Contents lists available at ScienceDirect

Energy Reports

journal homepage: www.elsevier.com/locate/egyr

Research paper

A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection

Saddam Hussain [a], Mohd. Wazir Mustafa [a], Touqeer A. Jumani [b], Shadi Khan Baloch [c], Hammad Alotaibi [d], Ilyas Khan [e,*], Afrasyab Khan [f]

[a] School of Electrical Engineering, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[b] Department of Electrical Engineering, Mehran University of Engineering and Technology SZAB Campus Khairpur, Mirs 66020, Pakistan
[c] Department of Mechatronics Engineering, Mehran University of Engineering and Technology Jamshoro, Sindh, 76062, Pakistan
[d] Department of Mathematics, College of Science, Taif University, P.O. Box, 11099, Taif, 21944, Saudi Arabia
[e] Department of Mathematics, College of Science Al-Zulfi, Majmaah University, Al-Majmaah 11952, Saudi Arabia
[f] Institute of Engineering and Technology, Department of Hydraulics and Hydraulic and Pneumatic Systems, South Ural State University, Lenin Prospect 76, Chelyabinsk, 454080, Russian Federation

ARTICLE INFO

Article history:
Received 3 March 2021
Received in revised form 5 June 2021
Accepted 5 July 2021
Available online 24 July 2021

Keywords:
CatBoost algorithm
NTL detection
Smart meters
Feature engineering
Machine learning model interpretation

ABSTRACT

This paper presents a novel supervised machine learning-based electric theft detection approach using the feature engineered-CatBoost algorithm in conjunction with the SMOTETomek algorithm. Contrary to the previous literature, where the missing observations in data are either ignored or imputed with average values, this work utilizes k-Nearest neighbor technique for missing data imputation; thus, an accurate and realistic estimation of the missing data is achieved. To mitigate the biasness to the majority data class, the proposed model utilizes the SMOTETomek algorithm, which neutralizes the mentioned effect by managing a proper balance between over-sampling and under-sampling techniques. Feature Extraction and Scalable Hypothesis (FRESH) algorithm is utilized at the later stage of the proposed NTL detection framework to extract and select the most relevant data features from the provided dataset. Afterward, the model is trained using the CatBoost algorithm to classify the consumers into two distinct categories, i.e., genuine and theft. Finally, to interpret the model's decision for the corresponding predictions, the tree-SHAP algorithm is utilized. To validate the efficacy of the proposed ML based theft detection approach, its performance is compared with that of the traditional gradient boosting ML algorithms such as XGBoost, lightGBM, Ensemble bagging, boosting ML models, and other conventional ML models using five of the most widely used performance metrics, i.e., precision, accuracy, F1score Kappa and MCC. The proposed technique achieved an accuracy of 93% and a detection rate of 92%, which is significantly higher than all the considered competing algorithms under identical dataset and hyperparameters.

1. Introduction

1.1. Background

The transmission and distribution (T&D) of electricity suffers from two major categories of losses, i.e., technical and non-technical. The technical losses account for the energy losses that occur in equipment that is essential for implementing the T&D of electricity. On the other hand, the non-technical losses (NTL) in any power system account for power theft, billing irregularities, and corruption within utility workers. According to a report, utilities around the globe are losing approximately US$96 billion every year due to NTLs (Northeast Group, 2017). The mentioned scenario is precisely depicted in Fig. 1, which shows the intensity of the NTL issue in different parts of the world.

Owing to such massive economic loss, the power utilities and researchers in the field of data mining, computer science, and electrical engineering are trying several intelligent and effective methods to minimize NTLs. One of the efficient methods to counter the electric theft issue is the implementation of smart meters. Such energy meters can monitor and record the consumers' consumption data remotely and precisely and provide the information to the utility directly in case of any suspicious activity. However, despite the vast number of benefits, smart meters are not feasible for countries suffering from severe economic issues due to huge expenditures associated with their implementation and operation. Furthermore, increasing cyber threats still

[*] Corresponding author.
E-mail addresses: i.said@mu.edu.sa (I. Khan), khana@susu.ru (A. Khan).

https://doi.org/10.1016/j.egyr.2021.07.008

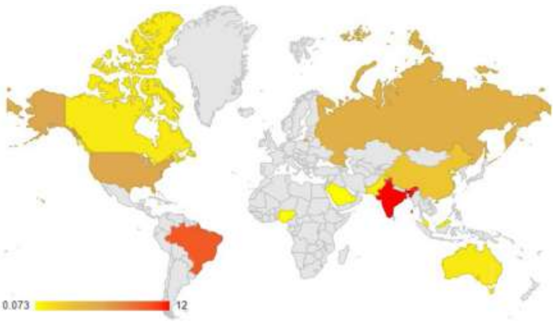


Fig. 1. Non-technical loss in billion dollars country-wise.

need to be addressed appropriately for the wide-scale implementation of such devices. In addition, the high-frequency data gathered from smart meters pose some serious data storage and analysis issues. It was estimated that the volume of data obtained from two million consumers' smart meters might exceed 22 GB per day (Rusitschka et al., 2010). Therefore, it is extremely challenging to identify suspicious consumers' profiles from such a huge dataset.

The NTL detection approaches can be broadly classified into three major categories, i.e., theoretical, hardware, and non-hardware based methods (Viegas et al., 2017). The theoretical methods utilize the relationship between the socio-economic and demographic factors for framing the policies to counter NTLs (Winther, 2012; Never, 2015; Yurtseven, 2015; Mwaura, 2012). On the other hand, the hardware-based or state-based methods utilize physical instruments such as sensors, detection devices, transformers, and other electrical devices to detect NTLs (Chen et al., 2013; Xiao et al., 2013; Jaiswal and Ballal, 2020; Saad et al., 2017). In these methods, the voltage, power, and current sensors are installed at various network nodes, that triggers an alarm whenever the malicious customers attempt to manipulate the actual grid characteristics at any network point. Despite an effortless working mechanism, these methods are not feasible for various power utilities due to additional maintenance and sensor deployment costs. Contrary to the hardware-based methods, the non-hardware-based energy-theft detection approaches do not require any additional NTL detection device. These methods are generally classified into two major categories, i.e., game-based and data-driven systems. In the former approach, the theft detection method is developed as a game between the power thief and the service provider using game theory. Even though these approaches require a comparatively lower cost, they pose a severe challenge in identifying the key position of players, offenders, regulating authorities, and distributors; thus, making it too complex to implement. The second category of the non-hardware-based machine learning techniques is data-driven methods. These methods are further classified into unsupervised and supervised machine learning approaches. The former methods utilize a clustering approach to segment consumers' load profiles based on similarity or dissimilarity metric measures (Badrinath Krishna et al., 2016; Ferreira et al., 2013; Passos Júnior et al., 2016; Hussain et al., 2020).

On the other hand, the supervised or classification-based theft detection methods utilize pre-labeled data (i.e., "Genuine" and "Theft") to train the model at the initial stage. Based on the information acquired from the training process, the model is made to classify the unlabeled data into two mentioned distinct categories; thus, minimizing the expenses and labor of site-inspections. Since this research work proposes a supervised ML-based approach, a detailed description of the most relevant literature in the mentioned research field is provided in the subsequent subsection.

1.2. Positioning of our work in literature

The supervised-based NTL detection methods generally face five major challenges, i.e., handling missing data values during data pre-processing, data class unbalancing, selecting the most relevant features, choosing an appropriate classifier, and interpreting the model's prediction. This subsection reviews the most relevant literature pertaining to the challenges mentioned above in conjunction with the significance of the current research work.

Paria et al. (Jokar et al., 2016) presented a consumption pattern-based energy theft detection (CPBETD) algorithm to identify the malicious consumption patterns in a smart grid network. The proposed CPBETD algorithm was made to detect the high energy theft areas at the transformer level by utilizing the data collected from the various distribution transformer meters. In another study (Jindal et al., 2016), the authors developed a highly accurate energy theft detection framework by utilizing the support vector machine (SVM) intelligence in conjunction with the decision tree algorithm. Even though both the studies have proposed very effective theft detection frameworks, however, none of them has tackled the missing data issue. Furthermore, the authors in Tureczek and Nielsen (2017), after a detailed review of 34 research papers on theft detection based on supervised ML methods, concluded that only half of the considered articles had addressed the issue of missing data values. Since current research work has detailly handled the mentioned problem, it is essential to emphasize its repercussions.

The consumption data obtained through the smart meters is generally inconsistent and often contains null values. Several factors, such as smart meter malfunction, inaccurate estimation of data transferred, unplanned device repair, and storage problems, can be the root cause of this problem. It is extremely difficult for a learning classifier to handle and learn patterns from such data types. To overcome the stated issue in ML based classification methods, various data imputation strategies have been proposed in the literature, such as Hot deck imputation method (Joenssen and Bankhofer, 2012), data clustering based imputation (Zhang et al., 2008), Monte Carlo missing values imputation method (Roth and Switzer, 1995) etc. Two of the most widely practiced solutions to counter this issue are to delete the missing entries from the original data (listwise or pairwise) or to impute the missing datapoints with mean values between the adjacent data entries as witnessed in references (Buzau et al., 2018a; Adil et al., 2020). The mentioned data adjusting methods are elementary and reasonable; nevertheless, the former method produces a significant information loss while the second provides noisy, inconsistent, and outlying data values. To overcome the stated issues, this study utilizes the k-Nearest neighbor-based imputer which imputes the average value from pre-selected kth number of nearest neighbors in a given sample of data, thus providing very reliable estimates.

Another critical issue in smart meters' labeled data sets for NTL detection application is the data class unbalancing. It causes difficulties for the learning systems to learn the concept related to the minority class (theft cases); thus, causing biasness of ML models towards the majority samples. In order to achieve an effective and unbiased ML model performance, a balanced set of the dataset is essentially required. Two of the prominent studies that have tackled the mentioned issue includes Hasan et al. (2019) and Gunturi and Sarkar (2020). Both studies have utilized the Synthetic minority oversampling technique (SMOTE) to balance the data class with reasonable accuracy. Since the SMOTE algorithm oversamples the minority class randomly, it results in overfitting and low generalization ability of the model. In another study (Buzau et al., 2018b), the authors have utilized an under-sampling technique where few samples of the majority class are removed to balance the data class. Such data balancing techniques are easy to implement; however, they may cause substantial data loss, resulting in lower accuracy of the developed NTL detection model. To avoid this issue, the current study utilizes an efficient statistical technique called SMOTETomek (Batista et al., 2004a). It combines the intelligence of SMOTE (oversampling) and Tomek link (under-sampling) to balance the data class distribution.

As discussed earlier in this section, the third critical issue in supervised-based NTL detection methods is the selection of the most relevant features for the model training. The efficiency of classification-based theft detection methods is highly dependent on the type of input features selected. Since the smart meters' data is generally high dimensional data containing many redundant and irrelevant features, it is essential to extract and select the most relevant features and discards the unnecessary ones. In this study, the mentioned issue is solved by using efficient feature extraction and selection process. Feature extraction and selection procedure is an effective practice for reducing the increased data dimensions, and redundant information in ML-based NTL approaches. It is worthwhile to mention that unlike most of the ML-based NTL detection approaches in literature where either feature extraction or selection process is adopted for model training, this research work utilizes both for acquiring highly relevant features from the considered smart meter dataset. The proposed approach utilizes the intelligence of one of the most intelligent algorithms called the Feature Extraction and Scalable Hypothesis (FRESH) algorithm to accomplish the mentioned task. It does so by utilizing more than 60 time-series analytical methods to capture 794 features from each dataset sample. The extracted features are reduced to 300 most relevant features through the Benjamini–Yekutieli statistical test. The resulting final set of features are a combination of essential user consumption and newly extracted features.

Once the feature engineering process is completed, the next challenge is to select an appropriate classifier for efficiently segregating the genuine and theft consumers. In this study, the CatBoost algorithm is utilized for the model training due to its efficient handling of the categorical features. These categorical features are handled during the pre-processing phase in most of the traditional ML models, which consequently increase the computational time and complexity. On the other side, the CatBoost efficiently handle these features during the training process, thus avoids the mentioned problems faced by conventional classifiers. Furthermore, it utilizes the intelligence of ordered boosting, which avoids the prediction shift problem faced by XGBoost and its variants. Also, by enabling the overfitting detector feature in its framework, the trained model can achieve an improved generalization ability.

Another important aspect of the proposed theft detection model is its novel interpretability of the model outcomes. Mostly, site inspections are initiated on the list of suspected consumers generated by the trained model on genuine and theft consumers' data. However, a model's prediction to place the consumer in a particular category based on a given input feature set is not justified logically. Nevertheless, few studies in literature such as Batista et al. (2004b) and Christ (2018), have employed simplistic decision tree diagrams to interrupt the model outcomes. However, the latest state-of-the-art theft detection models employing deep learning, gradient boosting machines and ensemble ML techniques incorporate a diverse range of complex prediction strategies, making themselves extremely difficult to comprehend through simplistic tree diagrams. To deal with the mentioned issue, tree-SHhapley Additive exPlanations (SHAP) is utilized in the current study. It assists in opening the black-box ML model's outcomes in terms of explaining how the model concluded a decision for a particular prediction.

It is fair to mention and highlight the most relevant studies on the current research work available in the literature. One of such studies was carried out by Gunturi and Sarkar (2020), where the authors have developed an ensemble machine learning-based theft detection model. In another study, Punmiya and Choe (2019) proposed a gradient boosted theft detector framework, which employs the latest XGBoost, lightGBM, and CatBoost for model training. The current study differentiates itself from the mentioned research works in its novel data class balancing and feature engineering approach. Furthermore, unlike the quoted studies where the model's outcome interpretability was not evaluated, this research work utilizes the tree-SHAP algorithm to accomplish the mentioned task.

Concluding the detailed discussion, the list of steps executed sequentially in order to accomplish the proposed supervised ML-based NTL detection framework is presented as follows.

i. k-Nearest Neighbors imputation technique is employed to handle the missing and erroneous data values in the acquired dataset.
ii. SMOTE-Tomek based resampling technique is utilized to tackle the data class imbalance issue.
iii. The FRESH algorithm is used to extract and select the most relevant statistical features from raw smart meter data.
iv. The implementation of the state-of-the-art CatBoost algorithm and its comparative analysis with other well-known ML classifiers is carried out for identifying the NTLs.
v. Interpretation of the model outcomes is performed through the tree-SHAP algorithm.
vi. To validate the effectiveness of the proposed theft detection framework, an extensive performance evaluation is made based on five of the most widely utilized performance metrics.
vii. The proposed NTL framework achieves the highest detection rate and the lowest false positive rates among all the compared algorithms.

The rest of the paper is divided into three sections. Section 2 presents the proposed research methodology and is further sub-categorized to discuss the CatBoost algorithm's theoretical background, considered performance metrics, and proposed framework results and interpretations. In Section 3, the proposed model's comparative analysis against the latest gradient boosting decision trees (GBTDs) and traditional ML models is discussed in detail. Finally, the conclusion is made in Section 4 of this research work.

2. Research methodology

In this section, the proposed NTL detection framework is presented. The overall framework is broadly classified into three major stages, i.e., data pre-processing stage, feature engineering stage, model training-testing, and interpretation stage. Each of the stages is depicted in Fig. 2. and detailly described in subsequent subsections.

2.1. Stage-1: Data pre-processing stage

Data pre-processing is required to transform the raw data into a meaningful data structure. The electricity consumption data acquired from the State Grid Corporation of China (SGCC) (Zheng et al., 2018) is used for testing the efficacy of the proposed theft detection model. Table 1 presents the metadata information of the acquired dataset.

As presented in Table 1, the daily electricity consumption of 42372 consumers for approximately 1035 days (2014-Jan to

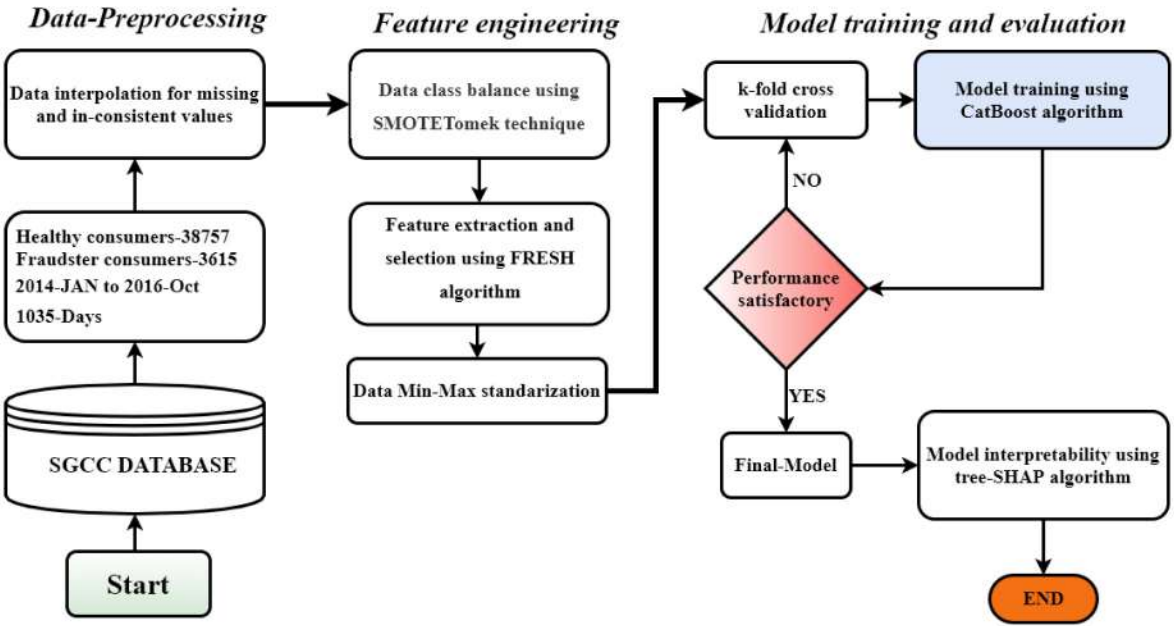

Data-Preprocessing
Feature engineering
Model training and evaluation
Data interpolation for missing and in-consistent values
Data class balance using SMOTETomek technique
k-fold cross validation
Model training using CatBoost algorithm
NO
Healthy consumers-38757
Fraudster consumers-3615
2014-JAN to 2016-Oct
1035-Days
Feature extraction and selection using FRESH algorithm
Performance satisfactory
Data Min-Max standarization
YES
SGCC DATABASE
Final-Model
Model interpretability using tree-SHAP algorithm
Start
END

Fig. 2. Proposed theft detection framework.

Table 1
Statistics of obtained SGCC data.

Description	Value
The time window for electricity consumption	2014-01-01 to 2016-10-31 1035- days
Number of total consumers	42372
Number of electricity thieves	3615
Number of genuine consumers	38757
Total data records	42372 * 1035 = 43855020

2016-Oct). It comprises of 91.46% genuine and 8.54% theft consumers. Fig. 3 and Fig. 4 depict the electric power consumption patterns for few of the theft and the genuine consumers respectively. It can be observed from the mentioned Figures that the theft consumption patterns of the theft consumers are highly irregular and contains low periodicity. On the other hand, the patterns for the genuine consumers are periodical and exhibits a correlation between the identical periods of the consecutive years.

To check the missing information in the data, the NaN values were computed for each consumer. It was found that 25.6% of 43855020 data entries contains NaN or missing values, which is significantly higher for any data set in the field of data mining. The distribution of computed null values in terms of the histogram is shown in Fig. 5. The histogram bar values depict the number of consumers falling in the missing values range.

The computed histogram illustrates that 22.6% of total consumers fall into the range of more than 700 missing values per consumer. To correctly estimate these consumers' missing data values becomes extremely challenging since a significant portion of the information is unavailable in the acquired dataset. Therefore, a viable option left is to drop such highly inadequate entries from the rest of the dataset. The missing values in the remaining consumers are imputed using the kNN interpolation technique (Troyanskaya et al., 2001). The kNN is a non-parametric and lazy learner algorithm that matches an observation in multidimensional space to its nearest kth neighbors. The kNN's capability of dealing with almost all types of missing data makes it a suitable candidate for the missing value imputation. It accomplishes the imputation task by utilizing the Euclidean distance metric to initially find the consumer's *k*th nearest neighbors and then imputes the missing feature value using the mean of selected k-neighbors. The current study utilizes the KNN-imputer module available in the Scikit-learn ML package to impute the missing data slots (Pedregosa et al., 2011). A few random consumers' consumption samples are plotted to visualize the newly imputed values in consumers' consumption data, as shown in Fig. 6.

2.2. Stage-2: Data class balancing and feature engineering

This stage is further divided into two sub-stages, i.e., data class balancing and feature engineering, as depicted in Fig. 2. Each of the mentioned sub-stages is explained in subsequent subsections.

2.2.1. Data class balancing

For an efficient and unbiased classifying performance of a supervised ML classifier, it is essential to extract and select the most suitable features from a balanced dataset. Since the considered smart meter dataset for the current study is unbalanced, as it occurs in most NTL detection data set, it is necessary to balance the class distribution before the feature extraction and selection process. In order to solve this issue, the SMOTETomek (Batista et al., 2004b) algorithm is utilized in the current study. SMOTETomek combines the intelligence of SMOTE and Tomek links techniques to over and under-sample data classes simultaneously. It accomplishes the mentioned task by discarding the majority class links until both classes reach an equal number of entities. Even though the SMOTE technique alone can mitigate the imbalanced data class distribution issue, it skews the class distributions. Since in most of the real-world smart meter datasets, clusters formed by different data classes are not well expressed. Therefore, a set of samples belonging to the minority or majority class is expected to be dominated during the SMOTE technique's oversampling period. Consequently, feeding such biased data to the learning classifier will lead to model overfitting.

On the other hand, SMOTETomek does not only helps in producing well-defined data class distribution, but it also generates data class clusters equally. The data class distribution for the current study before and after using SMOTETomek is shown in Fig. 7.

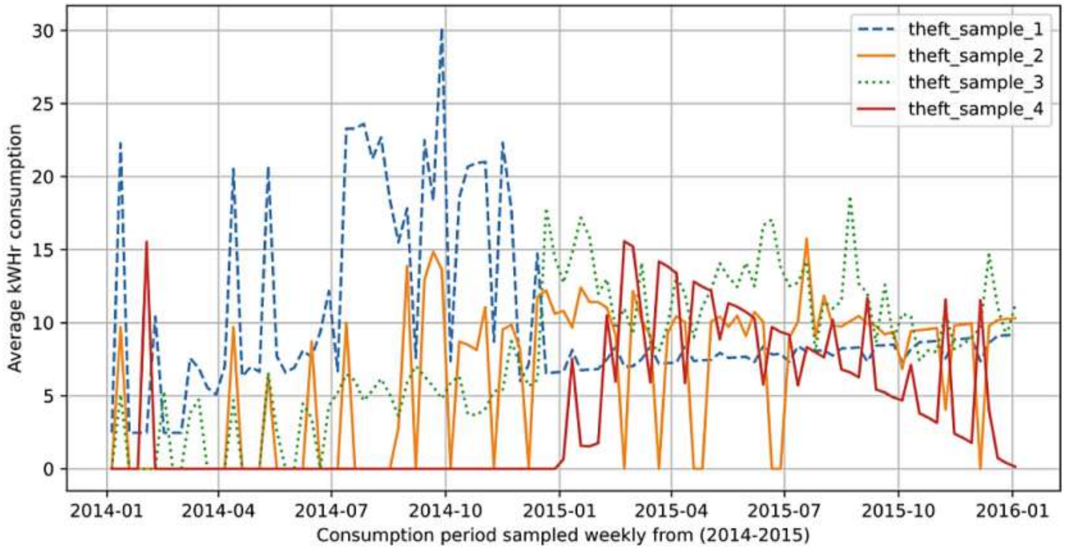


Fig. 3. Electric consumption samples of consumers involved in power theft.

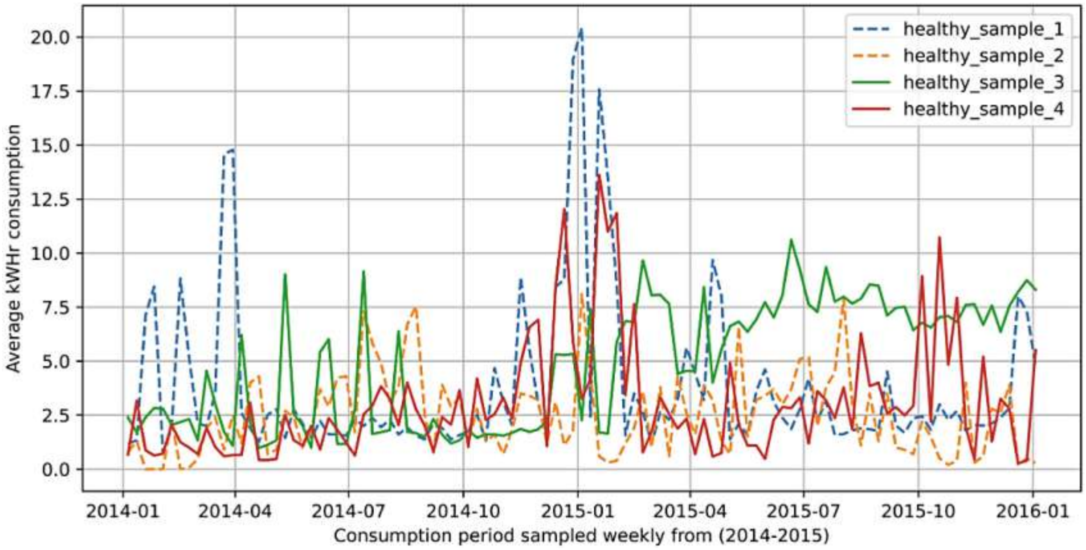


Fig. 4. Electric consumption samples of genuine consumers.

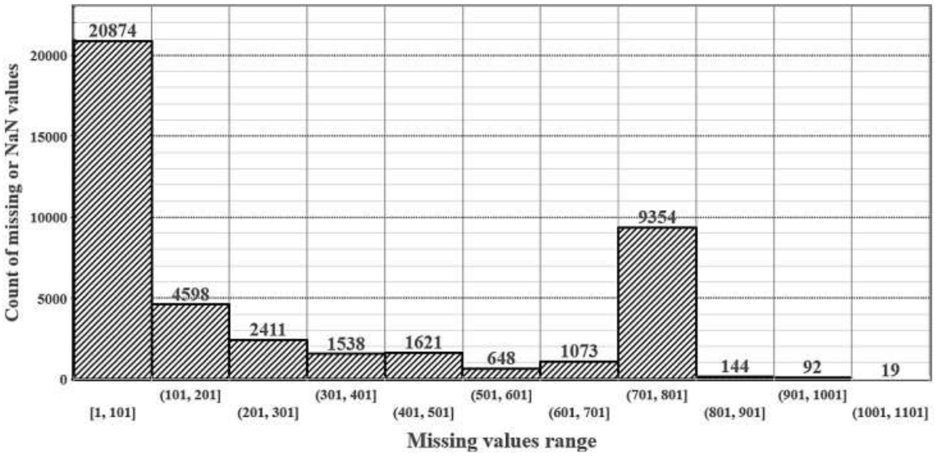


Fig. 5. Histogram of missing values present in SGCC dataset.

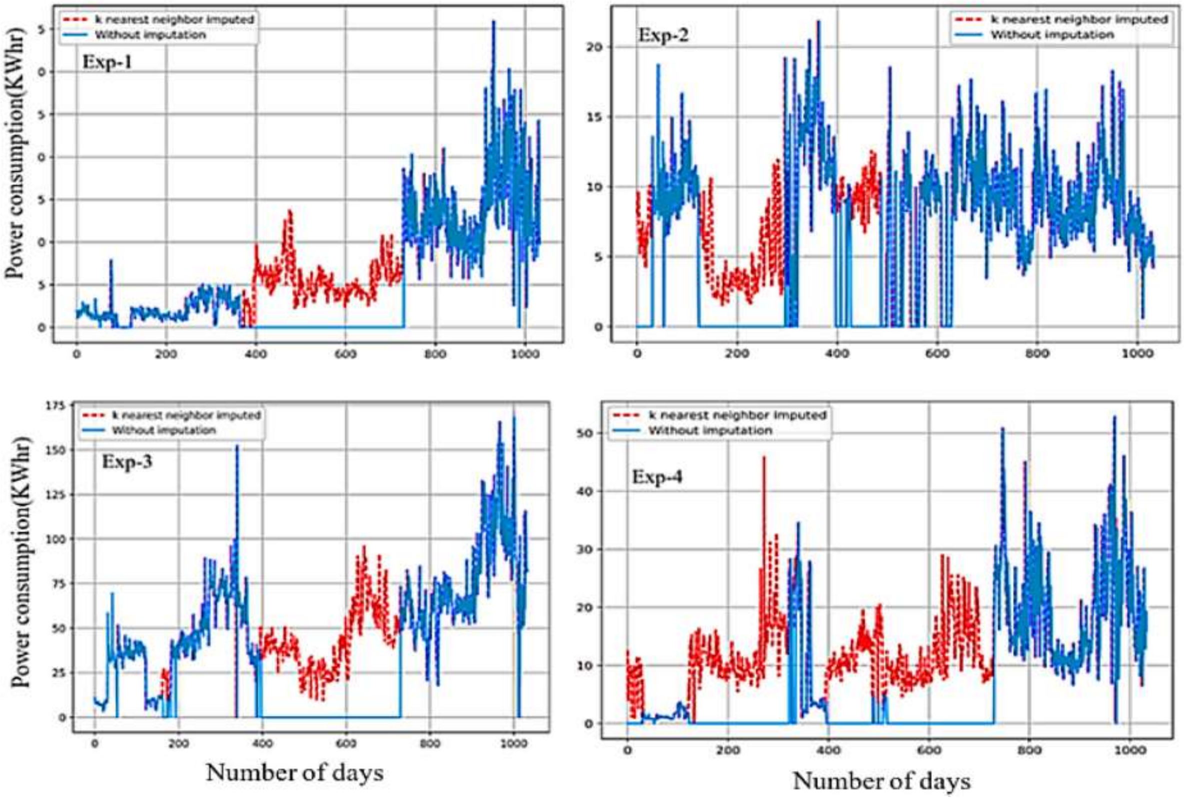
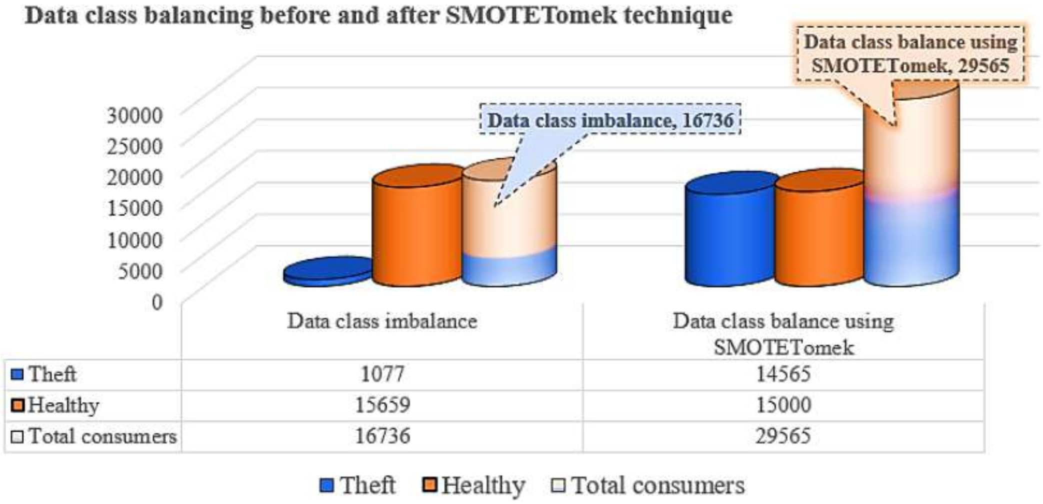


Fig. 6. Missing value imputation using the K-nearest neighbor technique.



	Data class imbalance	Data class balance using SMOTETomek
■ Theft	1077	14565
■ Healthy	15659	15000
□ Total consumers	16736	29565

Fig. 7. Data class distribution before and after using the SMOTETomek technique.

Fig. 7, shows that genuine consumers are significantly higher in number than those engaged in fraud before applying the SMOTETomek. In contrast, both the classes are well balanced after employing the proposed technique.

2.2.2. Feature engineering

In this section, the proposed feature engineering process is discussed in detail. Feature engineering is the process of extraction and selection of the most important features from given data typically done to enhance the ML model's learning ability. It is important to note that the dataset acquired from the smart meters lack statistical characteristics. For a theft detection model to be efficient, features fed to the model must reflect appropriately underlying abnormalities in consumers' consumption data. Therefore, the additional characteristics of the provided dataset are extracted using the feature extraction and selection process. In this study, both the tasks are accomplished using the FRESH algorithm, which simultaneously extracts and selects useful features from the given balanced dataset. For ease in computation, the FRESH algorithm authors have developed a standardized python-based package called "ts-fresh", which makes use of the FRESH algorithm within its framework. The source code and GitHub page of the ts-fresh package can be found in the link provided in Christ (2018). A complete list of extracted features and their mathematical description can be found in Christ et al. (2016),

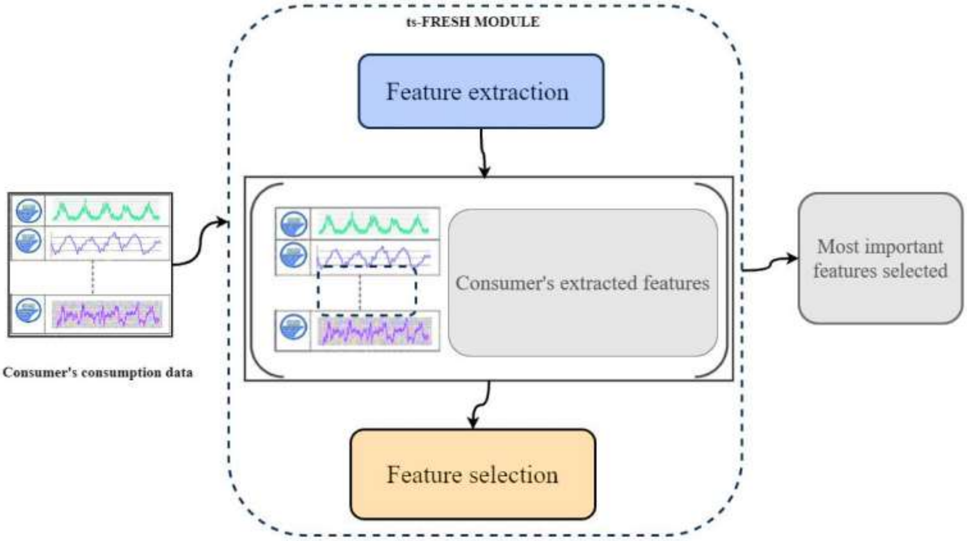

ts-FRESH MODULE
Feature extraction
Consumer's extracted features
Feature selection
Most important features selected
Consumer's consumption data

Fig. 8. Feature extraction and selection process using the FRESH algorithm.

while the simplified pictorial version of the feature extraction and selection process employing the FRESH algorithm is presented in Fig. 8.

The FRESH algorithm implementation using the ts-FRESH module is carried out in two steps, as depicted in Fig. 8. Initially, 794 features are extracted automatically from each consumer's consumption data using more than 60 time-series characterization methods. These extracted features can be broadly classified into temporal, statistical and spectral domains as depicted in Fig. 9.

Features such as entropy, zero-crossing points, spectral variation, Mel-Frequency Cepstral Coefficients (MFCC), skewness, kurtosis, trend, linear and non-linear characteristics, correlation, and various statistical test-based features provide in-depth knowledge of each consumer consumption sample. Due to the space limitation all the extracted features are not shown in Fig. 9, for the interested reader as mentioned above the detailed documentation of each feature along with source code for its implementation can be found in authors provided webpage (Christ, 2018).

In the second step, the derived features and consumers' actual consumption data are combined to select only highly important feature. This selection process is made by initially arranging the features in descending order based on their significance gauged through various statistical tests. Afterwards, the Benjamini and Yekutieli (2001) procedure is employed that sets a threshold for feature selection criteria; thus, the features with the negligible contribution to the target variable are discarded automatically. Since the feature-set selected by the FRESH algorithm contains diverse data points scattered over a wide range, the features with higher magnitudes will cause biasness during the model training. Therefore, it is crucial to standardize the accumulated features to a common scale. The current study utilizes the feature-wise Min–Max data standardization method to overcome the mentioned issue. Min–Max converts each numerical attribute to the range of 0 to 1 by using the following mathematical expression.

$$\mathrm{f}\,(\mathrm{x_i}) = \frac{\mathrm{x_i} - \min(\mathrm{X})}{\max\,(\mathrm{X}) - \min(\mathrm{X})} \tag{1}$$

where X is a vector composed of x_i daily electricity consumption while the min(X) and max(X) are the minimum and maximum values of X respectively.

2.3. Stage-3: Model training and evaluation stage

In this section, the training and evaluation of the proposed NTL detection model are discussed in detail. For ease of understanding and interpretation, this section is divided into three sub-sections.

2.3.1. Performance evaluation metrics

In any supervised ML technique, the labeled data is provided to the learning classifier for its training purpose initially. The trained model is then evaluated for its ability to predict and generalize the un-labeled data efficiently. The performance of such models is assessed based on a number of performance evaluation metrics, such as mentioned in Messinis and Hatziargyriou (2018). However, it is not feasible to evaluate and analyze all the metrics mentioned in the stated study; therefore, few of the most important metrics such as accuracy (Acc), recall, confusion matrix (CM), precision (P), Cohen's kappa coefficient (kappa), Matthews correlation coefficient (MCC), and $\mathrm{F1_{score}}$ are utilized to evaluate the performance of the proposed classifier. The mathematical expressions for calculating the mentioned metrics are depicted in Eqs. (2)–(9).

$$\text{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{2}$$

$$\text{Recall or Detection rate} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\text{False} - \text{positive rate} = \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{4}$$

$$\text{False} - \text{negative rate} = \mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} \tag{5}$$

$$\text{Precision} = \mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

$$\mathrm{F1_{score}} = 2*\frac{\text{Precision} * \mathrm{DR}}{\text{Precision} + \mathrm{DR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{7}$$

$$\text{Kappa} = \frac{\rho_o - \rho_e}{1 - \rho_e} \tag{8}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP}^{*}\mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}, \tag{9}$$

where FP and TP denote the false positive and true positive respectively, while FN and TN represent false negative and true negative respectively. ρ_o is the predicted value and ρ_e is the actual value.

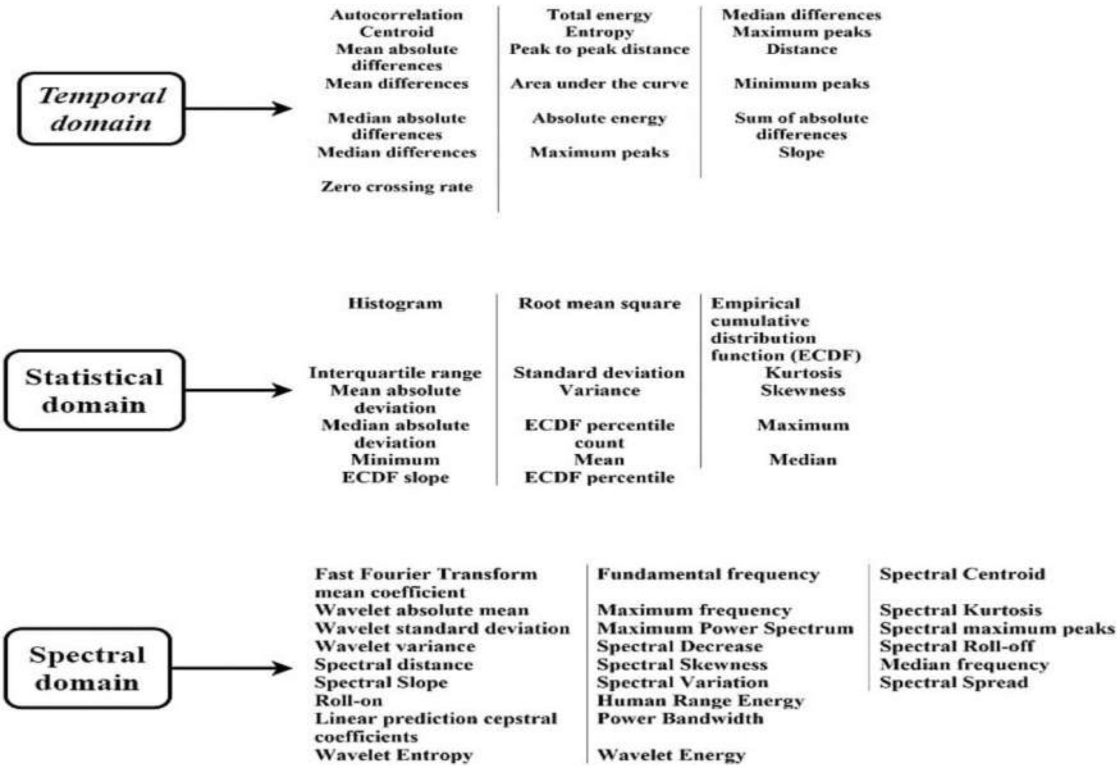


Fig. 9. Extracted features using the FRESH algorithm.

In addition to the appropriate selection of performance assessment metrics, the performance evaluation of the considered ML model on different test datasets is also important. Therefore, k-fold cross-validation technique is recommended in most of the literature (Salman Saeed et al., 2020; Saeed et al., 2019). In the k-fold cross-validation technique, the entire dataset is divided into the k-number of folds initially. Afterwards, the first k_1 fold is used to train the model, and the remaining (k-k_1) folds are used for validation purpose. Finally, the outcomes of all the considered evaluation metrics are averaged to depict the performance of the learning classifier.

2.3.2. CatBoost classification algorithm: Theoretical background and its implementation in current classification problem

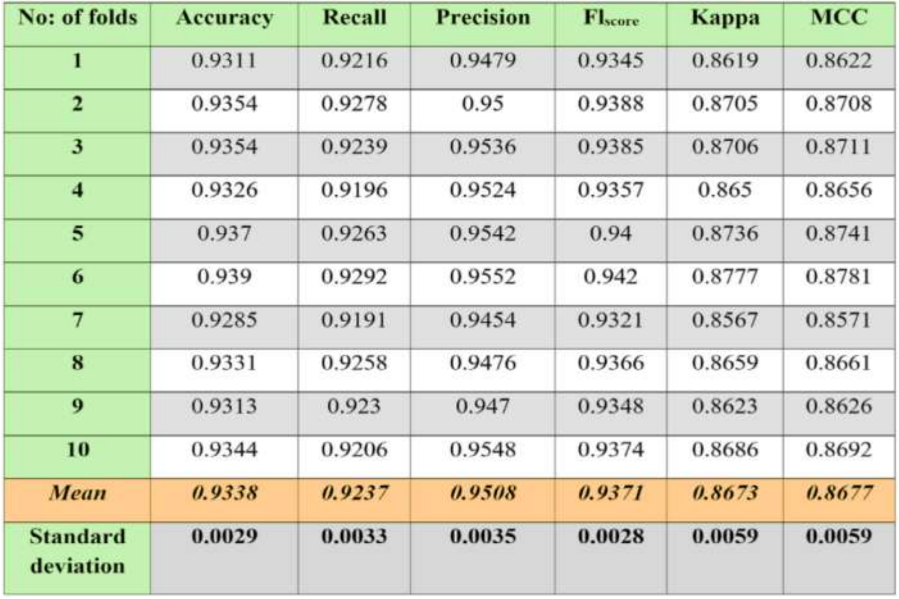
In this study, the CatBoost classification algorithm is utilized for model training and evaluation purpose. CatBoost is a refined version of the GBDTs, which utilizes a complex ensemble learning technique based on the gradient descent framework. During model training, a set of decision trees (DTs) are constructed sequentially to create each subsequent tree with decreased loss. In other words, each DT learns from the preceding tree and influences the next tree to boost the model performance, thus builds a strong learner. CatBoost algorithm differs from the rest of GBTs in terms of having two prominent features, i.e., efficient handling of categorical features and ordered boosting (Prokhorenkova et al., 2018). The learning classifiers handle numerical features quite efficiently during the model training phase; however, interpreting categorical features is complicated for them. Therefore, in conventional approaches, categorical features are transformed into useful information using the one-hot encoding technique (Daniele, 2001) or gradient statistics (Ke et al., 2017). In the former technique, each category of the original categorical features is replaced by the binary values, while in the later technique, an estimated value is generated by using gradient statistics to replace the original categorical feature at each boosting step. Nevertheless, in the case of the categorical features with high repeatability, both the mentioned techniques require large memory and other computational resources. To avoid the mentioned problem, the CatBoost algorithm utilizes efficient modified target-based statistics to appropriately handle the categorical features during training time, thus saving considerable computational time and resources.

Another important aspect of the CatBoost algorithm is its ordered boosting mechanism. In traditional GBTs, all the training samples are provided to construct a prediction model after executing several boosting steps. This approach causes a prediction shift in the constructed model, which consequently leads to a special kind of target leakage problem. The CatBoost algorithm avoids the stated issue by utilizing the ordered boosting framework. Furthermore, contrary to the conventional learning classifiers, the CatBoost algorithm eloquently handles the overfitting issue by using several permutations of the training dataset; hence it turns out to be the key motivation behind utilizing its intelligence in the current study.

For the effective implementation of the proposed CatBoost algorithm in the current NTL detection problem, the designed model is initially trained on the data developed in Stage-2. Afterward, a10-folds cross-validation (CV) technique employing the considered performance metrics is utilized for performance evaluation of the designed model. The corresponding outcomes are depicted in Table 2.

As can be seen from Table 2 that the CatBoost model attained an average accuracy and precision of 0.9338 and 0.9508 with a standard deviation (SD) of 0.0029 and 0.0035, respectively. It is essential to mention that in almost all data-oriented NTL detection systems, accuracy, and precision are two of the most widely used metrics. Nevertheless, these metrics cannot be considered as a conclusive measure to assess NTL detection-based classifiers' performance. For example, precision is a critical performance metric; however, it lacks significant information regarding False-negative (FN) instances. The FN value implies consumers involved

Table 2
10-folds cross-validation results achieved using the proposed model.



No: of folds	Accuracy	Recall	Precision	F1$_{score}$	Kappa	MCC
1	0.9311	0.9216	0.9479	0.9345	0.8619	0.8622
2	0.9354	0.9278	0.95	0.9388	0.8705	0.8708
3	0.9354	0.9239	0.9536	0.9385	0.8706	0.8711
4	0.9326	0.9196	0.9524	0.9357	0.865	0.8656
5	0.937	0.9263	0.9542	0.94	0.8736	0.8741
6	0.939	0.9292	0.9552	0.942	0.8777	0.8781
7	0.9285	0.9191	0.9454	0.9321	0.8567	0.8571
8	0.9331	0.9258	0.9476	0.9366	0.8659	0.8661
9	0.9313	0.923	0.947	0.9348	0.8623	0.8626
10	0.9344	0.9206	0.9548	0.9374	0.8686	0.8692
Mean	*0.9338*	*0.9237*	*0.9508*	*0.9371*	*0.8673*	*0.8677*
Standard deviation	0.0029	0.0033	0.0035	0.0028	0.0059	0.0059

in theft yet classified as genuine; hence the failure of this kind can cause permanent financial loss.

For that reasons, the proposed approach's performance is further authenticated by computing the recall, F1score, kappa, and MCC, Recall or detection rate (DR) value specifies a classifier's hit rate in accurately classifying the theft instances. The proposed technique attained a high average DR value of 0.9237 with standard deviation (SD) of 0.0033. On the other hand, MCC is a more balanced and informative statistical metric, which provides a high score only if the prediction has achieved good scores in all four confusion matrix categories. MCC score ranges from −1 (total conflict between outcome and observation) to +1 (perfect prediction). The average value of MCC attained in this study is 0.8677 with SD of 0.0059, which implies that the proposed technique correctly classifies most of the theft and genuine cases from the provided dataset.

2.3.3. Proposed model's outcomes interpretability using the tree-SHAP algorithm

In this section, the proposed theft detection model outcomes or predictions are interpreted using Shapley values computed by the tree-SHAP algorithm. The Shapley values assist in opening the black-box ML model outcomes extensively. These values provide a solution for fairly assigning the gains and costs to several features working in alliance for predicting the model outcomes. In simple words, these values assist in explaining how model has concluded a decision for a particular prediction. In this study, the Shapley values are computed using a recently introduced technique called tree-SHAP developed by Lundberg et al. (2020). The tree-SHAP algorithm is specially designed for tree-based models, and ensemble gradient boosted machines. One of the important features of this algorithm is that it computes the local feature interaction, which in-turn facilitates the interpretation of the global model structure for each prediction. A detailed explanation and source code of the tree-SHAP technique is presented tree-SHAP GitHub webpage (https://shap.readthedocs.io/). Fig. 10 shows the summary plot generated by the tree-SHAP algorithm that helps in interpreting the predicted outcomes of the proposed theft detection model.

The summary plot shown in Fig. 10, plots the consumers' extracted features against the computed Shapley values. The Shapley values are computed for every consumer's each feature value and plotted against the selected features in order to evaluate its impact on the model outcome. Since its quite challenging to show all the features and their corresponding Shapley values in the summary plot, therefore, only 20 most essential features are depicted in ascending order based on their significance in predicting the model outcomes. For example, the entropy feature attained the highest importance in terms of predicting the target variable, as shown in Fig. 10. It implies that most of the consumers with high entropy values (i.e., red color) obtain a positive SHAP value; thus, impacting the model outcomes positively. Further aspects of interpreting the ML model using the SHAP technique can be found in this source (Molnar, 2018).

3. Comparative analysis of proposed method with conventional ML classification methods

In this section, the performance of the proposed theft detection framework is compared against the latest GBTDs and other well-known conventional ML models under an identical feature set. The 10-fold cross-validation technique is employed in conjunction with the five most widely utilized performance metrics, i.e., precision, accuracy, F1$_{score}$Kappa, and MCC, to evaluate the performance of all studied classifiers. The proposed framework is sequentially implemented using the 8th generation, Intel Core-i5, RAM-8-GB unit. It took approximately 280 s for the model training and testing, while the feature extraction and selection process took around 600 s. Since the classifier utilized in the proposed framework is a modified variant of tree-based models, therefore its performance is compared with other tree-based models such as RF, ET, Ada Boost, XGBoost light and GBM. The outcomes of this comparison are depicted in Fig. 11.

As evident from Fig. 11, the proposed technique outperforms all the conventional ML methods in terms of accuracy, recall, precision, F1score, Kappa, and MCC; thus, proving its effectiveness and significance. Another performance evaluation-based comparison of the proposed method with a few of the well-known conventional ML methods is made on identical performance evaluation metrics. The corresponding outcomes are depicted in Fig. 12. Once again, the proposed method's performance superiority can be observed from outcomes depicted in Fig. 12. It achieves an accuracy, recall, precision, F1$_{score}$, Kappa, and MCC of 93.38%, 92%, 95%, 93.7%, and 87%, respectively, which are significantly higher than all the competing models.

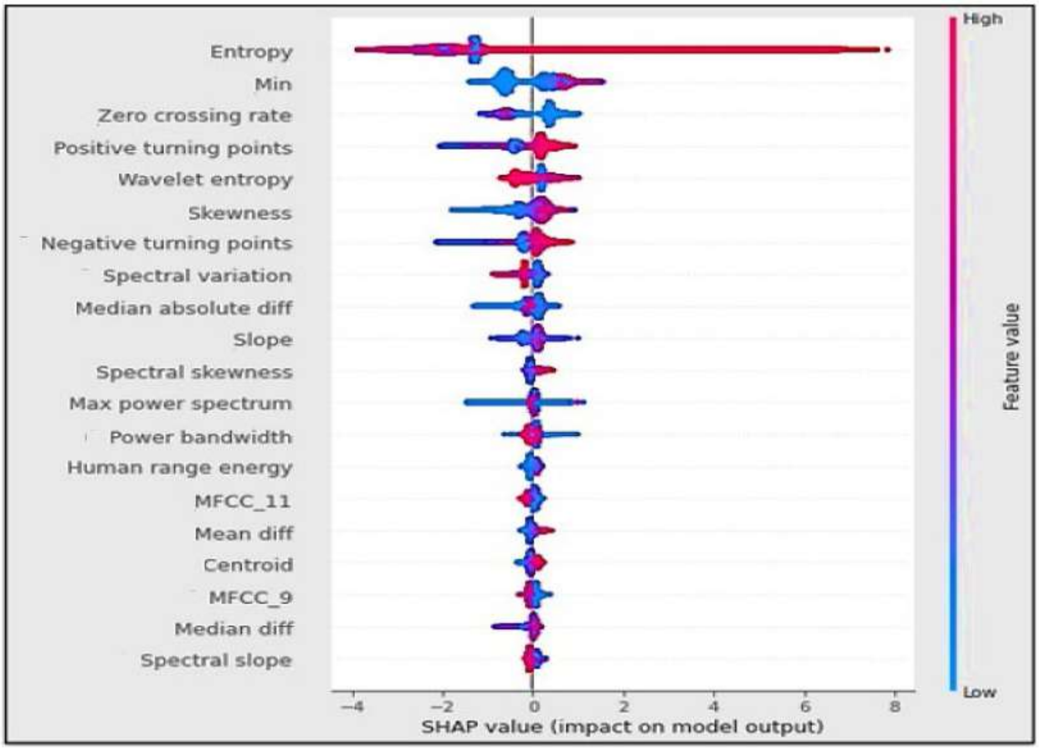
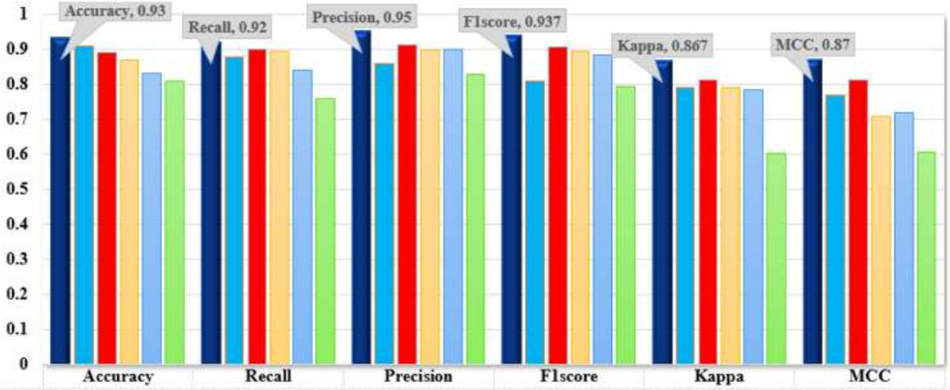


Fig. 10. SHAP value of the proposed model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



	Accuracy	Recall	Precision	F1score	Kappa	MCC
Proposed Model	0.93	0.92	0.95	0.937	0.867	0.87
Extreme Gradient Boosting	0.91	0.877	0.86	0.81	0.79	0.77
Extra Trees Classifier	0.89	0.8994	0.9114	0.9054	0.812	0.8121
Random Forest Classifier	0.87	0.8941	0.8964	0.8952	0.7908	0.71
Light Gradient Boosting Machine	0.83	0.8417	0.9	0.886	0.7834	0.72
Ada Boost Classifier	0.81	0.7601	0.8292	0.7928	0.6042	0.6066

Fig. 11. Performance evaluation of studied tree-based ML models.

4. Conclusion

In this paper, a novel feature engineered CatBoost-based NTL detection framework is developed. At the initial stage of the proposed NTL detection framework, the missing slots in the acquired data set were imputed using kNN missing value imputer. To avoid the data class imbalances, the SMOTETomek algorithm was utilized which simultaneously over and under-sample the data classes. The FRESH algorithm's intelligence was utilized at the later stage to extract and select the most relevant features from the acquired smart meter data set, which consequently led to lowering the computational time and enhancing the proposed classifier's learning capability. To classify data into genuine and theft consumers, the intelligence of the CatBoost algorithm was employed. Finally, the model's decision for a particular outcome was interpreted using the tree-SHAP algorithm. To prove the proposed framework's superior classification performance, its performance was compared with that of the latest gradient boosted machines and traditional ML models based on few of the well-known performance evaluation metrics. The proposed technique outperformed all the considered competing algorithms and achieved 93% accuracy, 92% recall and 95% precision.

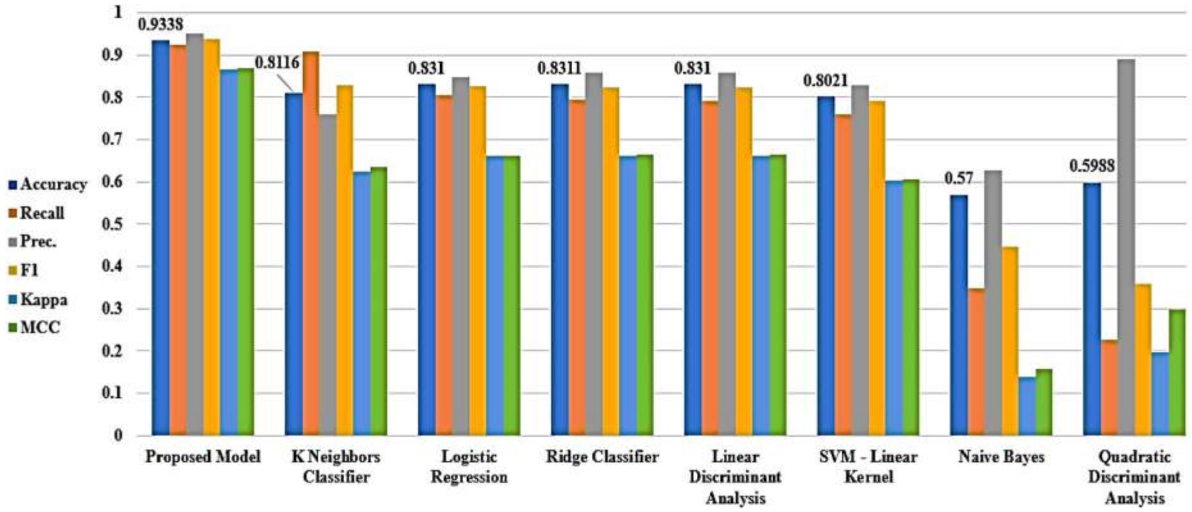


Fig. 12. Performance evaluation of studied conventional ML models.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Authors are agreed with this submission. They equally contributed in manuscript and its revision.

References

Adil, M., Javaid, N., Qasim, U., Ullah, I., Shafiq, M., Choi, J.-G., 2020. LSTM and bat-based RUSBoost approach for electricity theft detection. Appl. Sci. 10 (12), 4378. http://dx.doi.org/10.3390/app10124378, 2020-06-25.

Badrinath Krishna, V., Iyer, R.K., Sanders, W.H., 2016. ARIMA-Based Modeling and Validation of Consumption Readings in Power Grids. Springer International Publishing, pp. 199–210.

Batista, G.E., Prati, R.C., Monard, M.C., 2004a. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newsl. 6 (1), 20–29, %@ 1931-0145.

Batista, G.E.A.P.A., Prati, R.C., Monard, M.C., 2004b. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newsl. 6 (1), 20–29. http://dx.doi.org/10.1145/1007730.1007735.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Statist. 1165–1188, %@ 0090-5364.

Buzau, M.-M., Tejedor-Aguilera, J., Cruz-Romero, P., Gomez-Exposito, A., 2018a. Detection of non-technical losses using smart meter data and supervised learning. IEEE Trans. Smart Grid 1. http://dx.doi.org/10.1109/tsg.2018.2807925.

Buzau, M.M., Tejedor-Aguilera, J., Cruz-Romero, P., Gómez-Expósito, A., 2018b. Detection of non-technical losses using smart meter data and supervised learning. IEEE Trans. Smart Grid 10 (3), 2661–2670, %@ 1949-3053.

Chen, L., Chee-Wooi, T., Shiyan, H., 2013. Strategic FRTU deployment considering cybersecurity in secondary distribution network. 4, (3), pp. 1264–1274. http://dx.doi.org/10.1109/tsg.2013.2256939.

Christ, M., 2018. tsfresh, python library for FRESH algorithm-Documentation webpage. https://tsfresh.readthedocs.io/en/latest/index.html. (Accessed).

Christ, M., Kempa-Liehr, A., Feindt, M., 2016. Distributed and parallel time series feature extraction for industrial big data applications. arXiv, vol. abs/1610.07717.

Daniele, M.-B., 2001. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. SIGKDD Explor. Newsl. 3 (1), 27–32. http://dx.doi.org/10.1145/507533.507538, %@ 1931-0145.

Ferreira, A.M.S., Cavalcante, C.A.M.T., Fontes, C.H.O., Marambio, J.E.S., 2013. A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector. Int. J. Electr. Power Energy Syst. 53, 824–831. http://dx.doi.org/10.1016/j.ijepes.2013.06.001.

Gunturi, S.K., Sarkar, D., 2020. Ensemble machine learning models for the detection of energy theft. Electr. Power Syst. Res. http://dx.doi.org/10.1016/j.epsr.2020.106904.

Hasan, M., Toma, R.N., Nahid, A.-A., Islam, M.M., Kim, J.-M., 2019. Electricity theft detection in smart grid systems: a CNN-LSTM based approach. Energies 12 (17), 3310.

Hussain, S., Mustafa, M.W., Jumani, T.A., Baloch, S.K., Saeed, M.S., 2020. A novel unsupervised feature-based approach for electricity theft detection using robust PCA and outlier removal clustering algorithm. Int. Trans. Electr. Energy Syst. 30 (11), e12572, %@ 2050-7038.

Jaiswal, S., Ballal, M.S., 2020. Fuzzy inference based electricity theft prevention system to restrict direct tapping over distribution line. J. Electr. Eng. Technol. 15 (3), 1095–1106. http://dx.doi.org/10.1007/s42835-020-00408-7.

Jindal, A., Dua, A., Kaur, K., Singh, M., Kumar, N., Mishra, S., 2016. Decision tree and SVM-based data analytics for theft detection in smart grid. IEEE Trans. Ind. Inf. 12 (3), 1005–1016, %@ 1551-3203.

Joenssen, D.W., Bankhofer, U., 2012. Hot deck methods for imputing missing data. In: Machine Learning and Data Mining in Pattern Recognition. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 63–75, %@ 978-3-642-31537-4.

Jokar, P., Arianpoo, N., Leung, V.C.M., 2016. Electricity theft detection in AMI using customers' consumption patterns. IEEE Trans. Smart Grid 7 (1), 216–226. http://dx.doi.org/10.1109/tsg.2015.2425222.

Ke, G., et al., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, Vol. 30. NIPS 2017. pp. 3146–3154.

Lundberg, S.M., et al., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 (1), 56–67. http://dx.doi.org/10.1038/s42256-019-0138-9.

Messinis, G.M., Hatziargyriou, N.D., 2018. Review of non-technical loss detection methods. Electr. Power Syst. Res. 158, 250–266. http://dx.doi.org/10.1016/j.epsr.2018.01.005.

Molnar, C., 2018. A guide for making black box models explainable. URL: https://christophm.github.io/interpretable-ml-book.

Mwaura, F.M., 2012. Adopting electricity prepayment billing system to reduce non-technical energy losses in Uganda: Lesson from Rwanda. 23, pp. 72–79. http://dx.doi.org/10.1016/j.jup.2012.05.004.

Never, B., 2015. Social norms, trust and control of power theft in Uganda: Does bulk metering work for MSEs? Energy Policy 82, 197–206. http://dx.doi.org/10.1016/j.enpol.2015.03.020.

Northeast Group, 2017. Electricity Theft and Non-Technical Losses: Global Markets, Solutions and Vendors, 2017. Northeast Group, LLC, [Online]. Available: http://www.northeast-group.com/reports/Brochure-Electricity%20Theft%20&%20Non-Technical%20Losses%20-%20Northeast%20Group.pdf.

Passos Júnior, L.A., et al., 2016. Unsupervised non-technical losses identification through optimum-path forest. Electr. Power Syst. Res. 140, 413–423. http://dx.doi.org/10.1016/j.epsr.2016.05.036.

Pedregosa, F., et al., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830, %@ 1532-4435.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. pp. 6638–6648, https://arxiv.org/abs/1810.11363v1.

Punmiya, R., Choe, S., 2019. Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing. IEEE Trans. Smart Grid 10 (2), 2326–2329. http://dx.doi.org/10.1109/tsg.2019.2892595.

Roth, P.L., Switzer, F.S., 1995. A Monte Carlo analysis of missing data techniques in a HRM setting. J. Manage. 21 (5), 1003–1023. http://dx.doi.org/10.1177/014920639502100511, %U https://journals.sagepub.com/doi/abs/10.1177/014920639502100511.

Rusitschka, S., Eger, K., Gerdes, C., 2010. Smart Grid Data Cloud: A Model for Utilizing Cloud Computing in the Smart Grid Domain. IEEE, http://dx.doi.org/10.1109/smartgrid.2010.5622089, [Online]. Available: https://doi.org/10.1109/smartgrid.2010.5622089.

Saad, M., Tariq, M.F., Nawaz, A., Jamal, M.Y., 2017. Theft Detection Based GSM Prepaid Electricity System. IEEE, http://dx.doi.org/10.1109/ccsse.2017.8087973, [Online]. Available: https://doi.org/10.1109/ccsse.2017.8087973.

Saeed, M.S., Mustafa, M.W., Sheikh, U.U., Jumani, T.A., Mirjat, N.H., 2019. Ensemble bagged tree based classification for reducing non-technical losses in multan electric power company of Pakistan. Electronics 8 (8), 860.

Salman Saeed, M., et al., 2020. An efficient boosted C5.0 decision-tree-based classification approach for detecting non-technical losses in power utilities. Energies 13 (12), 3242. http://dx.doi.org/10.3390/en13123242.

Troyanskaya, O., et al., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17 (6), 520–525. http://dx.doi.org/10.1093/bioinformatics/17.6.520.

Tureczek, A.M., Nielsen, P.S., 2017. Structured literature review of electricity consumption classification using smart meter data. Energies 10 (5), 584.

Viegas, J.L., Esteves, P.R., Melício, R., Mendes, V.M.F., Vieira, S.M., 2017. Solutions for detection of non-technical losses in the electricity grid: A review. Renew. Sustain. Energy Rev. 80, 1256–1268. http://dx.doi.org/10.1016/j.rser.2017.05.193.

Winther, T., 2012. Electricity theft as a relational issue: A comparative look at Zanzibar, Tanzania, and the Sunderban Islands, India. Energy Sustain. Dev. 16 (1), 111–119. http://dx.doi.org/10.1016/j.esd.2011.11.002.

Xiao, Z., Xiao, Y., Du, D.H.-C., 2013. Exploring malicious meter inspection in neighborhood area smart grids. IEEE Trans. Smart Grid 4 (1), 214–226. http://dx.doi.org/10.1109/tsg.2012.2229397.

Yurtseven, Ç., 2015. The causes of electricity theft: An econometric analysis of the case of Turkey. Util. Policy 37, 70–78. http://dx.doi.org/10.1016/j.jup.2015.06.008.

Zhang, S., Zhang, J., Zhu, X., Qin, Y., Zhang, C., 2008. Missing value imputation based on data clustering. In: Transactions on Computational Science I. Springer, pp. 128–138.

Zheng, Z., Yang, Y., Niu, X., Dai, H.-N., Zhou, Y., 2018. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. IEEE Trans. Ind. Inf. 14 (4), 1606–1615. http://dx.doi.org/10.1109/tii.2017.2785963.