB.

### 3) Smote:

The Synthetic minority Oversampling Technique [2], [9] working for class imbalance datasets. SMOTE is quite possibly the most famous methodology for producing Synthetic examples [10]. SMOTE is an over-inspecting approach in which the minority class is over-tested by making synthetic models instead of over-examining with substitution.

### 4) Smoteenn:

Synthetic minority Oversampling Technique with Edited Nearest Neighbour method [11], [12]. ENN is a method for cleaning; it eliminates noisy instances from both majority and minority classes. Applying SMOTE create new node in the minority class, and try to solve imbalance problem.

### 5) Deep Smote:

The combination of deep learning and SMOTE is one of the effective ways to mitigate class imbalanced problem. Fusing Deep Learning [9], [12] is one of the techniques that provide high accuracy in less number of dataset.

## B. Motivation

Real world data are not always properly ready for analysis. Anlysis highly imbalanced data are lot more challenges compared to balanced dataset. In data anlysis lots of algorithm are there but they trained by the dataset, if this data set are highly imbalanced then algorithm will unable to get proper training, so output will not come upto the benchmark. Most algorithms aim to eliminate class imbalance form the dataset. They demonstrate that the algorithms are quite accurate, when the same technique is applied to different datasets, the accuracy falls short. As a result, algorithms are effective for a limited number of datasets. To minimize the class imbalance problem, solution that is easily adaptable to various types of datasets is necessary.

**Table II:** Literature Survey On Class Imbalance Problem By Major Researcher

| Author | Category | Dataset | Working Principal | Study /Remarks |
|---|---|---|---|---|
| rahman *et al.* [14] | 2 class (High Risk and Low Risk) | There are 7770 and 6593 training instance in Reuters and CiteSeer respectively | Modified cluster based under sampling method. SMOTE | Averaging the training time of the data set prepared by synthetic sample using SMOTE out of 10 runs. |
| ertekin *et al.* [15] | 2 type class | Reuters-21578 and CiteSeer, g-means metric | Increase the data of minority class | Compose pre-processing time and training time |
| weng et al. [16] | Medical data depends upon (0 to 1) 98 data sets across 4 different learner | Class imbalance problem, Weighted AUC and ROC curve, Design to cope with cost biases | Compare with precision recall curve | Changes the previous view about class independent evaluation |
| zhu et al. [17] | Class imbalance problem in WSD tasks 38 random chosen ambiguous noun used | 3 type of class (Ideal, Max Confidence, Min error), Stop sampling after a certain limit | Max-confidence an min error strategy in framework, Analysis resampling technique including oversampling and under sampling in active learning | Working with over sampling and under sampling in active learning |
| pouyanfaret al. [18] | Image/Video analysis, Multimedia system able to discover network camera those data is publicly available, Image include 19 semantic concept such as highway, intersection, yard 299*299 pixel. (70% training, 10% reference, 20% testing) | Deep Learning CNN, Utilize data augmentation and transfer learning technique to avoid overfitting towards the minority class | Improve in CNN model with less complexity | Model depends upon convolution neural network, It can identify objects and specify from real time visual data |
| wang et al. [19] | Capture classification error from both majority and minority class equally 20 new group is collection of approximately 2000 new group documents portioned. | Compared with convolution method. Mean square error, Mean false error and mean squared false error solved using neural network | MFE and MSFE approaches achieve higher AUC than MSE approaches under the same loss | Only work with few type of un-structural dataset |

The above table describes different type of comparison analysis of class imbalance problem.

## C. Contribution

In this paper, we proposed an FR-SMOTE for class imbalance datasets. We use Accuracy, Precision, Recall, F1 Score to analysis the performance in imbalanced data. here three kind of datasets used one is greater than 50000 dataset, second one for more than 10000 datasets and the third one less than 5000 datasets. Moreover, data sets are used by Logistics Tomek Links, Random forest Tomek links, Ada boost Tomek link, KNN Tomek link, D-Tree Tomek link, SVM Tomek link, Naive Bias Tomek link, Xgboost Tomek, Light GBM . We also show a prototype for FR-SMOTE, while taking three different kind of primary datasets, and compare with the other algorithm to reduce the imbalance problem. The key *contributions* of this paper are as follows:

- We propose an FR-SMOTE algorithm 1.
- We implement the FR-SMOTE algorithm in three individual datasets one greater than 50 thousand second greater than 10 thousands and last one less than 5 thousand.
- The performance results of the proposed algorithms checks with F1 Score, R1 Score, AUC, etc. and compared to the existing works.

## D. Organization

The rest of the paper is organized into four sections. Section II describes the related works. The proposed algorithm is presented and analysed in Section III. Results and discussions are described in section IV, and in Section V we conclude this paper.

## II. Related Works:

The industry and academics have collaborated flawlessly to create significant advancements in the field of class imbalance technique. The majority of algorithm are working to reduce class imbalance from the datasets. There are different datasets they are working with. They show the accuracy of those algorithms is high. If the same algorithm uses other datasets then the accuracy is not up to the mark. So algorithms are successfully working for some specific datasets. Table 1 describes different kind of machine learning algorithm which are depends upon the different kind of training and testing method. So we can easily identify that if the training data are highly imbalanced, some of the rare case event are neglected for the prediction though this are the play a important role for the prediction.

### A. Class Imbalance

An important part of data imbalance is vision classification, [13] traditional re-sampling and cost-sensitive learning schemes. The protocol used Deep embedding of Class-level clusters and class-level constraints. Firstly, it identifies the majority and minority class of training data, then the similarity between both classes using Triplet embedding and Quintuplet embedding. Assigning class costs and resampling classes in batches using the Network architecture. It Generates quintuplets from cluster class membership. For Feature learning, using Cluster-wise KNN search large margin of imbalance can resistant. Another author discusses the unbalanced learning problem is a difficult and crucial subject in the realm of knowledge discovery and data engineering. The data set is always preferred the balanced distribution of data, Equal costs of misclassification [1]. In real-world datasets, they are highly imbalanced and are a much more complex data set. In many cases, data are high dimensionality and small sample sizes like face recognition and gene expression. The issue of small sample size embedded absolute rarity and within-class Imbalance and the failure of generalizing inductive rules by learning algorithms. It is more challenging to form reasonable classification decision boundaries over more features but fewer samples, so it's overfitting the risk. There are few imbalances in learning like sampling, kernel, active learning methods cost-sensitive methods.

### B. Smote

Synthetic minority oversampling technique is a process to handle class imbalance problem [2], [9], [12], [20], [21] SMOTE is one of the popular technique used in classification problem of Imbalanced dataset. SMOTE is consider one of the most secure technique and reliable under the different circumstances. The main aim of the SMOTE is to increase the prediction accuracy. In the minority class there are very Less amount of data are there, now the above algorithm synthetically create data, so that minority class have enough data to get accurate result.

### C. Borderline Smote

Hybrid Clustered Affinitive Borderline SMOTE (HCABSMOTE) is a Minority class cases there are borderline SMOTE will need further training. It first identifies borderline minority cases, which are then used to generate synthetic examples with their chosen k nearest neighbours. Borderline SMOTE is used as an oversampling strategy in conjunction with a deep learning model comprised of two autoencoders and a SoftMax layer [22], [23].

### D. Smote-Nc

SMOTE-NC is an model for generating synthetic data in order to oversample a minority target class in an unbalanced dataset [24]. The parameters that may be connect include k-neighbours, which determines how many nearest neighbours to use to build the new sample, and sampling strategy, which specifies how many new samples to generate.

### E. Borderline Smote Svm

This method express Bagging of Extrapolation Borderline SMOTE SVM [25]. Using interpolation to generate some synthetic samples for the minority has been shown to be useful for reducing the level of imbalance and improving performance. It appears that samples near the decision boundary outnumber those further away. The ensemble method address the Bagging of Extrapolation Borderline-SMOTE SVM

### F. Smote Boost

The Synthetic Minority Oversampling Technique (SMOTE) and the traditional boosting process are combined in this approach. All misclassified instances are given equal weights in the usual boosting approach. The boosting method (Adaboost) handles both types of mistakes (FP and FN) similarly, sampling distributions in subsequent boosting rounds may contain a greater proportion of majority class instances [26].

### G. Fw-Smote

It is an abbreviation for Feature Weighted-SMOTE. To the best of our knowledge, It is the

sampling approach repetitive for incorporating a scheme weighted in nature. It also follows OWA operators. The weighted resampling approach FWSMOTE and its expansion as a technique of feature selection. FW-SMOTE is not intended to be a feature selection approach for classification tasks [27].

### H. C-Smote

It is intended to equalize an imbalanced data stream and may be used with almost any SML-model. C-SMOTE is an abbreviation for Continuous-Smote, ensuring that the updated Smote version is constantly used. It introduces a compromise between improving minority and majority class metrics. It also stores some data in a window and initiates a rebalancing process [28].

### I.Importance-Smote

Using this technique, only borderline and edge samples are oversampled from the minority class. Synthetic minority samples are created based on the composition and distribution of their nearest neighbours, which determines their relevance [29].

### J. Smote-Rknn

This algorithm based upon SMOTE and reverse k-nearest neighbours combinedly. It identifies noise based on its density information, which is gathered globally. To begin, the SMOTE technique is applied to the original training set in order to establish a balanced training set. The number of reverse k-nearest neighbours for every training sample is then counted within each class. Following that, the estimated probability density of each occurrence is determined using RkNN findings. In addition, a normalised approach is used to proportionately modify the approximation probability densities such that they are comparable across classes [30].

### K. Sshm

Smote-stacked hybrid model [31] SSHM technique, which combined SMOTE and Stacking, was used. This model give the high accuracy in different parameter in different dataset.

### L. Smote-Lmknn

A synthetic minority oversampling technique based on local means-based k-nearest neighbour [32], The local mean based KNN (LMKNN) is originally developed in SMOTELMKNN to explain the local feature of unbalanced data. Second, to eliminate noise and dangerous borderline samples, a novel LMKNN-based noise filter is presented. Third, to generate synthetic minority class samples, an interpolation between a base sample and its LMKNN is proposed.

### M. Adaptive Synthetic Sampling

The Adaptive Synthetic sampling (ADASYN) [33]–[35] data augmentation algorithm used to eliminate the interference of the unbalanced sample distribution on model training. By balancing the sample distribution, this strategy can successfully avoid the model being sensitive to large samples and disregarding small samples.

## III. Proposed Algorithm

The data are pre-processed by data cleaning unit (DCU) and data integration unit (DIU). Following that, the FR-SMOTE approach was used to the supplied dataset to address the class imbalance problem. Then different kind of machine learning and boosting algorithm are applied to the given dataset and their result are evaluated. Our proposed Frequent Relational Synthetic Minority Oversampling Technique (FRSMOTE) is a method that creates a synthetic node in the minority class based on the cluster head.

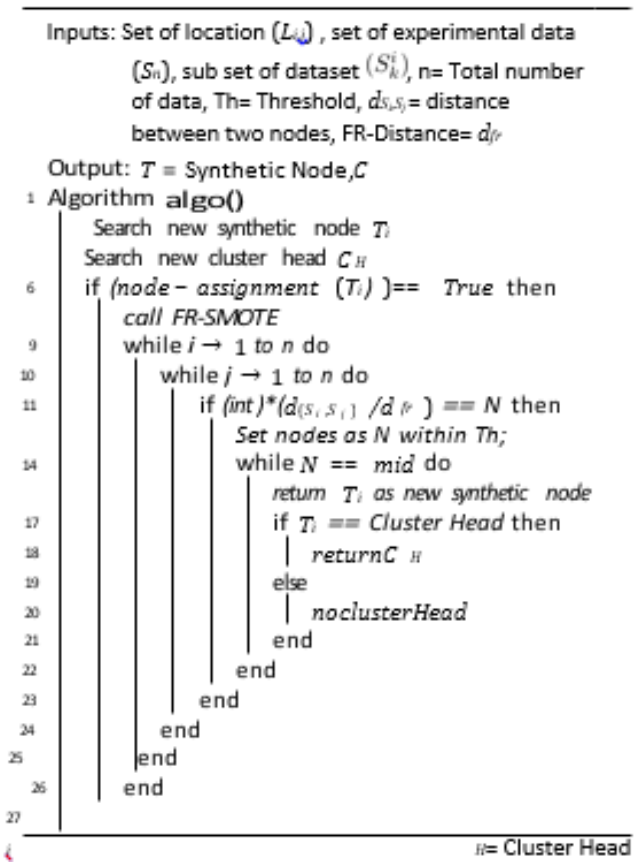



**Fig.1:** Algorithm 1: New node assignment and cluster head creation using FR-SMOTE

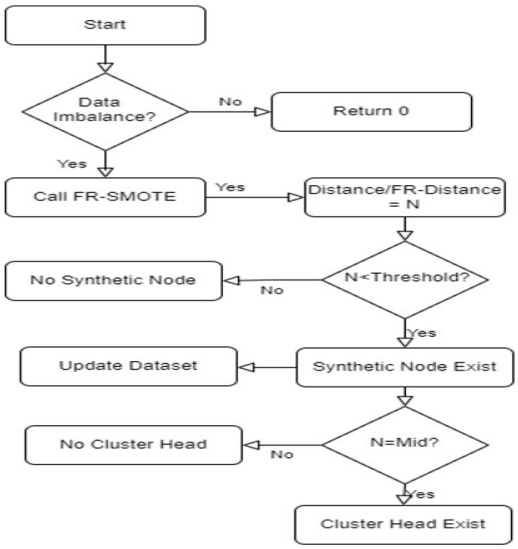



**Fig. 2:** Flowchart of proposed FR-SMOTE.

The Fig.2 describes the if two-node of the minority class distance is greater than FR-distance value, then depending upon threshold value, it creates multiple synthetic nodes. If the synthetic node is the cluster head then it from cluster head it defines the next node in the minority class. Finally, all the synthetic node stores are stored in the minority class and solve the minority problem of the datasets.

## A. Dataset

**Table III:** Sample Dataset For Analysis

| | Positive | Negative | Total |
|---|---|---|---|
| Surgical Deepnet | 3690 | 10946 | 14636 |
| Hotel Booking | 17 | 119373 | 119390 |
| Credit Risk | 981 | 269 | 1250 |

The dataset has been taken from Kaggle that consists of three completely different kind of Imbalanced dataset. At the very fast we find the accuracy, precision, Recall, F1 score of those three unbalanced datasets. Then in the same dataset apply under sampling, over sampling, Random over sampling, SMOTE individually and get accuracy precision, Recall, F1 score using different type of boosting algorithm of previous three dataset. Finally apply with FR-SMOTE individually with the same and compare with them.

The three dataset are in different size Surgical Deepnet is belonging more than 10000, Hotel Booking is belonging more than 50000, and Credit Risk is belonging less than 5000. $L_{i,j}$ represent the set of data location, where $L_{i,j}=l_{0,0}, l_{1,1}, l_{2,2},...,l_{i,j}$ and set of experimental data is represented as $S_N$, where $S_N = s_1, s_2, s_3,...,s_n$ where $50000 > n > 0$. The set of subset dataset is $S_k^i$, where $Ski=s1k, s2k, s3k,...,sik$.

## B. Parameter

For the result analysis we are use different kind of parameter like:
Precision/Specificity: The number of selected instances that are relevant. TP/(TP+FP)

Recall/Sensitivity: The number of relevant instances that are selected. TP/(TP+FN)

F1 score: Measure of precision and recall with a harmonic mean.

MCC: The coefficient of linkage between observed and expected binary classifications.

AUC: There is a correlation between true-positives and false-positives. AUC=(TP+TN)/(TP TN FP FN)

There we use different kind of boosting algorithm like

### 1) Adaboost:

AdaBoost, also known as Adaptive Boosting, is an Ensemble Method Machine Learning approach. One-level decision trees, also known as decision trees with one split, are the most common AdaBoost approach. These trees are often referred to as Decision Stumps.

### 2) Catboost:

It is working for gradient boosting on decision trees. It works in two sections one is Category, and another one is Boosting. It is a data boosting technique that can deal with category variables. Most machine learning algorithms cannot deal with input that contains different categories. Internally, CatBoost is capable of handling categorical variables in datasets. Various statistics on feature combinations are used to transform these variables into numerical values [36].

### 3) Xgboost:

It works with Gradient boosting in supervised learning. It gives better performance than CATBOOST. A Gradient Boosting Machine (GBM) merges the predictions from numerous decision trees to provide the final predictions. [37].

### 4) Light Gbm:

Instead of growing in levels, Light GBM grows leaves by leaves. When a leaf node is divided, only the one with the highest delta loss is divided again. Light GBM can easily manage large volumes of data. However, keep in mind that this algorithm struggles with a limited amount of data points. The trees in LightGBM grow leaf by leaf rather than level by level. Following the initial split, the following separation is performed exclusively on the leaf node with the highest delta loss. It is one of the boosting frameworks based on a decision tree algorithm [38].

### 5) Random Forest:

It works with the high dimensional dataset that creates faster learning than other in the case of feature selection. Random forest is a form of Supervised Machine Learning Algorithm used to solve regression and classification issues. It uses the majority vote for classification and the average for regression to create decision trees from many samples [39].

### 6) Logistic Regression:

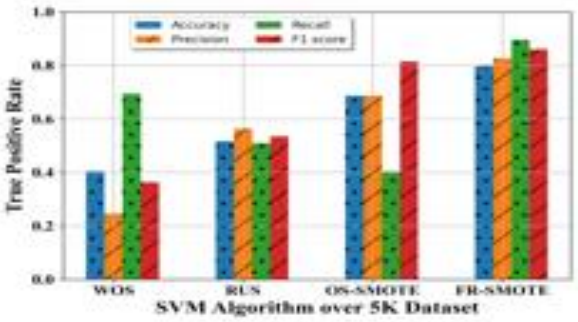
It is a statistical analysis method that works With mainly binary datasets. The classification approach of logistic regression employs supervised learning to predict the likelihood of an input variables. There are only two possible classes since the goal or dependent variable is binary in nature [40].

## C. System Model

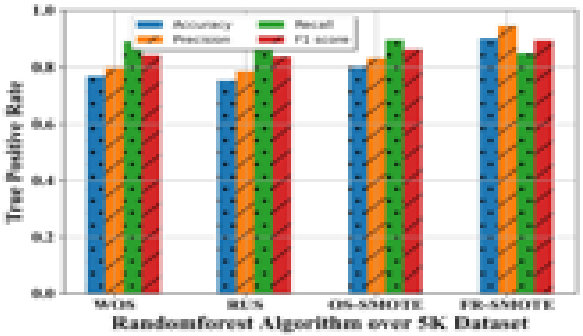
The proposed system works with Intel *i*5 $11^{th}$ generation computer with the 8GB RAM. and the complete experiment done with Python Jupiter notebook.

## Iv. Results And Discussion

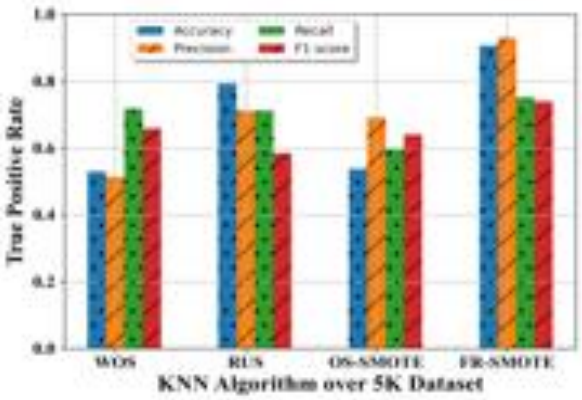
Here we apply some well-known machine learning algorithms such as XGBOOST, SVM, Random-forest, KNN, and GBM with different datasets. These datasets are taken from Kaggle, and Python COLAB does the execution.



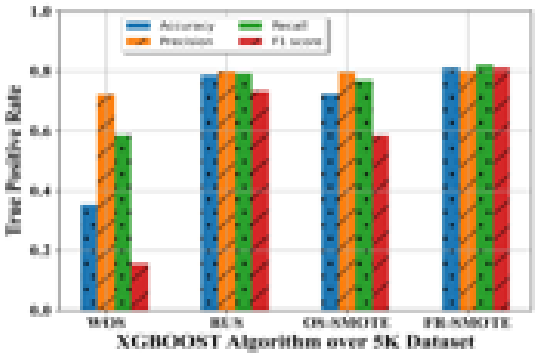

(a) SVM Algorithm over 5K Dataset



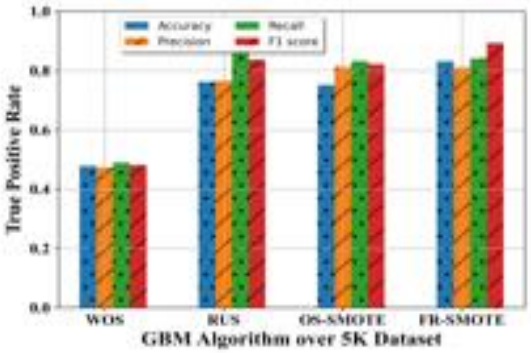

(b) Random forest Algorithm over 5K Dataset




(c) KNN Algorithm over 5K Dataset

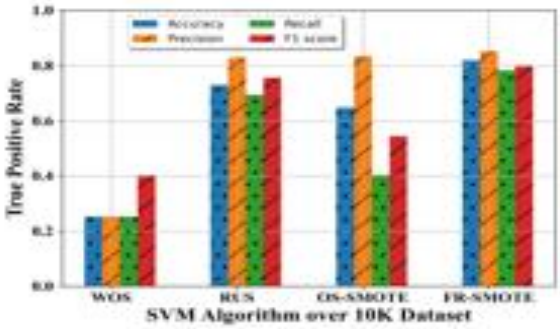



(d)XGBOOST Algorithm over 5K Dataset

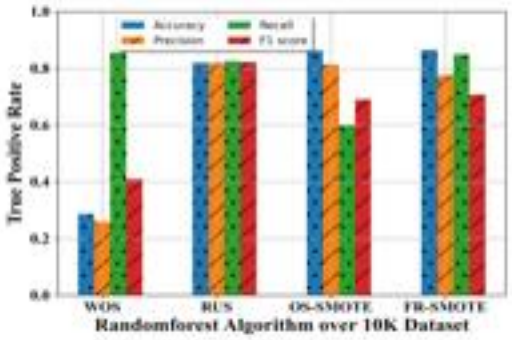



(e) GBM Algorithm over 5K Dataset

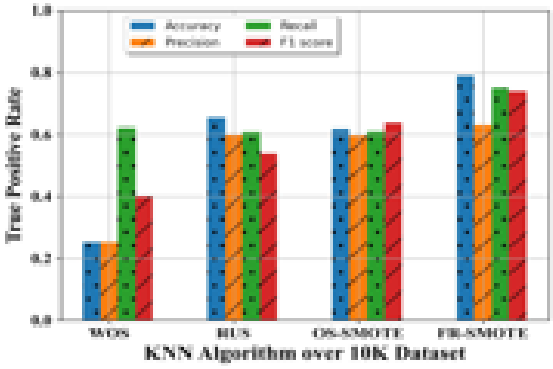
**Fig. 3:** Comparison analysis over 5000 dataset



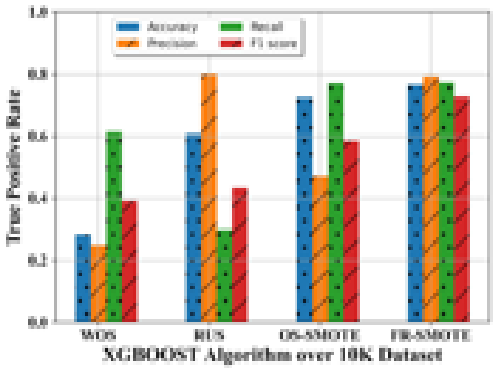

(a) SVM Algorithm over 10K Dataset



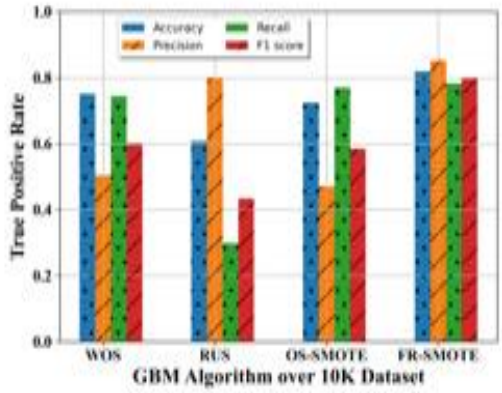

(b) Random forest Algorithm over 10K Dataset




(c) KNN Algorithm over10K Dataset

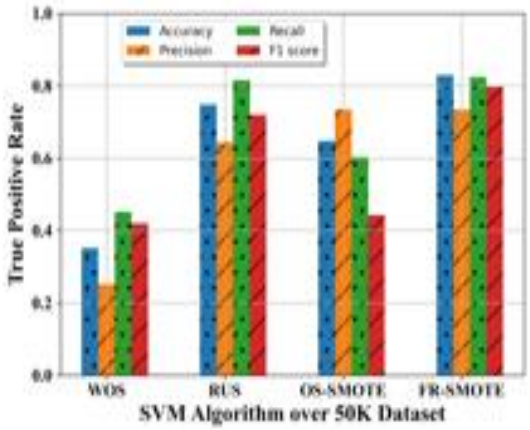



(d) XGBOOST Algorithm over 10K Dataset

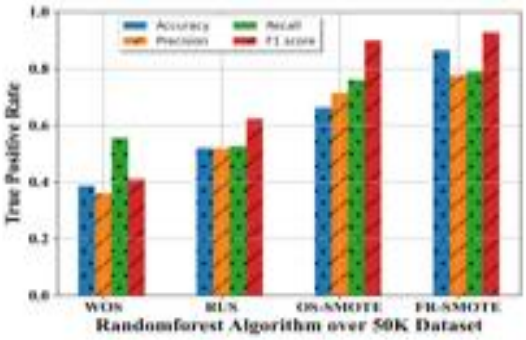



(e) GBM Algorithm over 10K Dataset

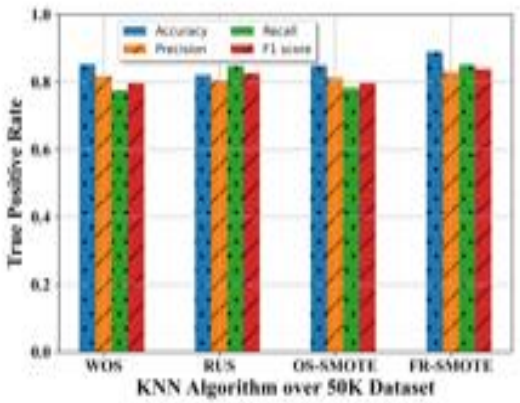
**Fig. 4:** Comparison analysis over 10000 dataset



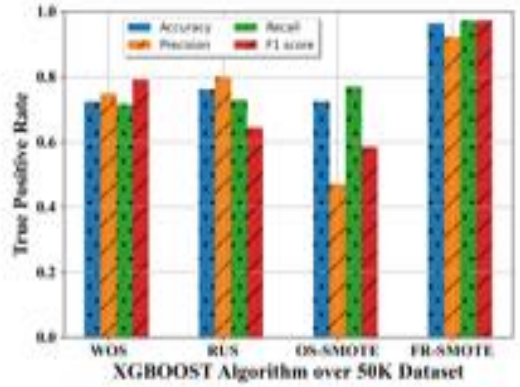

(a) SVM Algorihm over 50K Dataset



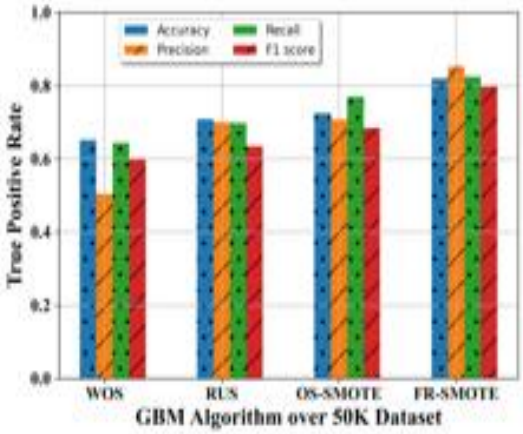

(b) Random forest Algorithm over 50K Dataset




(c) KNN Algorithm over 50K Dataset




(d) XGBOOST Algorithm over 50K Dataset




(e) GBM Algorithm over 50K Dataset
**Fig. 5:** Comparison analysis over 50000 dataset

Here we use Without sampling(WOS), Random Undersampling(RUS), Oversampling Synthetic Minority Over Sampling Technique(OS-SMOTE), and our implemented Frequent Relational Synthetic Minority Over Sampling Technique(FRSMOTE). Finally, it compares the accuracy, precision, recall, and F1 score. We, respectively, run our proposed class imbalance algorithm over 5000 Fig.2, 10000 Fig.3, and 50000 Fig.4 dataset III with existing well-known machine learning algorithm like SVM- Support Vector Machine, Random Forest, KNN- k-nearest neighbors, XGBOOST- Extreme Gradient Boosting, Light GBM- Light Gradient Boosting Machine.
Fig.3,4,5 describes four different evaluation method such as accuracy, Precision, Recall, F1-score executed in over four individual class imbalancing technique i.e. without sampling (WOS), Random Under sampling (RUS), Synthetic Minority Over-sampling Technique (SMOTE), and Frequent Relational Synthetic Minority Over-sampling Technique (FR- SMOTE). Where our proposed class imbalancing algorithm FR-SMOTE achieve average true positive rate of more than 85%.

In Fig. 2(a) 3(a) 4(a) SVM algorithm in the 5000 datasets, 10000 datasets, 50000 datasets respectively. Initially it uses without sampling data and find accuracy, Recall, Precision, F1 score. Then using same data set use different class imbalance algorithm like Random under sampling, Oversampling SMOTE(OS-SMOTE), and our proposed Frequent Relational Synthetic over sampling Technique (FR-SMOTE) and compare with accuracy, Recall, Precision, F1 score. Finally found FRSMOTE give better performance.

In Fig. 2(b) 3(b) 4(b) Random Forest algorithm in the 5000 datasets, 10000 datasets, 50000 datasets respectively. Initially it uses without sampling data and find accuracy, Recall, Precision, F1 score. Then using same data set use different class imbalance algorithm like Random under sampling, Oversampling SMOTE(OS-SMOTE), and our proposed Frequent Relational Synthetic over sampling Technique (FR-SMOTE) and compare with accuracy, Recall, Precision, F1 score. Finally found FRSMOTE give better performance.

In Fig. 2(c) 3(c) 4(c) KNN algorithm in the 5000 datasets, 10000 datasets, 50000 datasets respectively. Initially it uses without sampling data and find accuracy, Recall, Precision, F1 score. Then using same data set use different class imbalance algorithm like Random under sampling, Oversampling SMOTE(OS-SMOTE), and our proposed Frequent Relational Synthetic over sampling Technique (FR-SMOTE) and compare with accuracy, Recall, Precision, F1 score. Finally found FRSMOTE give better performance.

In Fig. 2(D) 3(d) 4(d) XGBOOST algorithm in the 5000 datasets, 10000 datasets, 50000 datasets respectively. Initially it uses without sampling data and find accuracy, Recall, Precision, F1 score. Then using same data set use different class imbalance algorithm like Random under sampling, Oversampling SMOTE(OS-SMOTE), and our

proposed Frequent Relational Synthetic over sampling Technique (FR-SMOTE) and compare with accuracy, Recall, Precision, F1 score. Finally found FRSMOTE give better performance.

In Fig. 2(e) 3(e) 4(e) GBM algorithm in the 5000 datasets, 10000 datasets, 50000 datasets respectively. Initially it uses without sampling data and find accuracy, Recall, Precision, F1 score. Then using same data set use different class imbalance algorithm like Random under sampling, Oversampling SMOTE(OS-SMOTE), and our proposed Frequent Relational Synthetic over sampling Technique (FR-SMOTE) and compare with accuracy, Recall, Precision, F1 score. Finally found FRSMOTE give better performance.

## V. Conclusion

This paper presented the FR-SMOTE algorithm for the class imbalance problem to find practical training in oversampling solutions using the synthetic method on imbalanced data distributions. It may be viewed as a data-level approach to class imbalance because it generates synthetic nodes that may be used to equalize the training dataset to balance any unbalanced dataset. In this context, FR-SMOTE met two critical features of a successful class imbalanced handling algorithm: the creation of synthetic node balances the minority class and generate high accuracy without boosting and bagging. In ensemble learning if we implement this algorithms, it can give the better result than the only FR-SMOTE.
Our next attempt will be to improve FR-SMOTE's basis of instance level challenges, allowing it to better handle problematic regions of the future data. [41]

## References

[1]. H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on knowledge and data engineering, vol. 21, no. 9, pp. 1263–1284, 2009.

[2]. M. Sowjanya and O. Mrudula, "Effective treatment of imbalanced datasets in health care using modified smote coupled with stacked deep learning algorithms," Applied Nanoscience, pp. 1–12, 2022.

[3]. Yang, C. Heiselman, J. G. Quirk, and P. M. Djuric, "Class-´ imbalanced classifiers using ensembles of gaussian processes and gaussian process latent variable models," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 3775–3779.

[4]. V. Lopez, S. Del R´ ´ıo, J. M. Ben´ıtez, and F. Herrera, "Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data," Fuzzy Sets and Systems, vol. 258, pp. 5–38, 2015.

[5]. Seliya, T. M. Khoshgoftaar, and J. Van Hulse, "A study on the relationships of classifier performance metrics," in 2009 21st IEEE international conference on tools with artificial intelligence. IEEE, 2009, pp. 59–66.

[6]. R. Longadge and S. Dongre, "Class imbalance problem in data mining review," arXiv preprint arXiv:1305.1707, 2013.

[7]. X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in 2008 Fourth international conference on natural computation, vol. 4. IEEE, 2008, pp. 192–201.

[8]. S. Guo, S. Wang, M. Wei, R. Chen, C. Guo, and H. Li, "Combining imbalance learning strategy and multiclassifier estimator for bug report classification," Mathematical Problems in Engineering, vol. 2020, 2020.

[9]. M. Waqar, H. Dawood, H. Dawood, N. Majeed, A. Banjar, and R. Alharbey, "An efficient smote-based deep learning model for heart attack prediction," Scientific Programming, vol. 2021, 2021.

[10]. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321–357, 2002.

[11]. S. Picek, A. Heuser, A. Jovic, S. Bhasin, and F. Regazzoni, "The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations," IACR Transactions on Cryptographic Hardware and Embedded Systems, vol. 2019, no. 1, pp. 1–29, 2019.

[12].Dablain, B. Krawczyk, and N. V. Chawla, "Deepsmote: Fusing deep learning and smote for imbalanced data," IEEE Transactions on Neural Networks and Learning Systems, 2022.

[13].C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5375–5384.

[14].M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," International Journal of Machine Learning and Computing, vol. 3, no. 2, p. 224, 2013.

[15].S. Ertekin, J. Huang, and C. L. Giles, "Active learning for class imbalance problem," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 823–824.

[16].C. G. Weng and J. Poon, "A new evaluation measure for imbalanced datasets," in Proceedings of the 7th Australasian Data Mining Conference-Volume 87, 2008, pp. 27–32.

[17].J. Zhu and E. Hovy, "Active learning for word sense disambiguation with methods for addressing the class imbalance problem," in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007, pp. 783–790.

[18].S. Pouyanfar, Y. Tao, A. Mohan, H. Tian, A. S. Kaseb, K. Gauen, R. Dailey, S. Aghajanzadeh, Y.-H. Lu, S.-C. Chen et al., "Dynamic sampling in convolutional neural networks for imbalanced data classification," in 2018 IEEE conference on multimedia information processing and retrieval (MIPR). IEEE, 2018, pp. 112–117.

[19].S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in 2016 international joint conference on neural networks (IJCNN). IEEE, 2016, pp. 4368–4374.

[20].J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," Journal of Big Data, vol. 6, no. 1, pp. 1–54, 2019.

[21].Ramentol, Y. Caballero, R. Bello, and F. Herrera, "Smote-rsb*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory," Knowledge and information systems, vol. 33, no. 2, pp. 245–265, 2012.

[22].H. Al Majzoub, I. Elgedawy, O. Akaydın, and M. K¨ ose Uluk¨ ok, "Hcab-¨ smote: A hybrid clustered affinitive borderline smote approach for imbalanced data binary classification," Arabian Journal for Science and Engineering, vol. 45, no. 4, pp. 3205–3222, 2020.

[23].S. Smiti and M. Soui, "Bankruptcy prediction using deep learning approach based on borderline smote," Information Systems Frontiers, vol. 22, no. 5, pp. 1067–1083, 2020.

[24].C. Gok and M. O. Olgun, "Smote-nc and gradient boosting imputa-¨ tion based random forest classifier for predicting severity level of covid19 patients with blood samples," Neural Computing and Applications, vol. 33, no. 22, pp. 15693–15707, 2021.

[25].Wang, Z. Luo, J. Huang, Y. Feng, and Z. Liu, "A novel ensemble method for imbalanced data learning: bagging of extrapolation-smote svm," Computational intelligence and neuroscience, vol. 2017, 2017.

[26].N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "Smoteboost: Improving prediction of the minority class in boosting," in European conference on principles of data mining and knowledge discovery. Springer, 2003, pp. 107–119.

[27].S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, "Fw-smote: A feature-weighted oversampling approach for imbalanced classification," Pattern Recognition, vol. 124, p. 108511, 2022.

[28].Bernardo and E. Della Valle, "An extensive study of c-smote, a continuous synthetic minority oversampling technique for evolving data streams," Expert Systems with Applications, vol. 196, p. 116630, 2022.

[29].J. Liu, "Importance-smote: a synthetic minority oversampling method for noisy imbalanced data," Soft Computing, vol. 26, no. 3, pp. 1141– 1163, 2022.

[30].Zhang, H. Yu, Z. Huan, X. Yang, S. Zheng, and S. Gao, "Smoterknn: A hybrid re-sampling method based on smote and reverse knearest neighbors," Information Sciences, vol. 595, pp. 70–88, 2022.

[31].J. Nanda and J. K. Chhabra, "Sshm: Smote-stacked hybrid model for improving severity classification of code smell," International Journal of Information Technology, pp. 1–7, 2022.

[32].S. Liu, "Smote-lmknn: A synthetic minority oversampling technique based on local means-based k-nearest neighbor," International Journal of Pattern Recognition and Artificial Intelligence, p. 2250019, 2022.

[33].He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE, 2008, pp. 1322–1328.

[34].T. Xu, G. Coco, and M. Neale, "A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning," Water research, vol. 177, p. 115788, 2020.

[35].Z. Hu, L. Wang, L. Qi, Y. Li, and W. Yang, "A novel wireless network intrusion detection

method based on adaptive synthetic sampling and an improved convolutional neural network," IEEE Access, vol. 8, pp. 195741–195751, 2020.
[36].J. T. Hancock and T. M. Khoshgoftaar, "Catboost for big data: an interdisciplinary review," Journal of big data, vol. 7, no. 1, pp. 1–45, 2020.
[37].T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen et al., "Xgboost: extreme gradient boosting," R package version 0.4-2, vol. 1, no. 4, pp. 1–4, 2015.
[38].D. Wang, Y. Zhang, and Y. Zhao, "Lightgbm: an effective mirna classification method in breast cancer patients," in Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics, 2017, pp. 7–11.
[39].M. Pal, "Random forest classifier for remote sensing classification," International journal of remote sensing, vol. 26, no. 1, pp. 217–222, 2005.
[40].D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, Logistic regression. Springer, 2002.
[41].Taherkhani, G. Cosma, and T. M. McGinnity, "Adaboost-cnn: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning," Neurocomputing, vol. 404, pp. 351–366, 2020.