IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON RELIABILITY IN SENSOR-CLOUD SYSTEMS AND APPLICATIONS (SCSA)

Received May 3, 2021, accepted June 4, 2021, date of publication July 6, 2021, date of current version July 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3095145

Treating Class Imbalance in Non-Technical Loss Detection: An Exploratory Analysis of a Real Dataset

KHAWAJA MOYEEZULLAH GHORI[1,2], MUHAMMAD AWAIS[3], (Member, IEEE), AKMAL SAEED KHATTAK[4], MUHAMMAD IMRAN[5], (Member, IEEE), FAZAL-E-AMIN[6], (Senior Member, IEEE), AND LASZLO SZATHMARY[7]

[1]Department of Computer Science, National University of Modern Languages (NUML), Islamabad 44000, Pakistan
[2]Doctoral School of Informatics, University of Debrecen, 4002 Debrecen, Hungary
[3]Department of Computer Science, Edge Hill University, Ormskirk L39 4QP, U.K.
[4]Department of Computer Sciences, Quaid-i-Azam University, Islamabad 15320, Pakistan
[5]College of Applied Computer Science, King Saud University, Riyadh 11451, Saudi Arabia
[6]Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[7]Department of IT, Faculty of Informatics, University of Debrecen, 4002 Debrecen, Hungary

Corresponding author: Muhammad Imran (cimran@ksu.edu.sa)

This work was supported in part by the European Union and the European Social Fund under Project EFOP-3.6.1-16-2016-00022, and in part by the Deanship of Scientific Research at King Saud University through Research Group under Project RG-1441-490.

ABSTRACT Non-Technical Loss (NTL) is a significant concern for many electric supply companies due to the financial impact caused as a result of suspect consumption activities. A range of machine learning classifiers have been tested across multiple synthesized and real datasets to combat NTL. An important characteristic that exists in these datasets is the imbalance distribution of the classes. When the focus is on predicting the minority class of suspect activities, the classifiers' sensitivity to the class imbalance becomes more important. In this paper, we evaluate the performance of a range of classifiers with under-sampling and over-sampling techniques. The results are compared with the untreated imbalanced dataset. In addition, we compare the performance of the classifiers using penalized classification model. Lastly, the paper presents an exploratory analysis of using different sampling techniques on NTL detection in a real dataset and identify the best performing classifiers. We conclude that logistic regression is the most sensitive to the sampling techniques as the change of its recall is measured around 50% for all sampling techniques. While the random forest is the least sensitive to the sampling technique, the difference in its precision is observed between $1\% - 6\%$ for all sampling techniques.

INDEX TERMS Class imbalance, non-technical loss detection, sampling techniques, under-sampling, over-sampling, cost-sensitive learning.

I. INTRODUCTION

Typically, many companies from the energy sector face financial crises due to Non-Technical Loss (NTL). It is a loss that is endured by the electric supplier and caused by the unusual suspect activities from the electric consumers. The suspect activities include illegal hooking of the wires, incorrect meter reading, meter bypassing, or even reversing the meters. The objective of these activities is the reduction of the bill amount. These activities are mainly practiced in those cities and industrial areas where manual infrastructure is still used. The Advanced Metering Infrastructure (AMI) has made many illegal activities hard to practice. However, many countries use manual electricity infrastructure, and hence, the monthly manual meter reading is still practiced in countries, including India, Pakistan, Brazil, etc. The multibillion-dollar annual loss is reported from such countries. For e.g., a yearly loss of 12 billion dollars is estimated for India due to the occurrences of NTL. An estimated loss of 58.7 billion dollars occurs every year on account NTL in power industries of the top 50 emerging countries [1], which shows the significance of NTL detection.

To combat the NTL, many techniques have been tested over the past decade to successfully detect NTL occurrences

The associate editor coordinating the review of this manuscript and approving it for publication was Md Zakirul Alam Bhuiyan.

and analyze the measures to avoid the losses incurred by NTL. The network-oriented, data-oriented, and hybrid techniques are commonly tested for this purpose [2]. The network-oriented techniques include the installation of separate hardware for the detection of NTL. The data-oriented techniques use the consumption data to identify the occurrences of NTL. Data-oriented techniques mainly focus on applying several machine learning classifiers to a pre-processed consumption data, including Support Vector Machine (SVM), Decision Trees (DT), K- Nearest Neighbors (KNN), CatBoost, XGBoost, LightGBoost, etc. These techniques use various performance evaluation metrics to measure the number of potential fraudsters detected correctly. In our previous contribution [2], we have identified the best metrics that can be used to evaluate the performance of the classifiers considering the characteristics of the datasets used in NTL detection. In another contribution [3], we have identified the best individual classifier and the best type of classifiers that outperformed others in NTL detection.

Many datasets which are used in machine learning have a problem of class imbalance. It is the characteristic of the dataset where the samples of one class are heavily represented while the samples of the other class are least represented. As reported in [4], due to this imbalanced distribution of the classes, a biases in the dataset are observed, resulting in a correct prediction of the majority class but an incorrect prediction of the minority class. The sensitivity of this issue is increased when the focus is in the identification of the minority class. NTL detection in the energy sector is also an application of the class imbalance domain. As the number of normal electric consumers outnumber the fraudsters, a dataset pertaining to NTL is characterized by the drawbacks of data bias, and hence, an imbalanced distribution of classes.

In order to deal with the problem of class imbalance in the datasets, multiple techniques are proposed in the literature. One of the techniques is under-sampling the majority class. An alternate is the over-sampling of the minority class [5]. Instance weighting schemes are also proposed in the literature to address the issue of class imbalance. Like other class imbalance problems, the datasets pertaining to NTL detection have also been tested in order to balance out the number of samples of the minority and the majority class in the pre-processing step. For this, under-sampling and over-sampling techniques have been tested separately. However, there is still a need to thoroughly explore the impact of using under-sampling, over-sampling, hybrid of both, and cost-sensitive approaches by applying them individually as well as in combination in a real dataset. In this work, we have used a real dataset of an electric supplier company operating in Pakistan. The dataset comprises 71 features and 80, 244 records. These are monthly meter readings of a specified neighborhood for a period of 15 months. The main contribution of the paper is to treat the class imbalance problem in NTL-oriented real datasets using a variety of different sampling techniques, such as under-sampling, over-sampling and penalized classification models. The dataset used in this study is collected in the real-life scenario by an electric supplier company in Pakistan. The dataset is then used to evaluate the performance of a range of classifiers. In the end, we present an exploratory analysis of the impact of using different sampling techniques on NTL detection in a real dataset and identify the best-performing classifiers.

The rest of the paper is as follows: Section II describes the literature review of the recent contributions in NTL detection. Section III presents the description of class imbalance problems. Section IV first describes the dataset used in this contribution, and it outlines different class imbalance methodologies and the description of the performance evaluation metrics used. Experiments are explained in Section V. Finally, conclusion and future work are presented in Section VI.

II. LITERATURE REVIEW

During the past decade, the financial deficit caused by electricity theft has been an alarming situation for many countries. Hence, the problem of NTL detection has become a thoroughly studied area. There are many sub-processes which the research community is trying to use to lessen the impact of NTL. For example, an increasing interest is found in using multiple classifiers, using multiple evaluation metrics to correctly figure out the losses, and the post-processing phase, etc. Multiple combinations of classifiers have been tested for the detection of NTL, which includes wide and deep CNN [6], recurrent neural network (RNN) [7], fuzzy logic [8], CatBoost [3], etc. Apart from classification, several other techniques have been tested, which include association rule mining [9], hierarchical clustering [10], outlier detection techniques [11], etc.

A. RECENT TRENDS IN PRE-PROCESSING STEPS OF NTL DETECTION

With the increase in recognition of the importance of NTL detection over the years, the research community has been equally showing an increasing interest in the pre-processing steps of the dataset before it is used in the training and the testing of the classifiers. The pre-processing steps include handling the imbalance characteristic of the datasets, feature selection, feature extraction, and data merging from the other sources. These recent trends in the pre-processing steps of NTL detection are described in Figure 1.

B. HANDLING CLASS IMBALANCE PROBLEM

Historically, the research community has shown a deep interest in tackling the imbalanced nature of the datasets. For this, generally, two sampling approaches are followed, namely under-sampling and over-sampling. These techniques are defined in Section IV-B1. The use of the synthetic minority over-sampling technique (SMOTE) [5] remains an attraction in many fields. This technique generates synthetic records using the records of the minority class.

There have been some notable contributions in this regard. For example, Buzau *et al.* [12] have used two under-sampling

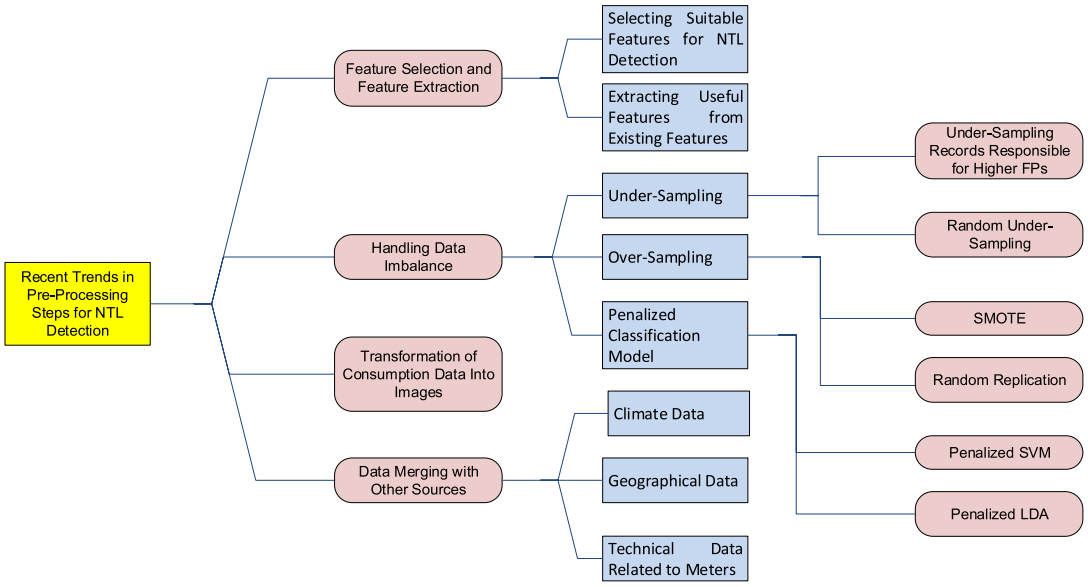

FIGURE 1. Recent trends in the pre-processing steps of NTL detection.

techniques to combat the imbalanced distribution of the classes. The first technique removes those normal consumers from the training set due to which the fraudulent consumers were wrongly identified as normal consumers. The second technique under-samples the training set randomly. The paper observers that there is not much of a difference in the output of the two under-sampling techniques. In addition, this contribution compares the results of SVM, logistic regression, KNN, and XGBoost using AUC (Area Under the Curve) as the performance evaluation metric. A dataset from a Spanish company is used in the experiments. In addition, this contribution has also included some geographical information and the technical information of the meters. However, the paper has not used any of the over-sampling techniques to compare the results of under and over-sampling.

In contrast, Hasan *et al.* [13] have used SMOTE as an over-sampling technique in the pre-processing step to overcoming the impact of the imbalanced distribution of the classes. The authors have used a dataset provided by an electric supplier in China. They have used long short-term memory (LSTM) and Convolutional Neural Network (CNN) to identify NTL. The performance evaluation measures used are precision, recall, F-1, and accuracy. The paper reports an accuracy of 89%. A somewhat different approach is used in [14]. The authors have combined the data of power and voltage measurements with the normal consumption data to train and test the SVM classifier to detect the faulty consumers and their time and the intensity of NTL in kW. This work has used an Irish dataset that includes half-hourly consumption records for 5000 consumers. The authors have used replication as an over-sampling technique to overcome the problem of class imbalance. The results are evaluated using AUC and accuracy. The paper observes an accuracy of 99.4%. However, this work has not performed any under-sampling techniques.

The same dataset is used in [15], but the paper focuses on detecting NTL in industrial supplies only. The authors have used a deep learning-based mechanism to extract some advanced features as a pre-processing step from the dataset, which are then used in a semi-supervised autoencoder for theft detection. Their results are compared with SVM, KNN [16], XGBoost [17], and multi-layer perceptron (MLP) using precision, recall, F-1, AUC score, and accuracy. The paper concludes that their proposed framework outperformed the other available classifiers. However, the article has not addressed the imbalance behavior in the pre-processing step. In our recent contribution [18], we have proposed an incremental feature selection algorithm that helps in selecting the minimum number of suitable features for NTL detection. The algorithm uses the feature importance [19] of every feature. The work has identified the top 9 features out of a total of 71 features in a real dataset. The precision, recall, and F-1 scores of the classifiers using selected features are comparable or better than all features. The classifiers used are CatBoost, KNN, and decision tree.

Another contribution using a dataset from a Chinese company is presented in [20]. The authors have used the time-series data to convert it into image form, which is helpful in the long run for analyzing the consumer's consumption behavior. The paper evaluates its results using precision, recall, F-1 score, and AUC curve and concludes that their work performs best when the labeled classes are few in the dataset. The imbalance behavior of the dataset is, however, not discussed in the paper.

A similar dataset from China is used in [21]. The authors have not dealt with the imbalance behavior of the data. However, they have taken into consideration the effect of climatic changes in the occurrence of NTL. The authors have combined the electric data with the climate data and concluded that the occurrence of NTL is more in the regions of

extreme weather. They have used different classifiers belonging to the neural networks (NN) and ensemble methods. The paper concludes that the ensemble methods perform better than NN.

Recent work in NTL detection [22] uses a binary masking scheme in the pre-processing step to fill the missing values. However, the paper has not dealt with the imbalance behavior of the data. This work has used CNN as a base classifier and evaluated the performance using AUC and F-1 scores. The paper concludes with a higher F-1 score as compared to other techniques.

There is a need to further explore the pre-processing steps in NTL detection by a comprehensive study comparing the effect of using the under-sampling and over-sampling approaches on a real imbalance dataset. A summary of the literature review is present in Table 1.

III. THE PROBLEM OF CLASS IMBALANCE

The imbalanced distribution of the classes is typical behavior in classification problems ([23], [24]). This class imbalance occurs in datasets having a disproportionate ratio of instances or examples in each class. In other words, it is caused by a skewed distribution of data between classes. Many real-life problems such as fake review detection, fake news detection, fraud detection, customer churn prediction, electricity loss prediction, and others appearing in different domains are prone to imbalanced data. Most classifiers are sensitive to real-life imbalanced data and suffer from achieving accurate results because state-of-the-art classification algorithms expect balanced class distribution ([25]–[28]).

The characteristic of class imbalance exhibits the difference in the class distribution in the training and the test set. In contrast, the conditional distribution of X in the training set is the same as the conditional distribution of X in the test set given the same class label. Let X be the observation sample, and Y be the target variable, then the class imbalance relation is shown in Equations 1 and 2 [29]:

$$P_{train}(X|Y = y) = P_{test}(X|Y = y) \quad (1)$$

$$P_{train}(Y) \neq P_{test}(Y) \quad (2)$$

where $P_{train}(X|Y = y)$ is the distribution of X given Y in the training set, $P_{test}(X|Y = y)$ is the distribution of X given Y in the test set, $P_{train}(Y)$ is the distribution of the target variable in the training set and $P_{test}(Y)$ is the distribution of the target variable in the test set.

IV. METHODOLOGY

A. DATASET

Many countries use the traditional metering infrastructure, which uses the on-site meter readings every month. For this, a monthly consumption record is entered in the database for each consumer. The dataset used in this contribution is a real dataset taken from a power supply company in Pakistan. This dataset also contains monthly readings of the electricity consumption of a neighborhood. The dataset contains 71 features

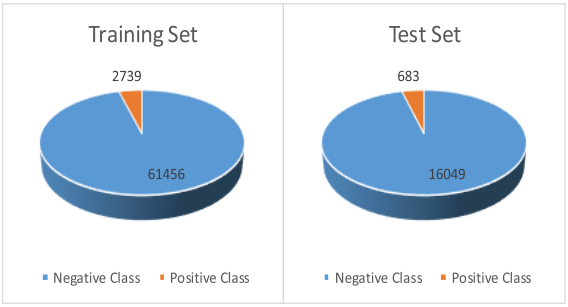

FIGURE 2. Class imbalance problem.

which include the numeric, string, and date data types. The total number of records used is 80, 244. The company also provided the records of consumption where NTL has been identified, due to which this is a labeled dataset where the classes are already known. The dataset is split into training and test sets with an 80% − 20% ratio. The training set contains 61, 456 records with negative class and 2, 739 records with positive class. Similarly, the test set contains 15, 366 records with negative class and 683 records with positive class. As Figure 2 depicts, a clear imbalance in the representation of the two classes in the training and the test sets is observed, making this dataset a perfect example for exploring the impact of class imbalance in NTL detection.

B. CLASS IMBALANCE METHODOLOGIES

To achieve high performance of the machine learning classifiers in the imbalanced datasets, care must be taken to balance out the representation of all the classes before the dataset is used for classification. The techniques to deal with the class imbalance problem belong to two main types. One type deals with the situation in the data-level phase. The second type, also termed cost-sensitive learning, deals with the imbalance problem at the algorithmic level [30].

1) DATA-LEVEL TECHNIQUES

In the pre-processing phase, the following three techniques are widely used:

1) *Under-Sampling Techniques:* These techniques attempt to reduce the number of samples of the majority class. The moist widely used under-sampling method is random under-sampling [31]. The drawback of using the under-sampling techniques is the loss of potentially useful records. However, the training time of the classifiers is improved with the reduction of the training set.
2) *Over-Sampling Techniques:* These techniques attempt to generate synthetically new samples or randomly duplicate the existing samples of the minority class. The synthetic generation of the minority class is termed SMOTE [5]. The drawback of over-sampling techniques is the increase in the size of the training set, and hence, an increase in the computation time of training the classifiers. However, in contrast with the under-sampling techniques, the over-sampling

TABLE 1. Summary of literature review.

S. #	Article	Sampling Tech.	Dataset	Classifiers	Metric	Deficiency
1	[12]	Under-Sampling	Synthesized, Spanish	SVM, LR, XGBoost, and KNN	AUC	i. No over-sampling ii. No comparison of sampling techniques
2	[13]	SMOTE	Real, Chinese	LSTM, CNN	Precision, Recall, F-1, and accuracy	i. No under-sampling ii.No comparison of sampling techniques
3	[14]	Over-Sampling	Synthesized, Irish	SVM	AUC and accuracy	i. No under-sampling ii. No comparison of sampling techniques
4	[15]	DL for advanced feature extraction	Synthesized, Irish	Semi- Supervised autoencoder	Precision, Recall, F-1, AUC, and accuracy	Imbalance behavior has not been addressed
5	[18]	Useful feature extraction algorithm	Real, Pakistan	CatBoost, KNN, and decision trees	Precision, Recall, and F-1	Imbalance behavior has not been addressed
6	[22]	Binary masking scheme	Real, Chinese	CNN	AUC and F-1	Imbalance behavior has not been addressed

techniques do not suffer the drawback of the loss of potentially useful records.

3) *Hybrid Techniques:* These techniques combine the under-sampling and the over-sampling techniques in the pre-processing step. Hybrid techniques tend to combine the benefits of using both the over-sampling and the under-sampling techniques.
4) *ADASYN:* Adaptive synthetic sampling technique [32] uses different weights for different minority records depending on the difficulty level of the records. For those records that are harder to learn, more synthetic examples are generated as compared to the records that are easier to learn. As a result, the decision boundary is shifted towards the difficult examples.

2) ALGORITHMIC-LEVEL TECHNIQUES

One of the schemes to counter the biases in the dataset is cost-sensitive learning (also termed as penalized classification model). This technique can be applied at the algorithmic level as well as at the data level. It assigns the weights (cost) to the training observation, which is responsible for the miss-classification. For example, if the ratio of class imbalance is 1 : 10 in favor of the negative class, then the cost of miss-classification of the positive class will be nine times as compared to the cost of miss-classification of the negative class [28]. These costs can be calculated for every observation using Equation 3 [4]:

$$w = \frac{P_{test}(Y)}{P_{train}(Y)} \tag{3}$$

where $P_{train}(Y)$ is the distribution of the target variable in the training set and $P_{test}(Y)$ is the distribution of the target variable in the test set. As the dataset that we have used is an application of class imbalance where the normal users are too many compared to the number of fraudsters, we have tested under-sampling, over-sampling and cost-sensitive technique to counter the class imbalance problem in our dataset.

C. CLASSIFICATION METHODS FOR NTL DETECTION

As discussed in Section II, many classification methods have been used to detect NTL in the electric power industry. The current work extends and advances our previous contributions ([2], [3]), where the same dataset has been used. The dataset is described in detail in Section IV-A. The earlier contributions ([2], [3]) identified the best performing classifiers and highlighted the suitable performance metrics that can be utilized to investigate NTL efficiently and effectively. In this work, we have analyzed the effect in the performance of nine classifiers after applying different sampling techniques and compared their performance with the performance of the original, untreated class imbalance dataset. The untreated class imbalance dataset is the one in which no sampling technique has been applied in the pre-processing step and original distribution of positive and negative classes is utilized, as shown in Fig. 2. SVM is well known for maximizing the boundaries between the classes. It shows good performance when used with high-dimensional datasets. One of the linear learning models used is Stochastic Gradient Descent (SGD). Sensitive to scaling, SGD also shows good results under high-dimensional data. Despite having the weakness of overfitting, a Decision Tree (DT) is a good option for some datasets due to its simple method of constructing if-else rules.

Random Forest (RF) is an ensemble of different decision trees. Overfitting is avoided in RF by using various training sets for each DT [33]. The classification of each instance in KNN is performed by majority voting in k-Nearest

TABLE 2. A comparison of confusion matrices of classifiers selected based on maximum recall using imbalance and balance dataset.

(a) Imbalance Dataset							
MLP	Predicted as ->	Theft	Normal	**CatBoost**	Predicted as ->	Theft	Normal
ActualClass	Theft	679	4	ActualClass	Theft	677	6
	Normal	18	15348		Normal	17	15349
XGBoost	Predicted as ->	Theft	Normal	**RF**	Predicted as ->	Theft	Normal
ActualClass	Theft	675	8	ActualClass	Theft	671	12
	Normal	15	15351		Normal	16	15350
(b) SMOTE							
CatBoost	Predicted as ->	Theft	Normal	**MLP**	Predicted as ->	Theft	Normal
ActualClass	Theft	680	3	ActualClass	Theft	682	1
	Normal	20	15346		Normal	36	15330
KNN	Predicted as ->	Theft	Normal	**SGD**	Predicted as ->	Theft	Normal
ActualClass	Theft	683	0	ActualClass	Theft	681	2
	Normal	36	15330		Normal	39	15327
(c) ADASYN							
CatBoost	Predicted as ->	Theft	Normal	**LogReg**	Predicted as ->	Theft	Normal
ActualClass	Theft	683	0	ActualClass	Theft	683	0
	Normal	22	15344		Normal	39	15327
KNN	Predicted as ->	Theft	Normal	**XGBoost**	Predicted as ->	Theft	Normal
ActualClass	Theft	683	0	ActualClass	Theft	681	2
	Normal	38	15328		Normal	50	15316
(d) Random Over-Sampling							
CatBoost	Predicted as ->	Theft	Normal	**XGBoost**	Predicted as ->	Theft	Normal
ActualClass	Theft	680	3	ActualClass	Theft	683	0
	Normal	29	15337		Normal	34	15332
KNN	Predicted as ->	Theft	Normal	**MLP**	Predicted as ->	Theft	Normal
ActualClass	Theft	681	2	ActualClass	Theft	683	0
	Normal	34	15332		Normal	39	15327
(e) Random Under-Sampling							
SVM	Predicted as ->	Theft	Normal	**MLP**	Predicted as ->	Theft	Normal
ActualClass	Theft	680	3	ActualClass	Theft	683	0
	Normal	36	15330		Normal	38	15328
SGD	Predicted as ->	Theft	Normal	**KNN**	Predicted as ->	Theft	Normal
ActualClass	Theft	681	2	ActualClass	Theft	683	0
	Normal	42	15324		Normal	55	15311

Neighbors. Care must be taken in setting the value of k, as with the increased value of k, the training time of the dataset also gets increased. Incorporated with multiple hidden layers, Multi-Layer Perceptron (MLP) is heavily used in classification problems. Although it requires several hyperparameters to be tuned, its added advantage of non-linear compatibility remains an attraction. The only difference between MLP and Logistic Regression (LR) is that LR contains only one intermediate layer between the input and the output layer. CatBoost and XGBoost are the boosting techniques, with CatBoost having the flexibility that it handles the categorical data on its own while for XGBoost all data is needed to be converted to numerical datatype [34].

D. PERFORMANCE METRICS

Three performance evaluation metrics are chosen to evaluate the classifiers, namely precision, recall, and F-score. Precision measures correctly classified True Positive (TP) instances out of the total predicted TP instances. The formula of precision is shown in Equation 4.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall gives a measure of correctly predicted TP samples out of total predicted samples. The formula for recall is shown in Equation 5.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

F-score is the metric used to prioritize between recall and precision. The formulae for F-score is shown in Equation 6.

$$F - score = \frac{\left(1 + \beta^2\right) Recall \times Precision}{\beta^2 \times Precision + Recall} \tag{6}$$

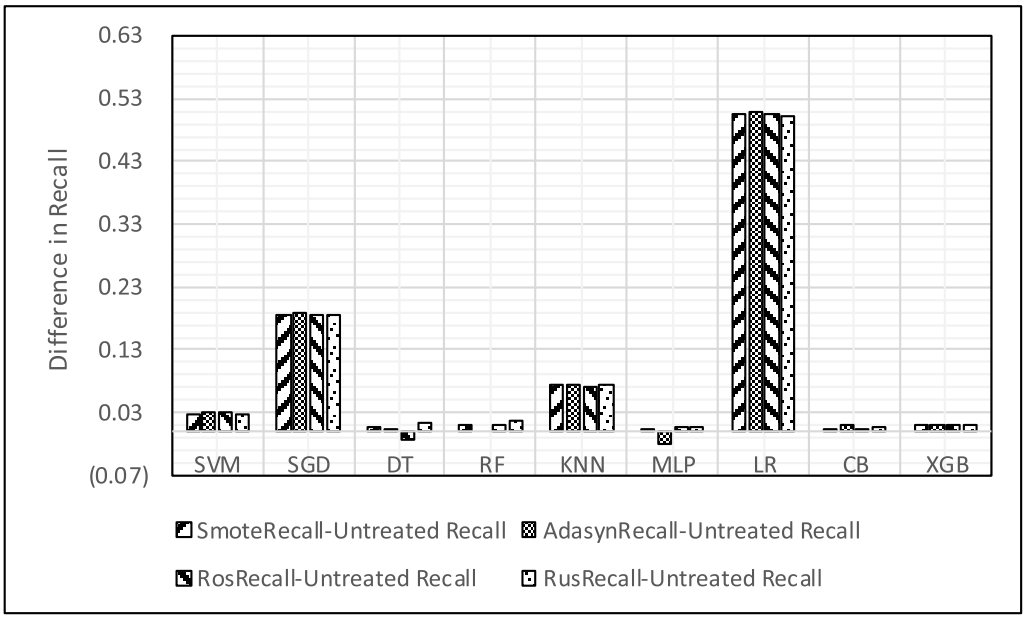

FIGURE 3. Recall.

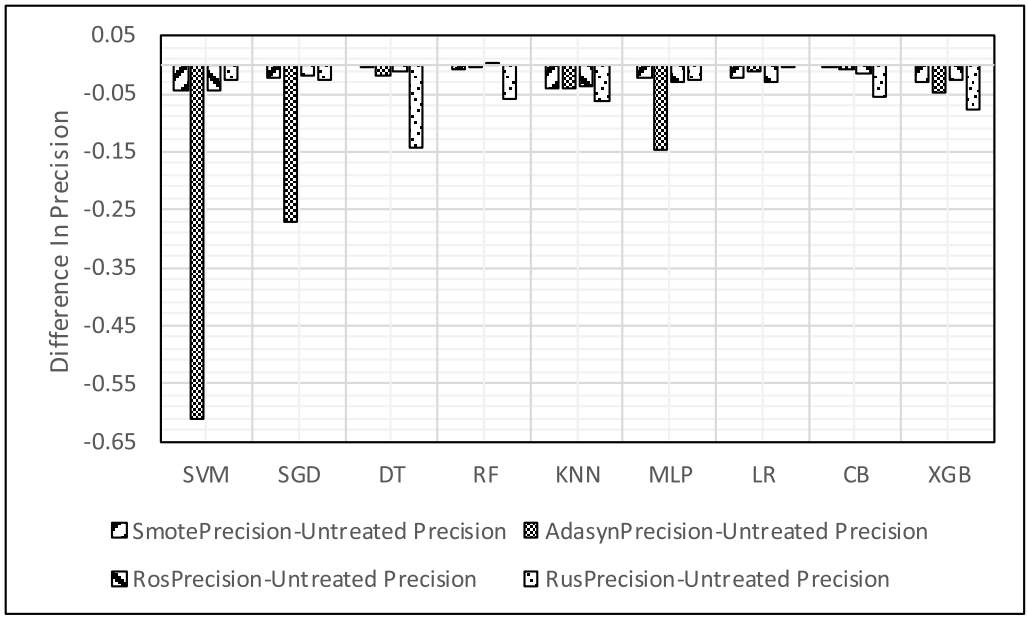

FIGURE 4. Precision.

When β is 0.5, recall and precision have equal priorities. When β is greater than 0.5, recall has the higher priority and when β is less than 0.5, precision has the higher priority.

As cited in our previous contribution [2], it is necessary to prioritize between FN and FP for NTL detection. Having a higher FP will result in an additional cost of on-site checking for NTL occurrence, but having a higher FN will directly affect the identification of NTL. We need the number of FN as low as possible. There is an indirect relation between FN and recall, i.e., with the decrease of FN, there is an increase in the recall. This leads to an interesting conclusion about NTL detection that as we need a lower FN, so, the classifier with a higher recall should be preferred.

V. RESULTS AND DISCUSSION

A detailed result containing TP, TN, FP, FN, precision, recall, and F-Score of nine classifiers tested with imbalance data and different sampling techniques is presented in Table 3 of Appendix VI. A comparison between confusion matrices of selected classifiers is shown in Table 2. Four classifiers with the best recall are chosen from each category of imbalance data, SMOTE, ADASYN, random over-sampling and random under-sampling. As discussed in Section IV-D, the classifier with the highest recall should be given priority for NTL detection. Considering this factor, with the original, untreated class imbalance dataset, MLP classifier, CatBoost and XGBoost has the best recall of 0.99

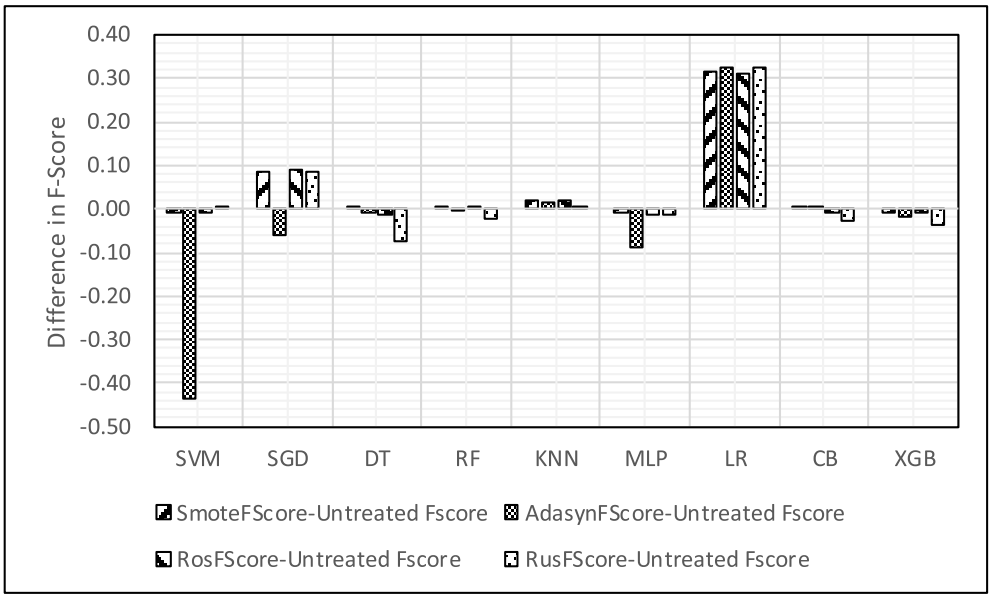

FIGURE 5. FScore.

each. Their corresponding numbers for FN are 4, 6, and 8, respectively.

A. IMPACT OF SAMPLING TECHNIQUES ON IMBALANCE DATA

Detailed comparison of various confusion matrices obtained from the set of classifiers achieving maximum recall using balanced and imbalanced datasets is presented in Table 2. It is pretty evident from the table that the recall of KNN and XGBoost jumps to 1 when applied to the sampled dataset obtained through SMOTE since the corresponding number for false negatives is reduced to zero. The best classifiers identified after applying ADASYN are Catboost, KNN, LR, and XGBoost. Considering recall, the most sensitive classifiers after applying random over-sampling are CatBoost, KNN, XGBoost, and MLP. Finally, the top four sensitive classifiers for random under-sampling are SVM, SGD, MLP, and KNN.

The classifier that is most sensitive to the sampling techniques is LR. As shown in Figure 3, the difference in LR recall is around 50% for SMOTE, ADASYN, ROS, and RUS. The next sensitive classifier found is SGD, in which the difference of recall is 0.23% for all the sampling techniques. One of the reasons behind this increase is the low recall of LR and SGD in imbalance data, which is 0.49 and 0.81. A difference of 8% is observed in the recall of KNN in all sampling techniques. KNN is susceptible to imbalance class distribution because it classifies an instance by a majority vote amongst the k-nearest neighbors. That's why KNN suffers in getting accurate results when it comes to dealing with imbalanced data. The recall is improved once the imbalance distribution of classes is removed by the sampling techniques.

A significant decrease of 63% in the precision of SVM is observed when ADASYN is applied. The top three percent decrease in the accuracy of the classifiers for ADASYN in SVM, SGD, and MLP, as depicted in Figure 4. Considering precision, the least sensitive classifier to sampling techniques is RF, which has a decrease of 1%, 0%, 0%, and 6% for SMOTE, ADASYN, ROS, and RUS, respectively.

As F-score evaluates the classifier with respect to both the precision and the recall, the most sensitive classifier for sampling techniques is LR, which has an increase of around $30\% - 34\%$ for all sampling techniques, as shown in Figure 5. In contrast, a decrease of 44.9% in F-score is observed for SVM when ADASYN was applied. Considering F-score, the top two insensitive classifiers concerning sampling techniques are RF and CB.

Considering recall as a performance metric and relevant measure, RUS is the best sampling technique. Eight out of nine classifiers resulted in a recall of 1 after applying RUS, while LR resulted in a recall of 0.99. On the contrary, as shown in Figure 4, a noticeable decrease in precision is observed in those classifiers after applying RUS hinting at the increase of FP. However, it is important to note that these findings are coupled with the specific dataset used in this study. The sampling methods implemented in this study might behave differently from our results, considering the nature of the dataset under investigation.

The precision of imbalanced data remains the highest compared to all sampling techniques because sampling techniques try to contract the decision boundaries. In doing so, many values which were treated as FN in imbalance data are treated as FP, resulting in the decrease of precision.

B. IMPACT OF COST-SENSITIVE LEARNING ON IMBALANCE DATA

In our experiments, we also performed cost-sensitive learning by applying a weighted SVM strategy. Every miss-classification resulted in penalizing the training samples, which are responsible for miss-classification. The recall

TABLE 3. Experimental results.

Sampling Techniques	Classifiers	TP	TN	FP	FN	Precision	Recall	F-Score
Imbalance Data (No Sampling)	SVM	661	15350	16	22	0.98	0.97	0.97
	SGD	554	15348	18	129	0.97	0.81	0.88
	DT	670	15347	19	13	0.97	0.98	**0.98**
	RF	671	15350	16	12	0.98	0.98	**0.98**
	KNN	632	15359	7	51	**0.99**	0.93	0.96
	MLP	679	15348	18	4	0.97	**0.99**	**0.98**
	LogReg	335	15351	15	348	0.96	0.49	0.65
	CatBoost	677	15349	17	6	0.98	**0.99**	**0.98**
	XGBoost	675	15351	15	8	0.98	**0.99**	**0.98**
SMOTE	SVM	679	15317	49	4	0.93	0.99	0.96
	SGD	681	15327	39	2	0.95	**1.00**	0.97
	DT	674	15346	20	9	**0.97**	0.99	**0.98**
	RF	677	15344	22	6	**0.97**	0.99	**0.98**
	KNN	683	15330	36	0	0.95	**1.00**	0.97
	MLP	682	15330	36	1	0.95	**1.00**	0.97
	LogReg	680	15317	49	3	0.93	**1.00**	0.96
	CatBoost	680	15346	20	3	**0.97**	**1.00**	**0.98**
	XGBoost	683	15330	36	0	0.95	**1.00**	0.97
ADASYN	SVM	683	14182	1184	0	0.37	**1.00**	0.54
	SGD	683	15072	294	0	0.70	**1.00**	0.82
	DT	671	15333	33	12	0.95	0.98	0.97
	RF	671	15348	18	12	**0.97**	0.98	**0.98**
	KNN	683	15328	38	0	0.95	**1.00**	0.97
	MLP	666	15227	139	17	0.83	0.98	0.90
	LogReg	683	15327	39	0	0.95	**1.00**	0.97
	CatBoost	683	15344	22	0	**0.97**	**1.00**	**0.98**
	XGBoost	681	15316	50	2	0.93	**1.00**	0.96
Random Over-Sampling	SVM	681	15317	49	2	0.93	**1.00**	0.96
	SGD	681	15329	37	2	0.95	**1.00**	0.97
	DT	660	15339	27	23	0.96	0.97	0.96
	RF	677	15350	16	6	0.98	0.99	**0.98**
	KNN	681	15332	34	2	0.95	**1.00**	0.97
	MLP	683	15327	39	0	0.95	**1.00**	0.97
	LogReg	680	15314	52	3	0.93	**1.00**	0.96
	CatBoost	680	15337	29	3	0.96	**1.00**	**0.98**
	XGBoost	683	15332	34	0	0.95	**1.00**	**0.98**
Random Under-Sampling	SVM	680	15330	36	3	**0.95**	**1.00**	**0.97**
	SGD	681	15324	42	2	0.94	**1.00**	**0.97**
	DT	680	15225	141	3	0.83	**1.00**	0.90
	RF	682	15306	60	1	0.92	**1.00**	0.96
	KNN	683	15311	55	0	0.93	**1.00**	0.96
	MLP	683	15328	38	0	**0.95**	**1.00**	**0.97**
	LogReg	678	15332	34	5	**0.95**	0.99	**0.97**
	CatBoost	682	15306	60	1	0.92	**1.00**	0.96
	XGBoost	682	15292	74	1	0.90	**1.00**	0.95
Cost-Sensitive Learning	Weighted SVM	682	15329	37	1	0.95	1.00	0.97

of weighted SVM is increased from 0.97 to 1.00 while the precision decreased from 0.98 to 0.95. The F-score remained stable at 0.97. Considering recall as the preferred performance evaluation metric for NTL detection, cost-sensitive learning resulted in an increase of 3% in SVM recall.

VI. CONCLUSION AND FUTURE WORK

In this paper, a real dataset of monthly consumption records was used for the experiments. The dataset includes approximately 80, 000 records, along with 71 features. The paper compares the performance analysis of 9 classifiers with four

sampling techniques applied on imbalanced data. The results are compared with the performance of untreated imbalance data. Additionally, the impact of cost-sensitive learning is also analyzed on imbalanced data for NTL detection.

One of the findings is that considering recall and logistic regression is the most sensitive classifier for all sampling techniques. The difference in the recall is observed around 0.50% for SMOTE, ADASYN, ROS, and RUS. A decrease of 63% is observed for SVM when ADASYN is applied. Another finding is that the top three percent decrease in the precision of the classifiers are for ADASYN in SVM, SGD and MLP, respectively. The random forest is observed as the least sensitive classifier with a percent decrease of $1\% - 6\%$ for all four sampling techniques.

The best sampling technique found is RUS, for which eight out of nine classifiers resulted in a recall of 1. Cost-sensitive learning was also applied by experimenting weighted SVM technique. Its recall increased from 0.97 to 1.00, while the precision decreased from 0.98 to 0.95. The F-score remained constant at 0.97.

In the future, we have a plan to explore the impact of sampling techniques on selected features for NTL detection. The features on which the class label is more dependent can be filtered out and tested for the sampling techniques. Another potential future direction for NTL detection is deep learning in the all-feature dataset compared with the selected features.

APPENDIX A
EXPERIMENTAL RESULTS

A detailed experimental results containing TP, TN, FP, FN, precision, recall and F-Score of selected classifiers using imbalance data and different sampling techniques are presented in Table 3.

REFERENCES

[1] G. Otchere-Appiah, S. Takahashi, M. S. Yeboah, and Y. Yoshida, "The impact of smart prepaid metering on non-technical losses in ghana," *Energies*, vol. 14, no. 7, p. 1852, Mar. 2021.

[2] K. M. Ghori, M. Imran, A. Nawaz, R. A. Abbasi, A. Ullah, and L. Szathmary, "Performance analysis of machine learning classifiers for non-technical loss detection," *J. Ambient Intell. Humanized Comput.*, pp. 1–16, Jan. 2020, doi: 10.1007/s12652-019-01649-9.

[3] K. M. Ghori, R. A. Abbasi, M. Awais, M. Imran, A. Ullah, and L. Szathmary, "Performance analysis of different types of machine learning classifiers for non-technical loss detection," *IEEE Access*, vol. 8, pp. 16033–16048, 2020.

[4] P. Glauner, P. Valtchev, and R. State, "Impact of biases in big data," 2018, *arXiv:1803.00897*. [Online]. Available: http://arxiv.org/abs/1803.00897

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[6] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1606–1615, Apr. 2018.

[7] S. Chatterjee, V. Archana, K. Suresh, R. Saha, R. Gupta, and F. Doshi, "Detection of non-technical losses using advanced metering infrastructure and deep recurrent neural networks," in *Proc. IEEE Int. Conf. Environ. Electr. Eng. IEEE Ind. Commercial Power Syst. Eur. (EEEIC/I&CPS Eur.)*, Jun. 2017, pp. 1–6.

[8] J. L. Viegas, P. R. Esteves, and S. M. Vieira, "Clustering-based novelty detection for identification of non-technical losses," *Int. J. Electr. Power Energy Syst.*, vol. 101, pp. 301–310, Oct. 2018.

[9] S. Singh and A. Yassine, "Big data mining of energy time series for behavioral analytics and energy consumption forecasting," *Energies*, vol. 11, no. 2, p. 452, Feb. 2018.

[10] C. C. S. Zuleta, U. de Medellín, J. P. F. Gutiérrez, C. C. P. Escobar, U. de Medellín, and U. de Medellín, "Identification of the characteristics incident to the detection of non-technical losses for two colombian energy companies," *Revista Facultad de Ingeniería Universidad de Antioquia*, no. 84, pp. 60–71, Sep. 2017.

[11] J. Yeckle and B. Tang, "Detection of electricity theft in customer consumption using outlier detection algorithms," in *Proc. 1st Int. Conf. Data Intell. Secur. (ICDIS)*, Apr. 2018, pp. 135–140.

[12] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gomez-Exposito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 2661–2670, May 2019.

[13] M. N. Hasan, R. N. Toma, A.-A. Nahid, M. M. M. Islam, and J.-M. Kim, "Electricity theft detection in smart grid systems: A CNN-LSTM based approach," *Energies*, vol. 12, no. 17, p. 3310, Aug. 2019.

[14] G. M. Messinis, A. E. Rigas, and N. D. Hatziargyriou, "A hybrid method for non-technical loss detection in smart distribution grids," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6080–6091, Nov. 2019.

[15] X. Lu, Y. Zhou, Z. Wang, Y. Yi, L. Feng, and F. Wang, "Knowledge embedded semi-supervised deep learning for detecting non-technical losses in the smart grid," *Energies*, vol. 12, no. 18, p. 3452, Sep. 2019.

[16] M. Awais, L. Palmerini, and L. Chiari, "Physical activity classification using body-worn inertial sensors in a multi-sensor setup," in *Proc. IEEE 2nd Int. Forum Res. Technol. Soc. Ind. Leveraging Better Tomorrow (RTSI)*, Sep. 2016, pp. 1–4.

[17] M. Raza, M. Awais, K. Ali, N. Aslam, V. V. Paranthaman, M. Imran, and F. Ali, "Establishing effective communications in disaster affected areas and artificial intelligence based detection using social media platform," *Future Gener. Comput. Syst.*, vol. 112, pp. 1057–1069, Nov. 2020.

[18] K. M. Ghori, A. R. Ayaz, M. Awais, M. Imran, A. Ullah, and L. Szathmary, "Impact of feature selection on non-technical loss detection," in *Proc. 6th Conf. Data Sci. Mach. Learn. Appl. (CDMA)*, Mar. 2020, pp. 19–24.

[19] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria," *Ann. Math. Artif. Intell.*, vol. 41, no. 1, pp. 77–93, May 2004.

[20] J. Li and F. Wang, "Non-technical loss detection with statistical profile images based on semi-supervised learning," 2019, *arXiv:1907.03925*. [Online]. Available: http://arxiv.org/abs/1907.03925

[21] W. Hu, Y. Yang, J. Wang, X. Huang, and Z. Cheng, "Understanding electricity-theft behavior via multi-source data," 2020, *arXiv:2001.07311*. [Online]. Available: http://arxiv.org/abs/2001.07311

[22] P. Finardi, I. Campiotti, G. Plensack, R. D. de Souza, R. Nogueira, G. Pinheiro, and R. Lotufo, "Electricity theft detection with self-attention," 2020, *arXiv:2002.06219*. [Online]. Available: http://arxiv.org/abs/2002.06219

[23] M.-C. Chen, L.-S. Chen, C.-C. Hsu, and W.-R. Zeng, "An information granulation based data mining approach for classifying imbalanced data," *Inf. Sci.*, vol. 178, no. 16, pp. 3214–3227, Aug. 2008.

[24] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[25] G. M. Weiss, "Learning with rare cases and small disjuncts," in *Machine Learning Proceedings*. Amsterdam, The Netherlands: Elsevier, 1995, pp. 558–565.

[26] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Mach. Learn.*, vol. 30, nos. 2–3, pp. 195–215, 1998.

[27] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data Min. Knowl. Discovery*, vol. 1, no. 3, pp. 291–316, Sep. 1997.

[28] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[29] J. Jiang, *A Literature Survey on Domain Adaptation of Statistical Classifiers*, vol. 3, 2008, pp. 1–12. [Online]. Available: http://sifaka.cs.uiuc.edu/jiang4/domainadaptation/survey

[30] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[31] M. A. Tahir, J. Kittler, K. Mikolajczyk, and F. Yan, "A multiple expert approach to the class imbalance problem using inverse random under sampling," in *Multiple Classifier Systems* (Lecture Notes in Computer Science), vol. 5519, J. A. Benediktsson, J. Kittler, and F. Roli, Eds. Berlin, Germany: Springer, 2009, doi: 10.1007/978-3-642-02326-2_9.
[32] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1322–1328.
[33] M. Awais, M. Raza, K. Ali, Z. Ali, M. Irfan, O. Chughtai, I. Khan, S. Kim, and M. Ur Rehman, "An Internet of Things based bed-egress alerting paradigm using wearable sensors in elderly care environment," *Sensors*, vol. 19, no. 11, p. 2498, May 2019.
[34] S. Rahman, M. Irfan, M. Raza, K. M. Ghori, S. Yaqoob, and M. Awais, "Performance analysis of boosting classifiers in recognizing activities of daily living," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, p. 1082, Feb. 2020.

KHAWAJA MOYEEZULLAH GHORI received the B.S. (CS) degree from International Islamic University (IIU), Islamabad, Pakistan, in 2002, and the M.S. (CS) degree from the FAST, National University of Computer and Emerging Sciences (NUCES), Islamabd, in 2004. He is currently pursuing the Ph.D. degree with the University of Debrecen, Hungary. He is currently an Assistant Professor with the Department of Computer Science, National University of Modern Languages, Islamabad. His research interests include data mining and machine learning. He is also working on the problem of detecting non-technical losses (NTL) in power sector.

MUHAMMAD AWAIS (Member, IEEE) received the B.S. degree in electronic engineering from Mohammad Ali Jinnah University, Pakistan, the M.S. degree in electrical and electronic engineering from Universiti Teknologi PETRONAS, Malaysia, and the Ph.D. degree in biomedical, electrical, and system engineering from the University of Bologna, Italy. He previously worked as a Research Fellow with the University of Hull, U.K., University of Leeds, U.K., and University of Bologna. He is currently a Senior Lecturer with the Department of Computer Science, Edge Hill University, U.K. His research interests include the domain of data mining, the Internet of Things, data analytics, signal processing, application-based machine learning and deep learning to develop ICT (information and communication technologies)-based systems for remote sensing, digital health, and industry 4.0. He is a member of the IEEE Engineering in Medicine and Biology Society (EMBS). He is a Reviewer of many well reputed journals, such as *Future Generation Computer Systems* (Elsevier), IEEE Journal of Biomedical and Health Informatics (J-BHI), IEEE Access, *IEEE Communication Magazine*, *Sensors* (MDPI), *JMIR*, and *CSSP*.

AKMAL SAEED KHATTAK received the B.S. (CS) degree from International Islamic University, Islamabad, Pakistan, in 2003, the M.S. degree in computer science from the National University of Computer and Emerging Sciences, Foundation for Advancement of Science and Technology (NUCES-FAST), Islamabad, in 2006, and the Dr.-Ing. degree (Ph.D.) in computer science engineering from Leipzig University, Germany, in 2014. During his Ph.D. studies, he worked in a research group Automatische Sprach Verarbeitung (ASV), as a Research Scientist. Since 2014, he has been working as an Assistant Professor with the Department of Computer Sciences, Quaid-i-Azam University, Islamabad. His current research interests include information retrieval systems, natural language processing, machine learning, text mining, and recommender systems.

MUHAMMAD IMRAN (Member, IEEE) received the Ph.D. degree in information technology from the Universiti Teknologi PETRONAS, Malaysia, in 2011. He is currently an Associate Professor with the College of Applied Computer Science, King Saud University, Saudi Arabia. His research is financially supported by several grants. He has completed a number of international collaborative research projects with reputable universities. He has published more than 250 research articles in peer-reviewed, well-recognized international conferences and journals. Many of his research articles are among the highly cited and most downloaded. His research interests include the Internet of Things, mobile and wireless networks, big data analytics, cloud computing, and information security. He has been involved in about 100 peer-reviewed international conferences and workshops in various capacities, such as the chair, the co-chair, and a technical program committee member. He has been consecutively awarded with Outstanding Associate Editor of IEEE Access, in 2018 and 2019, besides many others. He served as an Editor-in-Chief for *European Alliance for Innovation (EAI) Transactions on Pervasive Health and Technology*. He is serving as an Associate Editor for top ranked international journals, such as IEEE Network, *Future Generation Computer Systems*, and IEEE Access. He is also serving as a Guest Editor for about two dozen special issues in journals, such as *IEEE Communications Magazine*, *IEEE Wireless Communications Magazine*, *Future Generation Computer Systems*, IEEE Access, and *Computer Networks*.

FAZAL-E-AMIN (Senior Member, IEEE) received the B.S. degree in computer science from Hamdard University, the master's degree in information technology from Quid-i-Azam University, the master's degree in software engineering from International Islamic University, and the Ph.D. degree from the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia. He is currently serving as an Associate Professor with the Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Saudi Arabia. He secured several research grants and completed research projected successfully. He has published several research articles in reputable journals and papers in conferences. He served as a reviewer for many conferences and journals. His research interests include software project management, open-source software, software usability, the IoT, blockchain, and global software development.

LASZLO SZATHMARY received the B.Sc. and M.Sc. degrees in computer science from the University of Debrecen, Hungary, and the Ph.D. degree in computer science from Henri Poincaré University, Nancy, France.

Later, he was a Postdoctoral Research Fellow with the Université du Québec à Montréal (UQAM), Montreal, Canada. He is currently an Associate Professor with the University of Debrecen. His research interests include formal concept analysis, data mining, and artificial intelligence.