energies

MDPI

Article

Synthetic Theft Attacks and Long Short Term Memory-Based Preprocessing for Electricity Theft Detection Using Gated Recurrent Unit

Pamir [1], Nadeem Javaid [1,2,*], Saher Javaid [3,*], Muhammad Asif [1], Muhammad Umar Javed [1], Adamu Sani Yahaya [1] and Sheraz Aslam [4,*]

[1] Department of Computer Science, COMSATS University Islamabad, Islamabad 44000, Pakistan; pamirshams2011@yahoo.com (P.); muhammad.asif.comsat@gmail.com (M.A.); umarkhokhar1091@gmail.com (M.U.J.); asyahaya.it@buk.edu.ng (A.S.Y.)
[2] School of Computer Science, University of Technology Sydney, Ultimo, NSW 2007, Australia
[3] Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi City 923-1292, Ishikawa, Japan
[4] Department of Electrical Engineering, Computer Engineering and Informatics, Cyprus University of Technology, 3036 Limassol, Cyprus
* Correspondence: nadeemjavaidqau@gmail.com (N.J.); saher@jaist.ac.jp (S.J.); sheraz.aslam@cut.ac.cy (S.A.)

Abstract: Electricity theft is one of the challenging problems in smart grids. The power utilities around the globe face huge economic loss due to ET. The traditional electricity theft detection (ETD) models confront several challenges, such as highly imbalance distribution of electricity consumption data, curse of dimensionality and inevitable effects of non-malicious factors. To cope with the aforementioned concerns, this paper presents a novel ETD strategy for smart grids based on theft attacks, long short-term memory (LSTM) and gated recurrent unit (GRU) called TLGRU. It includes three subunits: (1) synthetic theft attacks based data balancing, (2) LSTM based feature extraction, and (3) GRU based theft classification. GRU is used for drift identification. It stores and extracts the long-term dependency in the power consumption data. It is beneficial for drift identification. In this way, a minimum false positive rate (FPR) is obtained. Moreover, dropout regularization and Adam optimizer are added in GRU for tackling overfitting and trapping model in the local minima, respectively. The proposed TLGRU model uses the realistic EC profiles of the Chinese power utility state grid corporation of China for analysis and to solve the ETD problem. From the simulation results, it is exhibited that 1% FPR, 97.96% precision, 91.56% accuracy, and 91.68% area under curve for ETD are obtained by the proposed model. The proposed model outperforms the existing models in terms of ETD.

Keywords: theft attacks; long short term memory; gated recurrent unit; deep learning techniques; machine learning techniques; electricity theft detection; smart grids

Citation: Pamir; Javaid, N.; Javaid, S.; Asif, M.; Javed, M.U.; Yahaya, A.S.; Aslam, S. Synthetic Theft Attacks and Long Short Term Memory-Based Preprocessing for Electricity Theft Detection Using Gated Recurrent Unit. *Energies* **2022**, *15*, 2778. https://doi.org/10.3390/en15082778

Academic Editors: Miguel Jiménez Carrizosa and Abu-Siada Ahmed

Received: 31 December 2021
Accepted: 30 March 2022
Published: 10 April 2022

1. Introduction

From global perspective, the traditional metering system is still a commonly applied system, especially in the residential sector [1]. However, the electrical meters used in the traditional metering system, i.e., electromechanical meters, do not perform as accurately as expected and their measurement ability is mostly affected by the waveform distortion, operating temperature, and other factors. Moreover, this category of meters allows unidirectional communication [2]. In addition, the consumed electricity measurement needs manual reading by the electric utility personnel, in which there are many chances of measurement errors. Furthermore, it should also be considered that manual meter readings lead the utilities to incur high operational cost, which is in fact charged from the energy users. Therefore, the advanced metering infrastructure [3] is introduced to overcome the issues caused by the conventional metering systems.

The electricity theft detection (ETD) is one of the major issues and a trending research area in the current era. It lies in the category of non technical losses (NTLs). Generally, NTLs along with technical losses (TLs) are the categories in which losses of electricity are grouped [4]. Majorly, TLs occur in power system's equipment owing their resistance against the power flow. While NTLs arise because of the electricity theft in terms of meter hacking, bypassing, or tampering. The theft of electricity leads to many dangerous issues like huge economic loss, operational inefficiency in electric grids, public safety hazards, etc. The economic loss arises due to the electricity theft that amounts around 100 million Canadian dollars per year according to the British Columbia power and hydro authority [5]. In addition, the revenue loss incurred due to the NTL throughout the world is around 96 billion USD yearly [6]. Hence, it is very crucial to have an efficient and effective ETD approach in the smart grids (SGs).

The detection approaches of the electricity theft used in the literature are grouped into three main categories: the hardware based approaches, the classification based approaches, and the game-theory based approaches. The ETD approaches based on the hardware devices [7,8] employ different hardware equipment to obtain higher theft detection accuracy. However, these techniques require high monetary cost for the installation and maintenance of the hardware equipment. ETD is referred as a game, in game-theory based approaches, between two players [9,10] where both players try to optimize their utility functions. Furthermore, a zero-sum game is introduced among the power entities for achieving the equilibrium state. These methods do not require extra payment. However, they are still not a suitable and optimal remedy to minimize electricity theft because the formulation of a suitable utility function in a real environment is a tiresome job in hand.

Several deep learning (DL) and machine learning (ML) based classification models are developed in [5,6,11–16] for ETD and they use energy consumption data stored in smart meters (SMs). Therefore, their costs are reasonable. However, there are some issues with the existing DL and ML based classification models, which negatively affect the classifiers in terms of false positive rate (FPR) and true positive rate (TPR). One crucial problem that causes the DL and ML based classifiers to perform poorly in detecting the electricity theft is imbalanced class problem. The imbalanced class or data imbalanced issue denotes that the count of the data points related to the abnormal consumers is not equal to the count of the normal consumers' present in a dataset. The data related to the normal electricity consumers is easily available in comparison with abnormal class data because the abnormal data are gathered in a limited amount from the real environment. Hence, the problem of data imbalance leads the DL and ML classifiers to be biased towards the majority class (normal users) when performing classification, which results in high FPR. In addition, another crucial issue that negatively affects the classification algorithms in terms of TPR, FPR, and overfitting is the curse of dimensionality. It occurs while dealing with the data of high dimensionality. In other words, the curse of dimensionality refers to a principle in which the increment in number of features (dimensions) is directly proportional to the increment in the classification error. Furthermore, another crucial issue that affects the ML and DL classification algorithms negatively is ignoring the drift. It refers to the irregular consumption of the electricity that occurs from non theft factors like changes in the number of family members, seasonal changes, changes in electric appliances in terms of their type or number, etc.

This paper presents the extended version of the work already published in [17]. This work uses six theft attacks (TAs) to produce theft data samples for balancing the data. The combined model having TAs, long short-term memory (LSTM) [18], and gated recurrent unit (GRU) [19], termed as TLGRU, is proposed for efficient ETD. Moreover, the proposed approach learns and pinpoints the real abnormal consumers instead of pinpointing the drift as theft. In this way, it reduces the FPR. The research article comprises the following major contributions.

- A TLGRU model is proposed for effective and reliable ETD in SGs. In the proposed model, the synthetic TAs are implemented to generate theft samples in the dataset

acquired from state grid corporation of China (SGCC) for tackling imbalance problem. Moreover, LSTM is employed to efficiently extract and maintain the vital characteristics from the huge time series data, which handle curse of dimensionality problem.
- A powerful recurrent memory network, termed as GRU, is utilized to initially investigate the electricity consumption (EC) profiles of consumers and then tackle the problem of drift accordingly.
- An efficient regularization technique, known as dropout, is integrated in the proposed TLGRU model to avoid overfitting and increase the convergence speed.

The remaining sections of the article are arranged as discussed. Section 2 comprises the study of the existing literature. While, the subject matter of Section 3 is the proposed system model. The outcomes obtained after performing extensive simulations are elaborated in Section 4. Finally, the concluding remarks are given in Section 5 .

2. Literature Review

Tuning the hyperparameters, data imbalance, ensuring privacy, and the dimensionality curse problems are the four broad categories in which the existing literature is divided and studied into four groups in this section. The research articles that deal with the hyperparameters' tuning of the ML techniques are given in initial category. In [4–6,11–14], those ML and DL techniques are under consideration that deal with efficient tuning of hyperparameters. In [4], a stacked sparse denoising autoencoder (SSDAE) is proposed to extract the most effective features to deal with the FPR and generalization issues. The low value for FPR, high detection rate, high robustness, and important feature extraction are achieved by introducing the noise and sparsity parameters into SSDAE. The hyperparameters of SSDAE are tuned using particle swarm optimization algorithm. Moreover, the SGCC hourly data are used for analysis in this work. Furthermore, in [5], to detect electricity theft, a wide and deep convolutional neural network (WDCNN) is proposed. The theft detection is performed by using both one dimensional and two dimensional data for model training and model testing. The wide component is used to process the global features and the deep component is utilized to find whether the periodicity exists between EC patterns or not. Moreover, the behavior of the data are checked using a statistical technique, i.e., pearson correlation coefficient. The proposed classifier is validated using area under the curve (AUC) and mean average precision (MAP).

However, manual hyperparameter tuning is done in the work, which reduces the accuracy and efficiency of the classifier.

To ensure correct identification and detection of the theft patterns, a hybrid model of two deep neural networks is proposed in [6]. The hybrid model comprises of multi layer perceptron (MLP) and LSTM. LSTM is responsible for analyzing EC data while MLP deals with the exogenous variables. Clearly, it is proved that the hybrid model outperforms the benchmark models. However, the proposed hybrid model is prone to overfitting issue due to the full connectivity of neurons. Furthermore, the class imbalance problem is not tackled, which reduces the model's generalization ability.

The adjustment of the hyperparameters' values and the imbalanced data problem are tackled by the studies conducted in [11–13]. The authors in [11] propose a consumption pattern based ETD (CPBETD) model. The data imbalance problem is resolved by generating a synthetic attacks' dataset. In the model, low sampling rate of EC values is considered to preserve customers' privacy. Furthermore, a sustainable energy authority of Ireland (SEAI) dataset is employed to check the performance of the model. The proposed model is validated using FPR, DR, and bayesian detection rate (BDR) performance metrics. However, the feature extraction step is not considered, which increases the computational complexity. One of the main performance indicators, i.e., accuracy is ignored in this paper. The authors in [12] utilize ensemble learning models for ETD. The ensemble ML models employed in the study are extreme gradient boosting (XGBoost), random forest (RF), adaptive boosting (AdaBoost), light gradient boosting (LGB), extra trees, and categorical boosting (CatBoost). The commission for energy regulation (CER) data are used for models'

evaluation. The data preprocessing step is also performed for the TPR improvement. The class imbalance problem is tackled using SMOTE. However, it requires high computational cost for training and testing processes. Another issue with synthetic minority oversampling technique (SMOTE) is that it lacks in capturing probability distribution curve from the complex EC data, which degrades the generalization ability of the classifiers. Similarly, the authors in [13] employ a variant of generative adversarial network (GAN) [20], named as wasserstein GAN (WGAN), and K-nearest neighbor for balancing imbalanced data and classifying data points that are near to support vector machine (SVM's) hyperplane, respectively. Moreover, a combined technique of supervised and unsupervised methods is proposed, i.e., decision tree combined with the K-nearest neighbor SVM (DT-KSVM) and WGAN techniques. In [14], the authors propose an XGBoost technique for detecting anomalies present in the dataset of SMs. The hyperparameters' values are adjusted using a grid search method in this paper. However, the grid search method is computationally expensive. Moreover, feature extraction is not efficiently performed in this work.

The second group of literature review focuses on imbalanced problem in SM data. In [15,21–25], the data imbalance problem is resolved by the authors using several sampling techniques. In [15], the authors tackle the data imbalance problem using a hybrid of k-means SMOTE (K-SMOTE) technique while RF is used for theft classification. The proposed model is evaluated and validated using the EC data obtained from Hebei province of China. The authors in [21] propose a framework that consists of maximal overlap discrete wavelet packet transform (MODWPT) and random undersampling boosting (RUSBoost) for feature extraction and theft classification, respectively. The data imbalance problem is tackled using a random undersampling (RUS) technique. The EC data of commercial and industrial users from Honduras is employed for the evaluation of the framework. Further, the proposed framework is validated using AUC, Matthews correlation coefficient (MCC), accuracy, precision, recall, F_β score, and specificity. However, some important information is lost because of random removal of observations from the majority class using RUS. This information loss results in a high FPR. In addition, the authors in [22] propose a combined model of CNN and LSTM for detecting energy theft in SGs. Furthermore, the class imbalance problem in the EC data is resolved using an oversampling technique, known as SMOTE. Eventually, validation of the proposed model is performed using MCC, recall, F1-score, accuracy, and precision. In addition, the EC data of Multan electric power company (MEPCO) is considered for conducting simulations. However, SMOTE based synthetic data generation leads to a class overlapping problem. Furthermore, in [23], the authors propose a methodology based on ensemble bagged tree (EBT) for detecting the fraudulent electricity consumers.

The research articles in the third group consider the electricity users' privacy preservation problem. The authors in [16,26–28] focus on dealing with the maintenance of the consumers' privacy. The study in [16] proposes a semi supervised auto encoder (SSAE) for extracting significant features from the SM data of the industrial consumers. The proposed SSAE model, for addressing the overfitting issue, makes use of the data that is unlabeled. Along with that, an adversarial module is also used. The proposed model's exceptional performance with a small set of samples and preservation of consumers' privacy are exhibited through simulations. In addition, the research conducted in [26] proposes an efficient ETD model. Privacy preservation is ensured using a functional encryption (FE) method used by the SMs of the users to encrypt their readings. The issue of imbalance between classes is tackled through adaptive synthetic (ADASYN) and by creating a malicious attack dataset. Similarly, in [27], the authors propose a privacy preserving based ETD (PPETD) algorithm. The generalized CNN (GCNN) classifier is employed for ETD using the encrypted small-sized consumption slots data. However, it has high computational complexity due to training of the excessive parameters of GCNN. Furthermore, in [28], a multiple linear regression model (MLRM) is proposed for NTL detection in SGs. A significant benefit of this method is that it detects electricity theft after performing comparison between the data recorded by an SM and the collector (a sensing device attached with the SM) without violating the customers' privacy.

The research articles in the final category deal with the curse of dimensionality issue using different feature generation, feature selection, and feature extraction strategies. The authors in [29] employ feature generation using mean, standard deviation, and minimum and maximum values of features. They develop a gradient boosting theft detector (GBTD) for electricity theft identification in SGs. A framework of practical feature engineering is proposed in [30]. The framework is the combination of the finite mixture model (FMM) based clustering for segmentation of the customers and genetic algorithm (GA) based feature generation from one or more already available features to improve prediction accuracy. A gradient boosting machine is used for classification. The model's validation is done using various performance measures. Furthermore, the authors in [31] focus on the important features' selection for pinpointing the anomalous consumers. They practice binary black hole algorithm (BBHA), which is a metaheuristic technique, for feature selection. The proposed algorithm is validated using accuracy and execution time metrics, which are not sufficient for a fair evaluation. Moreover, many novel optimization algorithms are developed that can be used for selecting the most suitable features in order to achieve better performance as compared to BBHA in terms of accuracy and execution time. In [32], a probabilistic technique is put forward for classifying patterns along with the mathematical formulation on the levenberg-marquardt technique's basis. Feature extraction is performed using the encoding algorithm. However, the model input parameters' tuning is ignored, which becomes the reason for overfitting.

The summarized form of the available literature review is given in Table 1. Moreover, the problems addressed from the aforementioned literature and motivation of this work are finalized as follows.

Table 1. Literature review summary.

Methodology	Objectives	Dataset	Performance Metrics	Limitations
SSDAE [4]	To tackle NTLs	SGCC hourly data	FPR, TPR and AUC	Inadequate evaluation metrics
WDCNN [5]	To secure SGs by detecting electricity theft	SGCC daily data	AUC and MAP	Data imbalance issue
LSTM-MLP [6]	To overcome NTLs	Endesa	AUC, precision, recall and precision-recall-AUC	Data imbalance issue
CPBETD [11]	To improve ETD performance	SEAI	TPR, FPR and BDR	No feature extraction is performed
RF, AdaBoost, XGBoost, LGB, ensemble tree and CatBoost [12]	To detect energy theft in power grids	CER	Precision, AUC and accuracy	Ensemble techniques are computationally complex
DT-KSVM [13]	To decrease power losses	SEAI	AUC and accuracy	Inadequate performance metrics
XGBoost [14]	To enhance ETD performance	Endesa	AUC, precision-recall and execution time	High computational time
RF [15]	To detect NTL behavior	Hebei province	AUC and accuracy	No feature extraction is done

Table 1. *Cont.*

Methodology	Objectives	Dataset	Performance Metrics	Limitations
SSAE [16]	To reduce NTLs by employing semi-supervised data	SGCC daily data	Accuracy, TPR, FPR, precision, recall and F1-score	Inappropriate hyperparameter tuning
MODWPT, RUSBoost [21]	To reduce NTLs	Honduras	F1-score, MCC, precision, recall, AUC and accuracy	Important information is lost due to RUS
CNN-LSTM [22]	To detect abnormal EC profiles of consumers	SGCC daily data	F1-score., MCC, precision, recall and accuracy	Classes overlap due to SMOTE
EBT [23]	To minimize NTLs	MEPCO	Accuracy, sensitivity, specificity, F1-score and FPR	Curse of dimensionality problem is not tackled
ETDFE [26]	To detect ET by preserving consumers' privacy	CER	Highest difference (HD), FPR, DR and Accuracy	High computational complexity due to improper hyperparameter optimization
PPETD [27]	To perform ETD while maintaining consumers' privacy	CER	HD, DR and FPR	Improper hyperparameter tuning
MLRM [28]	To overcome NTLs	Neighborhood area network dataset	Accuracy, sensitivity and specificity	Curse of dimensionality problem is not handled

The authors of [4] introduce SSDAE for ETD in SGs. The principal challenge they considered is high value of FPR because of the low generalization of classification techniques. The CPBETD technique is presented in [11]. CPBETD uses the data of the electric users as well as the transformer meters for ETD. The imbalanced data issue is handled by creating a synthetic TA dataset. Finally, motivated from [4,11], we focused on anomalies' prevention and detection due to nonmalicious intermediaries (drift), curse of dimensionality, and dealing with the data imbalance problem. The authors in [5,6] introduce new WDCNN [33] and LSTM-MLP based techniques for electricity theft classification, respectively. However, the unavoidable issue of data imbalance is not considered that leads to classifier's biasness with respect to the normal class that causes high FPR. In addition, the high value of the FPR is neglected by these research articles. In addition, in the wide module of the WDCNN model, only one layer, i.e., fully connected layer, is employed, which causes the model to be trapped into the local optima. Furthermore, the authors in [11] present CPBETD for efficient detection of NTL as previously metioned. Moreover, CPBETD obtained better performance even with a low granularity of EC data, which assists to maintain consumers' privacy. However, the curse of dimensionality issue is ignored, which leads to high FPR and overfitting issues.

3. Proposed System Model

The model designed for the underlying work is made of two main units and some subunits. The main units are (1) preprocessing unit and (2) theft classification unit. The details of the units and their relevant subunits are given in the subsequent sections. Figure 1 exhibits the graphical view of the system model.

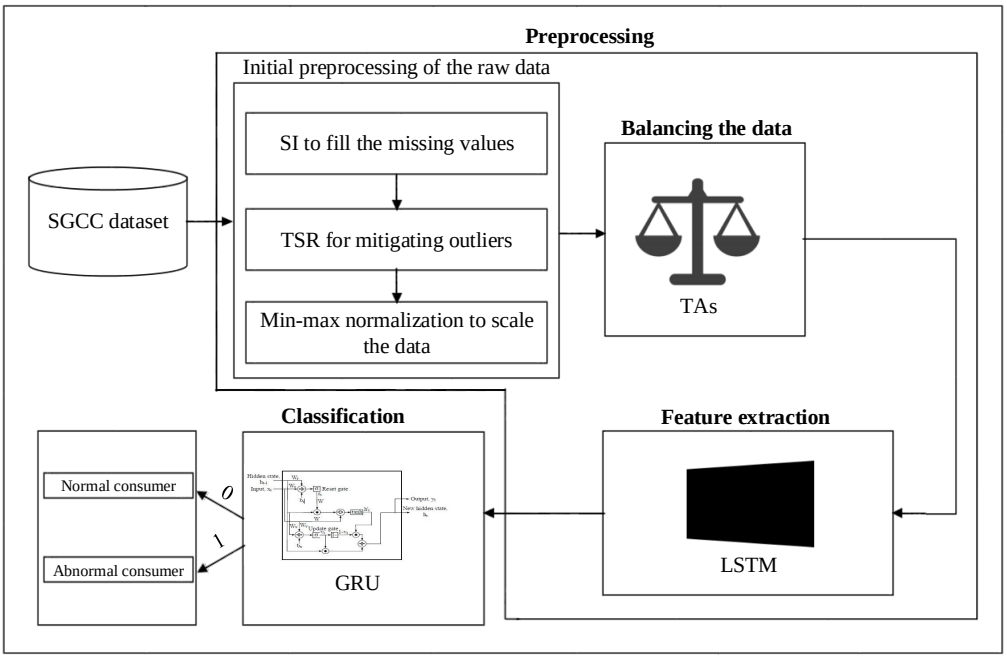

Figure 1. Overview of the proposed ETD model.

3.1. Dataset Description

The data of energy consumption acquired from SGCC (which is both realistic and easily accessible) is employed for the proposed model's validation via different performance metrics [5]. The SGCC possesses imbalanced EC data.

The total number of consumers' data records is 42,372 of which 38,752 are the normal users' records and 3615 are theft consumers' records. The sampling frequency of the dataset is set to daily. In the dataset, the overall EC values are represented in terms of rows while the EC value on a specific day is given in terms of a column. Furthermore, the data are collected after conducting onsite inspections. So, it contains NaN values, outliers, and data being dispersed on a huge scale. These abnormalities should be treated before proceeding to the development of the ETD model. In this regard, to recover the missing values, mitigate the outliers, and scale the data in a specific range, the preprocessing step is required, which is discussed in detail in the next subsection. Before preprocessing, the total number of consumers' data is 42,372, whereas, after the preprocessing, 5 rows are dropped by the simple imputer (SI) method because all the values in these rows are NaN values. In such cases, the SI does not know what value is to be imputed. The imputer will impute some values instead of deleting a record if it finds atleast one non-NaN value in the targeted record. It is also important to note that SI method works column-wise, so you need to take the transpose of your data before applying the imputer method. After imputation, again take the transpose of the data to revert it to the original shape. Table 2 presents details of the used dataset.

Table 2. Dataset detail.

Dataset Description	Values
Dataset acquisition intervals	2014–2016
Total abnormal users count before the data balancing	3615
Total benign users count before the data balancing	38,752
Total abnormal users count after the data balancing	21,183
Total benign users count after the data balancing	21,184
Total users count before the initial preprocessing of raw data	42,372
Total users count after the initial preprocessing of raw data	42,367

3.2. Preprocessing Unit

The data balancing, initial preprocessing and the feature extraction are the subunits of the preprocessing unit. These subunits are elaborated below.

3.2.1. Initial Preprocessing of the Raw Data

The preprocessing of the raw data is very important as the performance of any model is not only limited to the classification of electricity theft using ML models, but related to the data quality as well. Generally, the consumption data stored by the electric meters mostly has the missing values or the outliers. In our case, we considered the real EC data from a Chinese dataset, i.e., SGCC [5], that also contains the outliers and the missing values. The values exist because of various causes like the unreliable dispatch of the consumption data, the faulty meters, the storage related problems, etc., [5]. We utilized an SI technique for computing the mean of the consumption data present in the previous and the next cells to deal with the missing data by. The SI working mechanism is taken from [34] and is given in Equation (1).

$$f(x_{i,s}) = \begin{cases} \frac{x_{i,s-1}+x_{i,s+1}}{2}, & x_{i,s} \in NaN, x_{i,s-1}, \\ & x_{i,s+1} \notin NaN \\ 0, & x_{i,s} \in NaN, x_{i,s-1} \text{ or} \\ & x_{i,s+1} \in NaN, \\ x_{i,s}, & Otherwise, \end{cases} \tag{1}$$

where, i and s show a specific electricity customer and a time slot (day), respectively. $x_{i,s-1}$ and $x_{i,s+1}$ denote the EC data of a consumer for the previous day and the next day, respectively. Not a number (NaN) represents the missing values.

The availability of the outliers in the dataset negatively affects the classifier's performance and maximizes the FPR value. Hence, the outliers need to be removed from the dataset. Therefore, we use the three-sigma rule (TSR) of thumb method to remove the outliers from the dataset. The mathematical representation of TSR is taken from [5] and is given in Equation (2).

$$f(x_{i,s}) = \begin{cases} avg(X) + 2.\sigma(X), & x_{i,s} > avg(X) + 2.\sigma(X), \\ x_{i,s}, & Otherwise, \end{cases} \tag{2}$$

where, X shows a vector that is made of multiple $x_{i,s}$ values. The term $avg(X)$ and $\sigma(X)$ represent the average and standard deviation of X, respectively.

As up to now, the NaN and outlier values are successfully dealt with, so now, the dataset normalization is required because DL techniques are sensitive to the sparsed, diversed, and unscaled data. Therefore, we use min-max data scaling method in order to scale or normalize the data. The data scaling is performed using the Equation (3) [5].

$$f(x_{i,s}) = \frac{x_{i,s} - min(X)}{max(X) - min(X)} \tag{3}$$

where, $min(X)$ and $max(X)$ functions return the minimum and the maximum values of vector X, respectively.

3.2.2. Data Balancing by TAs' Implementation

The synthetic TAs are employed for balancing the data in the SGCC dataset. All the six TAs are introduced in [11] while the updated and revised version of the attacks are introduced in [29]. We select the updated TAs to create more practical and real abnormal consumption patterns to balance the dataset. The real consumption of a user is denoted by e_t, where, $(t \in [0, 1034])$. In this study, the employed dataset contains the total of 1035 days' consumption data. The mathematical representations of all TAs are taken from [29] and are presented in Equations (4)–(9).

$$t1(x_t) = x_t * random(0.1, 0.9), \tag{4}$$

$$t2(x_t) = x_t * r_t, r_t = random(0.1, 1), \tag{5}$$

$$t3(x_t) = x_t * random[0, 1], \tag{6}$$

$$t4(x_t) = mean(X) * random(0.1, 1), \tag{7}$$

$$t5(x_t) = mean(X), \tag{8}$$

$$t6(x_t) = x_{1034-t}, \tag{9}$$

where, $X = \{x_1, x_2, ..., x_{1034}\}$. In theft attack 1, $t1(.)$ multiplies the complete row (actual reading) by the same randomly generated value between 0.1 and 0.9. It is argued in [29] that it is not necessary that a theft might occur continuously in the real world but some discontinuous values may also be reported by the theft. Therefore, in theft attack 2, $t2(.)$ multiples each timestamp in a row with a different random value ranges from (and including) 0.1 to 1. Here, when the upper limit (1) of the random number is generated, the actual reading of that particular timestamp will be reported as theft. In theft attack 3, the theft consumer sometimes reports the real EC value and sometimes reports a 0 value. In theft attacks 4 and 5, the mean value of the actual readings is involved. Where, in attack 4, the mean value is multiplied with the same randomly generated value. Finally, theft attack 6 mimics the behaviour of a theft consumer when it sends the actual readings in a reverse order.

We applied the TAs on benign users' consumption data to establish an appropriate balance between the theft and honest data points. The SGCC dataset has 3615 abnormal and 38,757 normal consumers' data instances out of total of 42,372 records. The ratio of the normal and abnormal consumers in the dataset is 1:9. There are large number of data points in the dataset and it is difficult to use all of them for analysis due to the higher computational complexity problem. So, as a sample, 9999 records out of the 42,372 are employed for analysis. In this way, the last 900 real theft records out of 3615 (2714–3615) are selected and the other real theft records (0–2713) are ignored. Moreover, the remaining deficient abnormal records (4098) are synthetically created using the TAs, generation over the benign consumers' data starting from 901th and ending at 4998th record. Attacks 1–6 are implemented on the benign consumers' data ranges 901–1583, 1584–2266, 2267–2949, 2950–3632, 3633–4315, and 4316–4998, respectively. The TAs' pattern and the normal user's consumption pattern are shown in Figures 2 and 3, respectively. Furthermore, the data from 4999–9998 is the benign consumers' data. So now, the data are balanced (where 0–4998 is theft consumers' data and 4999–9998 is normal consumers' data) and is forwarded to LSTM for extracting the necessary features from them, which improves the classification performance. Finally, the GRU classifier is trained using the extracted data received for efficient and effective electricity theft classification. The dataset contains 1035 days' energy consumption data and the attacks are implemented on the entire data. However, in the

figures referred above, only 30 days' synthetic TAs and normal patterns are given as an example.

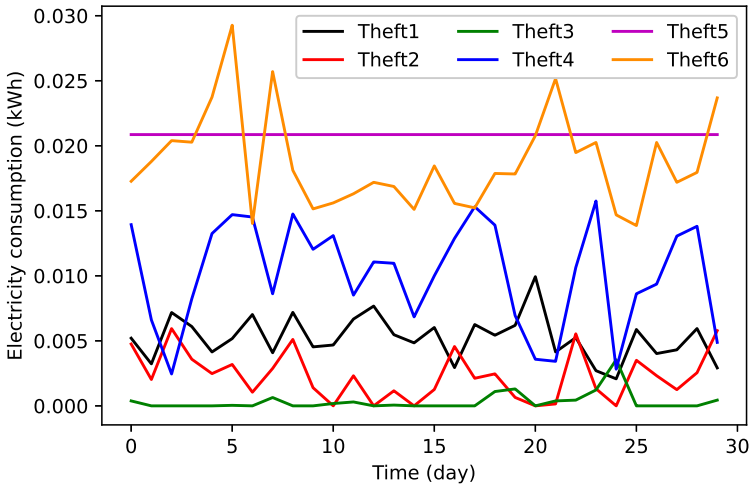

Figure 2. Synthetic theft attack patterns.

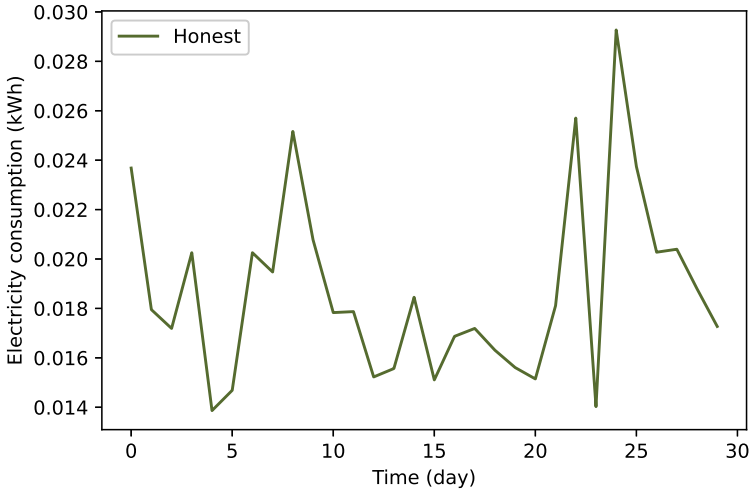

Figure 3. Normal pattern.

3.2.3. Long Short Term Memory Based Feature Extraction

Once the initial preprocessing and balancing of the data are performed, LSTM [35] is used for extracting important features. SGCC dataset contains huge and high dimensional (features) data and we need to perform dimensionality reduction. In such situations, the conventional recurrent neural network (RNN) [36] can not be employed as it faces the problems of the vanishing gradient and exploding gradient while dealing with the huge amount of data for dimensionality reduction. LSTM, an advanced RNN variant, is employed widely for successfully dealing with exploding and vanishing gradients issue. In training process, the temporal correlation between the current input and the previous state is found via using the previous data by RNN, and the output is finalized. However, because of its short and temporary memory, it fails to re-achieve and re-gain the previous information for the huge time series data and therefore, it fails to capture the temporal correlation between the current and the previous states. On the other hand, LSTM is able to easily and smoothly capture the temporal correlation. It also helps in dimensionality reduction of the huge time series data. The reason is that it has special and unique memory cells, which make use of the historical information. Furthermore, the significant features

from the huge time series data are retained and memorized using the cells. This information is kept and maintained by the cell state (long term memory) in the LSTM. The significant features enfold the necessary information of the whole dataset.

The LSTM comprises the forget gate (f_t), the input gate (i_t), and the output gate (o_t). The determination and decision that whether the information taken from the current input (x_t) and the previous hidden state (h_{t-1}) should be retained or discarded from the cell state is made by the forget gate. In this way, the information from h_{t-1} and x_t are passed through the sigmoid (σ) activation function and it will decide either to keep or discard the previous output from the cell state by generating 1 or 0, respectively. The input gate determines that which values or data are employed for updating memory state or cell state, denoted by (C_t). Again the details from h_{t-1} and x_t are passed from the second sigmoid and decision will be made that what to do with the information; either discard it or save it. Similarly, in the cell state (C_t), the *tanh* activation function is applied on h_{t-1} and x_t, and the output is stored into the cell state. The results from the cell state and the input gate are multiplied and added with the multiplication result of the forget gate and cell state. It is finally stored into the current cell state C'_t, which is now an updated cell state. Similarly, the final output gate (o_t) takes the x_t and h_{t-1} as the inputs, applies the sigmoid operation on them and the final result is stored in o_t. In order to calculate the next hidden state (h_t), the multiplication of $tanh(C'_t)$ and o_t is performed, sigmoid is applied on their multiplication result and the final output is stored into h_t. Moreover, the mathematical formulations of the forget, input, and output gates are given in Equations (10)–(15) [37].

$$f_t = \sigma(W_f(x_t, h_{t-1}) + b_f), \tag{10}$$

$$i_t = \sigma(W_i(x_t, h_{t-1}) + b_i), \tag{11}$$

$$C_t = tanh(W_c(x_t, h_{t-1})), \tag{12}$$

$$C'_t = (f_t * (C_t)) + (i_t * (C_t)), \tag{13}$$

$$o_t = \sigma(W_o(x_t, h_{t-1}) + b_o), \tag{14}$$

$$h_t = \sigma(o_t * tanh(C'_t)). \tag{15}$$

where, b_o, b_i, and b_f are the biases for the output, input, and forget gates, respectively. The W_o, W_i, and W_f denote the weights of the output, input, and forget gates, respectively. Moving ahead, for denoting previous hidden state information along with the updated cell state information, C_t and C'_t are used, respectively.

The optimal adjustments of the hyperparameters' values are very important to attain better results for feature extraction using LSTM. In order to perform better feature extraction, we perform manual parameter tuning. We set 200 and 100 neurons for each LSTM's layer. Whereas, the dense or fully connected layer has only one neuron. The dropout layer value is set to be 0.2 in order to protect the proposed TLGRU model from overfitting. The more detailed picture of the hyperparameters' values are given in Table 3. In our proposed LSTM feature extractor, we employ six layers, i.e., two LSTM, two LeakyReLU, one BatchNormalization, and one Dropout layer. We use 200 neurons in the first LSTM layer, the learning rate (Alpha) for both LeakyReLU layers is chosen to be 0.001, 100 neurons are used for the second LSTM layer, and finally, 0.2 is selected as the dropout probability for Dropout layer.

Table 3. Optimal settings of the hyperparameters' values.

Hyperparameters	Values
Units	200 and 100
Alpha	0.001
Dropout	0.2

3.3. Theft Classification Unit

TLGRU based classification is put forward for detecting the thefts in in electricity usage in SG. Furthermore, the popular benchmark models, LR, DT, SVM and GRU, are used for performance comparison with the proposed classifier. The details of these classifiers are given in the following subsections.

3.3.1. GRU

It was developed in 2014 [38]. GRU is a sub module of our proposed TLGRU model. GRU is used in two ways: (1) as a sub module of our proposed TLGRU model to tackle drift, ovefitting, high FPR, and local minima trap problems, (2) benchmark method for comparison purpose. GRU is faster in comparison with RNN and LSTM with respect to training time. It is employed widely in other research areas [39–42]; but, it is rarely employed for efficient ETD.

GRU is mainly developed to deal with the vanishing gradient problem with which the RNN fails to deal. The GRU merges the forget gates and the input gate of the LSTM into one gate known as the update gate. Moreover, the combination of both cell state and hidden state is made in GRU. The GRU's basic architecture is illustrated in Figure 4 [39,40].

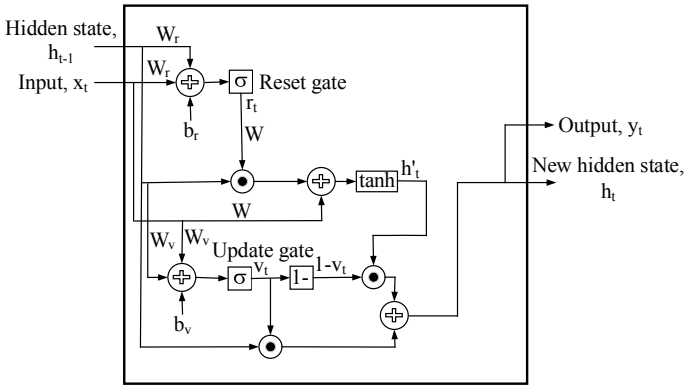

Figure 4. Gated recurrent unit architecture.

GRU consists of two gates: update gate (long term memory) and reset gate (short term memory), which are used to solve the gradient vanishing issue of RNN. These gates are two different vectors that finalize and determine that which information must be passed and proceeded to the output. One special and unique property about them is that the gates have the capability to be trained in order to retain information for a long duration, without discarding it with time or discarding information that is irrelevant and insignificant for prediction. By this feature GRU can clearly differentiate between the irregularity because of nonmalicious intermediaries (drift) and the irregularity due to the real theft factors, which consequently minimizes the FPR value. Therefore, we select the GRU classifier as the theft detection module of our proposed TLGRU classifier. The complete mathematical formulation of reset gate (r_t) and update gate (v_t) is given in Equations (16) and (17), respectively [38].

$$r_t = \sigma(W_r x_t + W_r h_{t-1} + b_r), \tag{16}$$

$$v_t = \sigma(W_v x_t + W_v h_{t-1} + b_v), \tag{17}$$

where, W_r, b_r, W_v, b_v, and x_t denote the weight related to reset gate, the bias value related to the reset gate, the weight related to update gate, bias related to the update gate, and the input vector, respectively. The reset gate is used in GRU to decide the amount of historical information to forget. Whereas, the update gate assists GRU to decide about previous information to be copied or passed to future. To compute the current or new hidden state (h_t), two steps are need to be performed. The first step is to calculate the candidate hidden

state (h'_t) while the second step is to compute the h_t. The mathematical forms for h'_t, and h_t are depicted in Equations (18) and (19), respectively [38].

$$h'_t = tanh(Wx_t + Wr_t \odot h_{t-1}), \tag{18}$$

$$h_t = v_t \odot h_{t-1} + (1 - v_t) \odot h'_t. \tag{19}$$

where, W and h_{t-1} denote the weights' values and the hidden state at the previous time-step, respectively. In addition, the symbol $\odot$ denotes the Hadamard product. The *tanh* and σ show the hyperbolic and sigmoid activation functions, respectively. In addition, the value of r_t is used to know about the role or influence of h_{t-1} on h'_t. If r_t value is 0, it means that the information from h_{t-1} is totally ignored or discarded. Likewise, if its value is 1, it means that all the information from h_{t-1} is considered. Furthermore, in the final h_t equation, unlike LSTM, instead of employing an independent or separate gate, in GRU only one update gate is employed to control both the previous information from h_{t-1} as well as the new information from the h'_t. Now, assume that the value of v_t is 0 or near to 0 in h_t's equation, the first term of h_t equation is going to vanish. It means that h_t will not contain a good amount of information from h_{t-1} and the h_t will have information from h'_t only. Similarly, if the value of v_t is 1 or close to 1 in h_t's equation, the second part of the equation will be 0; it means h_t will totally be dependent on the first part of the equation, i.e., h_{t-1}. Hence, it is proved that the value of v_t has a great importance and it ranges from 0 to 1.

In the GRU module of TLGRU, four layers are employed: GRU layer, Flatten layer, Dropout layer, and Dense layer. In the first layer, 50 neurons are used, 20% of dropout probability is selected, and only one neuron for dense layer is employed. In the second layer, the conversion of data from multi dimensional to single dimensional is performed. Furthermore, the 20% of the dropout probability value is used to randomly deactivate the 20% of neurons, so that the overfitting issue can be avoided.

3.3.2. Support Vector Machine

SVM is one of the most popular and widely used classification techniques that is employed by many researchers as their basic proposed or benchmark model in the existing literature [43]. In [11], SVM based CPBETD is proposed and in [35], SVM is employed as an existing or benchmark classifier for ETD. In our case, we also select SVM as the benchmark classifier for performance evaluation of our proposed TLGRU classifier. The SVM can be employed for classification as well as regression. The support vector regression (SVR) and the support vector classification (SVC) are the two classes of SVM that are used for regression and classification, respectively. However, as our task is theft detection (a classification task), so we use the SVC class of the SVM for classification. For classification purpose, the SVM finds a hyperplane that maximizes the margin of the hyperplane to support vectors. This is done in order to separate the benign data and the abnormal data from each other so that the data of both the classes can be more clearly classified. Furthermore, as the EC data in SGCC dataset is not linearly-separable, therefore, we need to use the kernel SVM. In this way, the radial basis function (RBF) kernel is utilized for the classification of the non linearly-separable data (SGCC). The C and γ are the hyperparameters of SVM and their values are selected by default. The curliness of the decision boundary in SVM is decided by γ. Whereas, C is used to control the misclassification error.

3.3.3. Logistic Regression

LR is the simplest supervised ML binary classification algorithm [44]. It is an extensively utilized classifier for ETD in SGs (binary classification problem) [35,45,46]. LR follows the same principles as neural network. Therefore, we can surely say that LR used for binary classification problems (binary LR) is analogous and similar to the neural network with only one hidden layer and a sigmoid activation function (spans from 0 to 1). Where, the value close to 0 is regarded the normal consumer and vice versa. During the coding stage of

LR, we have considered the optimal values for the hyperparameters, i.e., $random\_state = 5$ and $solver = liblinear$, where *random_state* is employed to control the random number generator and *solver* specifies that which algorithm to employ for optimization. The values for other hyperparameters are chosen as default.

3.3.4. Decision Tree

DT is also a popular classification method that divides the attributes into classes on the basis of their relevant features. DT is widely used by the researchers as the benchmark [45] and base model [47]. DT prepares a road map for the state of the art and advanced ensembled techniques like gradient boosting classifiers, random forest, and bagging techniques [48].

4. Simulation Results

In this study, simulations are performed using Google Colaboratory. The details of the selected SGCC dataset for validation of our proposed model are given in Section 3.1. Section 4.1 comprises the results obtained after performing extensive simulations. Moreover, the comparison between TLGRU's performance and the performance of the existing models, i.e., GRU, SVM, DT, and LR, is made and the validation is done in terms of ETD. Furthermore, recall, area under the curve (AUC), precision, and F1-score metrics are considered the appropriate measures in order to compute the classifiers' performance using the imbalanced data [35]. Based on the cases mentioned above, the accuracy metric is not an appropriate performance measure [48,49].

4.1. Proposed TLGRU and Other Benchmark Techniques' Results

The epoch variable is employed to control the training process of the proposed model. We run our model for 10 iterations or epochs. The convergence of our proposed TLGRU model with respect to accuracy and loss performance measures is exhibited in Figures 5 and 6, respectively. It is visible in the figures that the training accuracy (accuracy on seen data) of the TLGRU increases moderately at each iteration and finally it reaches to 91.77% at the final iteration. Whereas, using the testing data (accuracy using unseen data), the TLGRU accuracy gradually increases as well and it reaches 91.56% at the final iteration. The SGCC dataset contains some zero values due to which the proposed classifier can not learn it properly at the early iterations, therefore, during the first three epochs, the training accuracy is better than testing accuracy, which means that the overfitting issue has occurred. After the 3rd iteration, the proposed model efficiently learns the zero values and the overfitting issue is removed. The loss of the TLGRU is also computed and noted at different iterations. As shown in Figure 6, the training loss is minimizing at every iteration till it reaches to 0.2068 at the 10th iteration. While, the testing loss also reduces till it reaches to 0.2084 at the final iteration. During the first three iterations, the model overfits because of the zero values that exist in the dataset. After the third iteration, the model learns the dataset as well as the zero values and therefore, the overfitting issue is solved. Now, finally, from training and testing accuracy of the proposed model, it is concluded that the model generalizes well and avoids overfitting. The overfitting problem is tackled using the proper tuning of the dropout probability value and powerful and significant features' extraction ability of the LSTM in TLGRU model.

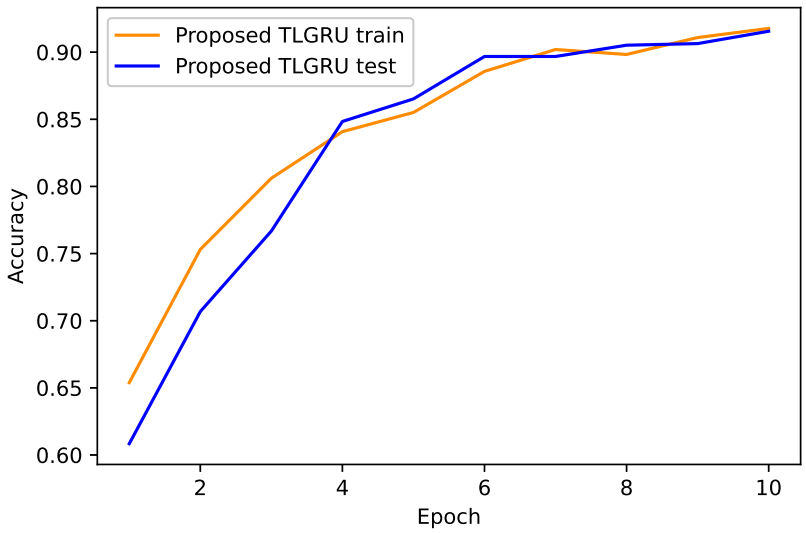

Figure 5. The proposed TLGRU model's training and testing accuracy.

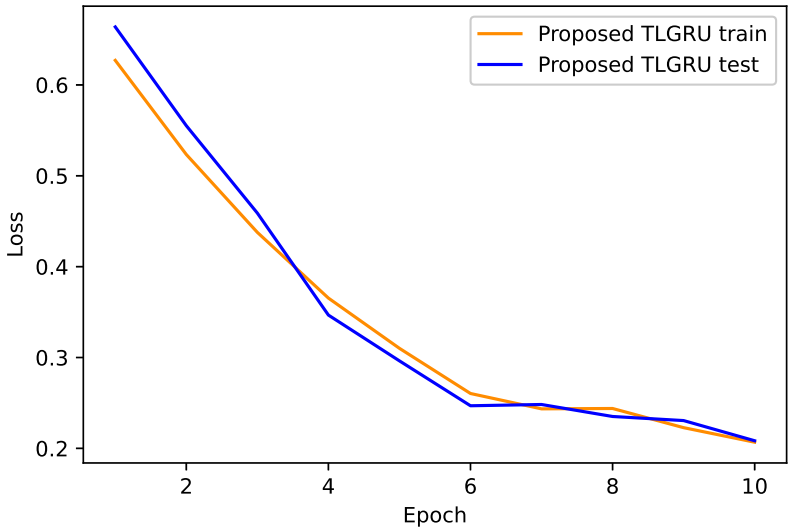

Figure 6. The proposed TLGRU model's training and testing losses.

The accuracy and loss convergence analysis results for the benchmark GRU model are shown in Figures 7 and 8, respectively. It is shown that at the final iteration, the training and testing accuracy value for GRU is 89.99% and 82.65%, respectively. Using the seen data, the accuracy slowly increases, whereas, on testing data, the accuracy fluctuates till the 7th iteration and after 7th iteration, it starts decreasing. Moreover, at the 3rd epoch, the GRU model learns and trains using a batch having some zero values that causes overfitting. However, after the 7th epoch, the overfitting again starts and continues till the final iteration. Hence, it is proved that the existing GRU model overfits. The main reasons for the occurrence of this issue are lack of important features' extraction to reduce the data dimensionality and no proper tuning of the dropout regularization probability value. Furthermore, the training and testing loss values for GRU at the last iteration are 0.2852 and 0.5136, respectively. The same trend continues in the loss as well. The training loss continuously decreases while the testing loss fluctuates till the 7th epoch and it starts increasing after the 7th iteration. At the 3rd epoch, the overfitting issue occurs due to the presence of zero values in the batch. Later on, after the 7th iteration, the overfitting

issue occurs and it continues till the end, hence, the model overfits. Finally, from the training and testing accuracy and loss plots, it is concluded that GRU model does not have good generalization ability and overfitting issue occurs because of neglecting the tuning of the dropout regularization probability value and features' extraction. Whereas, our proposed TLGRU model outperforms the benchmark GRU in terms of tackling the overfitting problem.

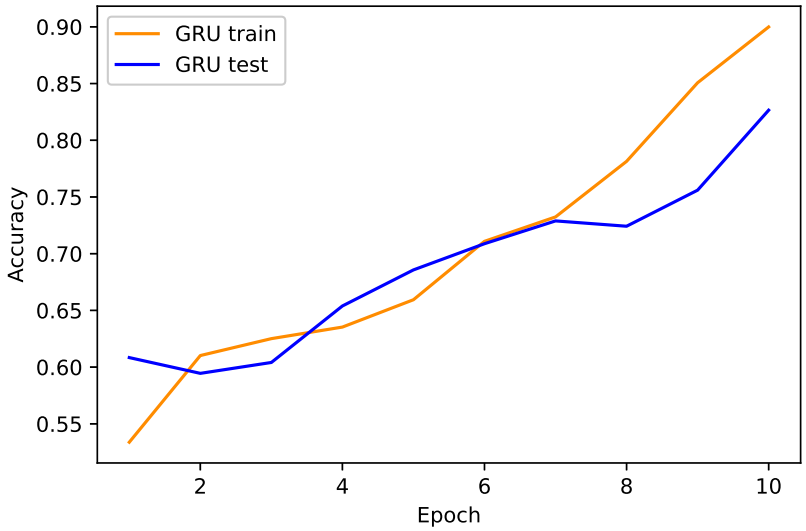

Figure 7. The GRU model's training and testing accuracy.

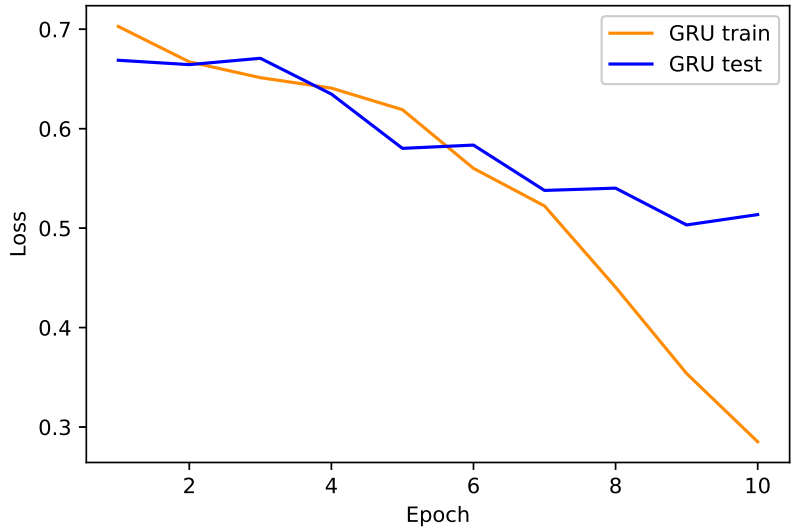

Figure 8. The GRU model's training and testing losses.

For TLGRU's comparison with DT, LR, SVM, and GRU models, the AUC, recall, precision, F1-score, and accuracy are considered, which are the most effective performance parameters. The comparison of the TLGRU's performance with the existing benchmark models is exhibited in Table 4 and Figure 9 with regard to various performance measures. From the results, the superiority of TLGRU model over all the other conventional models with regard to the already discussed indicators of performance due to the following reasons. The first reason is that TLGRU effectively tackles the imbalanced data issue using different TAs, the second reason is that the LSTM module is used to extract the necessary features and solve the high dimensionality issue, and finally, the classification using the GRU

module enhances TLGRU's performance. Moreover, in the GRU module, dealing with the drift, overfitting, and local optima trapping issues using the update gate (long term memory), dropout regularization, and Adam optimizer further enhances the performance of the model being put forward our proposed model with respect to the selected evaluators. Contrarily, SVM and LR perform very bad. The reason is that they can not tackle the large and huge time series data and that is why overfitting issue occurs. Whereas, the proposed TLGRU classifier smoothly handles the large time series data due to the feature extraction ability of the proposed model using LSTM algorithm and solves the issue of overfitting.

Table 4. Performance comparison of the proposed and existing schemes.

Classifier	Accuracy	AUC Score	Precision	Recall	F1-Score	FPR
DT	0.6701	0.6702	0.7019	0.6585	0.6795	0.1485
LR	0.5379	0.5365	0.5207	0.7097	0.6736	0.4370
SVM	0.6433	0.6423	0.4678	0.7162	0.5660	0.2646
GRU	0.8265	0.7552	0.8355	0.7176	0.7721	0.0818
Proposed TLGRU	0.9156	0.9168	0.9796	0.8659	0.9192	0.0100

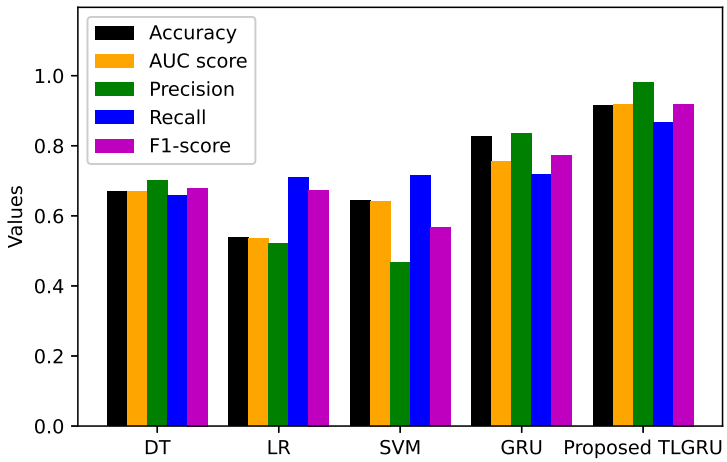

Figure 9. Performance comparison of different classifiers.

FPR is one of the significant performance metrics for ML techniques in which the benign electricity users are classified and shown as theft. The value of FPR is directly proportional to the on field inspection cost. There are many solutions or ways to minimize the FPR value. However, we have only considered and worked on three different ways to reduce the FPR value: data balancing, extraction of the important features from the raw data, and correctly identifying the drift. This article uses the GRU for classification purpose because it has a special quality of keeping the long historical sequence of data using its update gate (long term memory), and the data are then used for analyzing the long relationships among the consumption patterns. The GRU model identifies the anomalies or irregularities in consumption data that arise due to the non-theft reasons. The FPR for DT is 0.1485, for SVM is 0.2646, for LR is 0.4370, for GRU is 0.0818, and for TLGRU is 0.01. Hence, consequences are clear from the FPR numeric values that the TLGRU classifier exhibits the least FPR in comparison with all the other benchmark classifiers. The FPR for our proposed TLGRU and other benchmark models is shown in Figure 10.

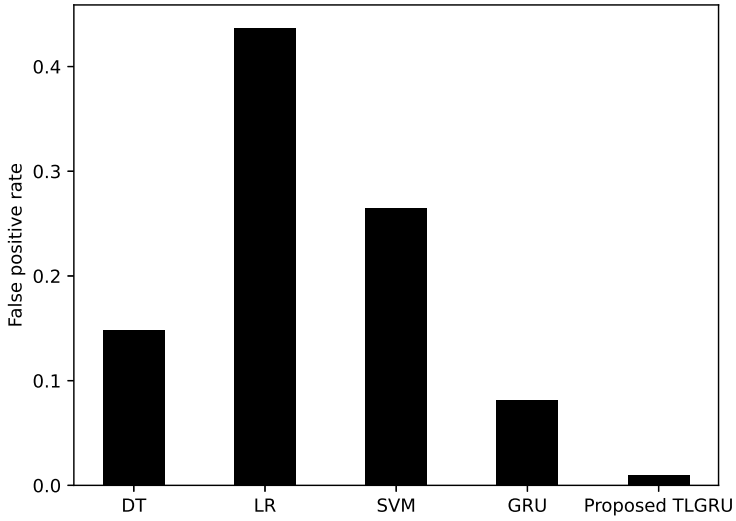

Figure 10. The proposed and benchmark models' FPR analysis.

4.2. Strengths and Weaknesses of the Proposed Work

The fundamental advantage of this work is to provide an efficient ETD model for power utilities, which helps them to reduce economic loss. Furthermore, the accurate and timely detection of energy thieves reduces the line losses in transformers and other grids' components. Besides, the proposed model has some shortcomings. The low-frequency EC data are used to train the model, which limits its performance towards capturing the most granular EC patterns. Consequently, the rate of misclassifying instances increases. Further, it may incur high computational time due to the absence of a hyperparameter tuning technique.

5. Conclusions and Future Work

In this research article, the TLGRU model is presented that consists of two main components and three sub-components. The main components are preprocessing unit and theft classification unit. The preprocessing component is further divided into three sub-components: initial preprocessing of the raw data, data balancing, and feature extraction. In the first sub-component, the NaN values, outliers, and unscaled data are dealt by employing SI, TSR of thumb, and min-max scaler, respectively. In the second sub-component, TAs are employed for creation of the synthetic abnormal data samples to solve the imbalanced data problem. Finally, LSTM classifier is employed for feature extraction and dealing with dimensionality curse problem. Furthermore, the classification component contains the GRU model for theft classification. Moreover, the GRU provides solution for drift identification, overfitting, and trapping in local optima problems. In addition, four popular benchmark models, DT, LR, SVM, and GRU, are implemented for performance comparison with the proposed TLGRU classifier. A realistic EC dataset (SGCC) is employed for simulations. From the simulation results, the superiority of TLGRU over benchmark models in terms of ETD is exhibited. The simulations provide us with 1% FPR, 97.96% precision, 91.92% F1-score, 91.68% AUC, 91.56% accuracy, and 86.59% recall, which are better than the benchmark schemes. Hence, we conclude that the newly developed TLGRU is an efficient model for ETD with minimum FPR. In future works, we will investigate the novel DL models for feature extraction and classification to more efficiently perform the ETD task. Moreover, automated tuning of the hyperparameters' of the models will also be performed using meta-heuristic optimization algorithms. In addition, the aim of the underlying study is the development of a novel deep learning based hybrid model, which helps electric utilities to detect energy frauds in SGs around the globe. Furthermore, the proposed model

is trained on a massive EC dataset. So, its real time practicability in terms of identifying the presence of thieves in the SGs is ensured. Further, the model introduced in the underlying work is a quite general solution to detect anomalies and frauds in any time series. It needs only a dataset for its training. So, the EC data collected by the conventional meters can be used to train the proposed model and then it will be applicable in SGs to detect energy frauds.

Author Contributions: Conceptualization, P., N.J. and S.A.; methodology, P.; software, P. and S.J.; validation, N.J., M.A. and M.U.J.; formal analysis, A.S.Y. and S.J.; investigation, N.J.; resources, S.J.; data curation, P.; writing—original draft preparation, P. and N.J.; writing—review and editing, N.J. and S.A.; visualization, N.J. and S.A.; supervision, N.J.; project administration, S.A. and N.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset employed in this research is available online at https://github.com/henryRDlab/ElectricityTheftDetection (accessed on 30 December 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Garcia Deluno, F.; Marafão, F.P.; de Souza, W.A.; da Silva, L.C.P. Power metering: History and future trends. In Proceedings of the 2017 Ninth Annual IEEE Green Technologies Conference (GreenTech), Denver, CO, USA, 29–31 March 2017; pp. 26–33.
2. Weranga, K.S.K.; Kumarawadu, K.; Chandima, D.P. *Smart Metering Design and Applications*; Springer: Singapore, 2014.
3. Foudeh Husam A.; Mokhtar, A.S. Automated meter reading and advanced metering infrastructure projects. In Proceedings of the 2015 9th Jordanian International Electrical and Electronics Engineering Conference, Amman, Jordan, 12–14 November 2015; pp. 1–6.
4. Huang, Y.; Qifeng, X. Electricity theft detection based on stacked sparse denoising autoencoder. *Int. J. Elect. Power Energy Syst.* **2021**, *125*, 106448.
5. Zheng, Z.; Yatao, Y.; Xiangdong, N.; Dai, H.; Zhou, Y. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. *IEEE Transact. Ind. Inform.* **2017**, *14*, 1606–1615.
6. Buzau, M.-M.; Tejedor-Aguilera, J.; Cruz-Romero, P.; Gómez-Expósito, A. Hybrid deep neural networks for detection of non-technical losses in electricity smart meters. *IEEE Trans. Power Syst.* **2019**, *35*, 1254–1263.
7. Khoo, B.; Ye, C. Using RFID for anti-theft in a Chinese electrical supply company: A cost-benefit analysis. In Proceedings of the 2011 Wireless Telecommunications Symposium (WTS), New York City, NY, USA, 13–15 April 2011; pp. 1–6.
8. McLaughlin, S.; Brett, H.; Fawaz, A.; Berthier, R.; Zonouz, S. A multi-sensor energy theft detection framework for advanced metering infrastructures. *IEEE J. Select. Areas Commun.* **2013**, *31*, 1319–1330.
9. Cárdenas; A.A.; Saurabh, A.; Schwartz, G.; Dong, R.; Sastry, S. A game theory model for electricity theft detection and privacy-aware control in AMI systems. In Proceedings of the 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 1–5 October 2012; pp. 1830–1837.
10. Amin, S.; Schwartz, G.A.; Tembine, H. Incentives and security in electricity distribution networks. In Proceedings of the International Conference on Decision and Game Theory for Security, Virtual Conference, 25–27 October 2012; pp. 264–280.
11. Jokar, P.; Arianpoo, N.; Leung, V.C.M. Electricity theft detection in AMI using customers' consumption patterns. *IEEE Trans. Smart Grid* **2015**, *7*, 216–226.
12. Gunturi Kumar, S.; Sarkar, D. Ensemble machine learning models for the detection of energy theft. *Electric Power Syst. Res.* **2021**, *192*, 106904.
13. Kong, X.; Zhao, X.; Liu, C.; Li, Q.; Dong, D.; Ye, L. Electricity theft detection in low-voltage stations based on similarity measure and DT-KSVM. *Int. J. Electr. Power Energy Syst.* **2021**, *125*, 106544.
14. Buzau, M.M.; Tejedor-Aguilera, J.; Cruz-Romero, P.; Gómez-Expósito, A. Detection of non-technical losses using smart meter data and supervised learning. *IEEE Trans. Smart Grid* **2018**, *10*, 2661–2670.
15. Qu, Z.; Li, H.; Wang, Y.; Zhang, J.; Abu-Siada, A.; Yao, Y. Detection of electricity theft behavior based on improved synthetic minority oversampling technique and random forest classifier. *Energies* **2020**, *13*, 2039.
16. Lu, X.; Zhou, Y.; Wang, Z.; Yi, Y.; Feng, L.; Wang, F. Knowledge embedded semi-supervised deep learning for detecting non-technical losses in the smart grid. *Energies* **2019**, *12*, 3452.
17. Ashraf Ullah, P.; Shoaib, M.; Muhammad, A.; Kabir, B.; Javaid, N. Synthetic Theft Attacks Implementation for Data Balancing and a Gated Recurrent Unit Based Electricity Theft Detection in Smart Grids. In Proceedings of the Conference on Complex, Intelligent, and Software Intensive Systems, Asan, Korea, 1–3 July 2021; Springer: Cham, Switzerland, 2021; pp. 395–405.

18. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
19. Cho, K.B.; Merrienboer, V.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259.
20. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inform. Proces. Syst.* **2014**, *27*.
21. Fabian, A.N.; Figueroa, G.; Chu, C. NTL detection in electric distribution systems using the maximal overlap discrete wavelet-packet transform and random undersampling boosting. *IEEE Trans. Power Syst.* **2018**, *33*, 7171–7180.
22. Hasan, M.; Toma, R.N.; Abdullah-Al, N.; Islam, M.M.; Kim, J. Electricity theft detection in smart grid systems: A CNN-LSTM based approach. *Energies* **2019**, *12*, 3310.
23. Saeed Salman, M.; Mustafa, M.W.; Sheikh, U.U.; Jumani, T.A.; Mirjat, N.H. Ensemble bagged tree based classification for reducing non-technical losses in multan electric power company of Pakistan. *Electronics* **2019**, *8*, 860.
24. Wang, X.; Yang, I.; Ahn, Su. Sample efficient home power anomaly detection in real time using semi-supervised learning. *IEEE Access* **2019**, *7*, 139712–139725.
25. Liu, H.; Li, Z.; Li, Y. Noise reduction power stealing detection model based on self-balanced data set. *Energies* **2020**, *13*, 1763.
26. Ibrahem, M.I.; Nabil, M.; Fouda, M.M.; Mahmoud, M.M.E.A.; Alasmary, W.; Alsolami, F. Efficient Privacy-Preserving Electricity Theft Detection with Dynamic Billing and Load Monitoring for AMI Networks. *IEEE Internet Thing. J.* **2020**, *8*, 1243–1258.
27. Nabil, M.; Ismail, M.; Mahmoud, M.M.E.A.; Alasmary, W.; Serpedin, E. PPETD: Privacy-preserving electricity theft detection scheme with load monitoring and billing for AMI networks. *IEEE Access* **2019**, *7*, 96334–96348.
28. Micheli, G.; Soda, E.; Vespucci, M.T.; Gobbi, M.; Bertani, A. Big data analytics: An aid to detection of non-technical losses in power utilities. *Comput. Manag. Sci.* **2019**, *16*, 329–343.
29. Punmiya, R.; Choe, S. Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing. *IEEE Trans. Smart Grid* **2019**, *10*, 2326–2329.
30. Razavi, R.; Gharipour, A.; Fleury, M.; Akpan, I.J. A practical feature-engineering framework for electricity theft detection in smart grids. *Appl. Energy* **2019**, *238*, 481–494.
31. Ramos, C.; Rodrigues, D.; de Souza, A.N.; Papa, J.P. On the study of commercial losses in Brazil: A binary black hole algorithm for theft characterization. *IEEE Trans. Smart Grid* **2016**, *9*, 676–683.
32. Ghasemi, A.A.; Gitizadeh, M. Detection of illegal consumers using pattern classification approach combined with Levenberg-Marquardt method in smart grid. *Int. J. Electr. Power Energy Syst.* **2018**, *99*, 363–375.
33. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*; MIT Press: Cambridge, MA, USA, 1995; Volume 3361.
34. Li, S.; Han, Y.; Yao, X.; Yingchen, S.; Wang, J.; Zhao, Q. Electricity theft detection in power grids with deep learning and random forests. *J. Electr. Comput. Eng.* **2019**, *2019*, 4136874.
35. Adil, M.; Javaid, N.; Qasim, U.; Ullah, I.; Shafiq, M.; Choi, J. LSTM and bat-based RUSBoost approach for electricity theft detection. *Appl. Sci.* **2020**, *10*, 4378.
36. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536.
37. Javaid, N. A PLSTM, AlexNet and ESNN Based Ensemble Learning Model for Detecting Electricity Theft in Smart Grids. *IEEE Access* **2021**, *9*, 162935–162950.
38. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
39. Zhang, Y.G.; Tang, J.; He, Z.; Tan, J.; Li, C. A novel displacement prediction method using gated recurrent unit model with time series analysis in the Erdaohe landslide. *Nat. Hazards* **2021**, *105*, 783–813.
40. Aniruddha, D.; Kumar, S.; Basu, M. A gated recurrent unit approach to bitcoin price prediction. *J. Risk Financ. Manag.* **2020**, *13*, 23.
41. Niu, Z.; Yu, Z.; Tang, W.; Wu, Q.; Reformat, M. Wind power forecasting using attention-based gated recurrent unit network. *Energy* **2020**, *196*, 117081.
42. Luo, H.; Wang, M.; Wong, P.K.; Tang, J.; Cheng, J.C.P. Construction machine pose prediction considering historical motions and activity attributes using gated recurrent unit (GRU). *Automat. Construct.* **2021**, *121*, 103444.
43. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
44. Available online: www.tutorialspoint.com (accessed on 5 June 2021).
45. Gul, H.; Javaid, N.; Ullah, I.; Qamar, A.M.; Afzal, M.K.; Joshi, G.P. Detection of non-technical losses using SOSTLink and bidirectional gated recurrent unit to secure smart meters. *Appl. Sci.* **2020**, *10*, 3151.
46. Aslam, Z.; Javaid, N.; Ahmad, A.; Ahmed, A.; Sardar Muhammad Gulfam. A Combined Deep Learning and Ensemble Learning Methodology to Avoid Electricity Theft in Smart Grids. *Energies* **2020**, *13*, 5599.
47. Jindal, A.; Dua, A.; Kaur, K.; Singh, M.; Kumar, N.; Mishra, S. Decision tree and SVM-based data analytics for theft detection in smart grid. *IEEE Trans. Ind. Inform.* **2016**, *12*, 5–1016.
48. Available online: www.machinelearningmastery.com (accessed on 17 April 2021).
49. Javaid, N.; Jan, N.; Javed, M.U. An adaptive synthesis to handle imbalanced big data with deep siamese network for electricity theft detection in smart grids. *J. Parallel Distrib. Comput.* **2021**, *153*, 44–52.