Shoaib Munawar[1], Muhammad Sajid Iqbal[1], Muhammad Adnan[1], Muhammad Ali Akbar[2], and Amine Bermak[2]

[1]National University of Computer and Emerging Sciences, Chiniot-Faisalabad Campus
[2]Computer Science and Engineering department HBKU

November 19, 2024

Posted on 19 Nov 2024 — CC-BY 4.0 — https://doi.org/10.36227/techrxiv.173203066.63774440/v1 — e-Prints posted on TechRxiv are preliminary reports that are not peer reviewed. They should not b...

A Critical Review of Technical Case Studies for Electricity Theft Detection in Smart Grids: A New Paradigm Based Transformative Approach

Shoaib Munawar[a,], Muhammad Sajid Iqbal[a], Muhammad Adnan[a], Muhammad Ali Akbar[b], Amine Bermak[b]

[a]*National University of Computer and Emerging Sciences, Chiniot-Faisalabad Campus, Pakistan*
[b]*Computer Science and Engineering department HBKU, Qatar*

Abstract

Electricity theft detection (ETD) is a vital global concern that affects utility providers (UPs), and non-technical losses (NTLs) are a major issue. While smart metering has helped to reduce conventional technical losses (TLs), NTLs caused by theft are still difficult to identify and have serious effects. Consumers frequently underreport their power consumption, which complicates detection attempts. This paper provides a comprehensive review of various ETD approaches, categorizing them into five areas: (i) NTL detection with synthetic data, (ii) sequential data-based schemes, (iii) non-sequential data analysis, (iv) neighborhood area-based networks (NANs), and (v) IoT and hardware-based solutions. Each area is presented using case examples that have been statistically, mathematically, and visually analyzed. The article also includes a summary table of known problems and possible solutions, making it a useful resource for academics and developers. A comparison analysis utilizing F1 scores is presented to assess the efficacy of different detection strategies. This is the first review of its type to convert technical articles into real-world case studies, offering useful insights for selecting the best theft detection technologies in smart grids.

Keywords: Electricity Theft Detection, Non-Technical Losses, SMOTE, BiGRU-BiLSTM, Feature Engineering, Feature Extraction, F1 Score.

1. Introduction

Nowadays electricity theft in smart meters' (SM) data is a serious issue in electrical distribution systems. SM is a device on consumer premises, which monitors their consumed energy. The SM data are observed by utility providers (UPs), who charge consumers for their consumed energy. Consumers tend to scratch their SM data to under-report the recorded data of SM. The under-reporting benefits the consumers in a sort of financial relief. To adopt such approaches, consumers use some mechanisms such as (i) use of shunt devices for data tampering [1]-[3] (ii) SMs double tapping and (iii) forced electronic faults [4]-[6]. SMs double tapping and forced electronic faults are the traditional approaches and are applied through handcrafting mechanisms. The handcrafting mechanisms are the physical linking and can be investigated easily through physical inspection. Tampering SM's data through sharp data manipulating approaches is a novel and vigilant approach. Such techniques are hard to identify and need a proper expert system to investigate. Both traditional SM data tampering and data manipulating approaches result in non-technical losses (NTLs) on consumer premises. Losses are categorized into two categories i-e technical losses (TLs) and NTLs [7]-[9]. NTLs occur due to consumers' intervention, however, TLs are the inherent system losses. The inherent system losses are the systems' internal losses due to known technical factors. Such factors are probably internal losses and environmental factors. This study focuses on the occurrence and investigation of NTLs. A power system is an integration of three main phases (i) generation phase (ii) transmission phase and (ii) distribution phase. NTLs occur on the distribution side where consumers interact with the electrical network [10]-[11].

The Ups install SMs to monitor the consumed energy and behavior of the consumer. The consumed energy is presented in the form of digital information. The UPs monitor the digital information as a pattern and behavior of the consumer. The behavior is based on the morphological assessment. Based on the morphological assessment, the consumers are divided into two categories (i) honest consumers and (ii) fraudulent consumers. The honest consumer is also named benign consumer whereas the fraudulent consumer is named anomalous or

Email address: iqbal.sajid@nu.edu.pk (Muhammad Sajid Iqbal)

theft consumer. A consumer with a fair data pattern of the consumed energy is a benign consumer whereas a consumer with a tampered consumption pattern is considered a theft, anomalous, or fraudulent consumer. The morphological assessment and investigation of the consumer's pattern investigate features of the consumed energy on a daily, weekly, monthly, and quarterly basis. Such analysis renders the data in a specific spectrum over the recorded time. Each consumer's spectrum is considered as a base or historic pattern for that specified consumer. Various comparative analyses are carried out and compared within the historic pattern to identify the anomaly in patterns of a consumer. The historic and consumed energy patterns over a specified time are compared and the differences are monitored. If there is any difference in historical and suspected data, the difference is calculated. The difference can be due to various factors such as geographical, topographical parameters, and demographic structure. Geographical, topographical, and demographic are the surface, environment, and family structure-based parameters, respectively. Such factors cause minor differences in patterns, which are considered thresholds. If the difference is much enough over the threshold, the consumer is considered a suspected one and referred to as an anomalous one. Data obtained and monitored from the aforementioned parameters is non-sequential data.

Data are of two types (i) sequential data and (ii) non-sequential data. Sequential data is temporal sequence-based data, which contains numerical data monitored by SM against the consumed energy. Non-sequential is predefined factors-based data. It may or may not be in numerical format. Sequential data of SM are number-based data and it is easy to manipulate its originality. To tamper with the SM data consumers use various data manipulating techniques such as theft cases and false data injection techniques (FDIs). There are six theft cases and six FDIs mentioned in the literature. FDIs are novel data manipulating techniques, which are introduced in contrast to six theft cases. To manipulate SMs' data, the benign consumer's data are considered and theft cases or FDIs are applied. Applying such approaches changes the data nature and under-reports the SM readings, which results in huge revenue losses to Ups. Literature in [12] reports losses of two decades during 1980-200. The revenue losses are increased from 11% to 16% due to NTLs. Similarly, 20% of the total supply is reported in India [13]. Revenue losses of 10% and 16% are reported in Russia and Brazil, respectively. Moreover, a 100 million dollar loss is reported in Canada due to NTLs [14]-[15]. Additionally, a loss of 96 billion dollars is reported worldwide [16]. Losses due to NTLs are huge enough to underestimate. Such losses cause huge revenue losses and suffer UPs financially. To tackle such an issue, is an important aspect of the research and many solutions have been provided in literature. The traditional causes for NTLs have been eliminated, however, the major issue of data manipulation is still a serious one [17]-[19]. The data manipulating techniques are vigilant and sharp, which are difficult to detect with 100% efficiency. The data manipulating techniques reported in the literature are as follows:

- Theft cases and
- FDIs.

1.1. Theft Cases Vs FDIs

Theft case-1: In theft case-1 the benign data of SM is multiplied with a random number. The random number ranges between (0.1-0.9). Multiplying SM data with a random number manipulates the originality of the data and under-reports it. Equation 1 shows the mathematical representation of Theft case 1.

$$T1(E_t) = E_t * random(0.1, 0.9) \tag{1}$$

Theft case-2: In theft case-2, discontinuity in the pattern is observed by multiplying the whole time series data with a series of random numbers. Unlike theft case-1, the series of random numbers is multiplied instead of a single random number. Such approaches depict a discontinuity and randomness in the data manipulation. Equation 2 shows the mathematical representation of Theft case-2.

$$T2(E_t) = E_t * E_t(E_t = random(0.1, 0.9)) \tag{2}$$

Theft case-3: The time series data of SM is multiplied by 0 or 1. It is a two-stage pattern. Multiplication with 0 voids the whole day's consumption and reports it as a zero. However, multiplication with 1 depicts the original consumption and no theft is reported. The stages are kept in such a pattern so that no proper behavior of the theft data is observed by the investigation team. It is a vigilant and worse type of theft case as no such evidence is left for suspicious activities. Equation 3 shows the mathematical representation of Theft case-3.

$$T3(E_t) = E_t * random[0, 1] \tag{3}$$

Theft case-4: In theft case-4, a total mean of the SM data is taken, and the mean is then multiplied by a random number. Here in this scenario, the whole day consumption is considered as a mean value. As the recorded consumption in the SGCC dataset is only 24 hours for a single day that's why we consider it as a single mean value. However, it can be weekly and monthly based. The random number ranges between (0.1-1). It is a severe case as well where two modes of data manipulation are applied. Initially, the mean of the whole consumed energy is presented, and later on, it is multiplied with a random number. Equation 4 shows the mathematical representation of Theft case-4.

$$T4(E_t) = mean(E_t) * random(0.1, 1.0) \tag{4}$$

Theft case-5: In theft case-5, the mean of the consumed data is presented, which is the average data of a specific period. The mean can be daily, weekly, and monthly based. This sort of manipulation is mild and it is hard to detect as no prominent manipulating factors are obvious to investigate. Equation 5 shows the mathematical representation of Theft case-5.

$$T5(E_t) = mean(E_t) \tag{5}$$

Theft case-6: In theft case-6, the time series data is swiped with one another. The OFF-peak hours and ON-peak hours are swiped. The scenario reports OFF-peak hours as ON-peak hours. Such manipulation reports the same data usage, however, the tariff schemes are different for on-peak and off-peak hours. Consumers tend to opt for minimal financial benefits through such schemes. Equation 6 shows the mathematical representation of Theft case-6.

$$T6(E_t) = E_{T-t} \tag{6}$$

Where T is the total consumption time and t is swipe time to exchange the time period of the consumed energy.
FDI-1: In FDI-1 the mean of the time series data is taken and multiplied with a random number, which ranges between (0.1-0.9). The product is then divided by total consumption, however, the consumption taken is less than the mean value. Equation 7 shows the mathematical representation of FDI-1.

$$FDI_1 = \frac{mean(E) * random(0.1 - 0.9)}{E} \tag{7}$$

Where $E > 1 \leq$ mean.
FDI-2: In FDI-2, the mean of the total consumed energy is multiplied by a random number likewise FDI1. Later on, the square root of the whole product is taken. The resultant data is shown as the total consumption for a specific billing month. Equation 8 shows the mathematical representation of FDI-2.

$$FDI_2 = \sqrt{(}mean(E)) * random(0.1 - 0.9) \tag{8}$$

FDI-3: In FDI-3 traits of theft cases are adopted similarly, however, an additional square-rooted statistical feature is opted for. The total consumption is multiplied by a random number ranging between (0.1-0.9). The product is then square rooted and the result is shown as final consumption. This approach is severe and hard to detect. Equation 9 shows the mathematical representation of FDI3.

$$FDI_3 = \sqrt{(}E) * random(0.1 - 0.9) \tag{9}$$

FDI-4: In FDI-4, the mean of the total consumption is taken and a random number is subtracted from the mean. The random subtraction can be hourly, daily, weekly, and monthly based. It is a sharp way to manipulate SM data as it leaves no specific pattern or behavior of consumer's identity to investigate. Equation 10 shows the mathematical representation of FDI-4.

$$FDI_4 = mean(E) - (\gamma) \tag{10}$$

where γ is a constant consumption and $\gamma \leq mean$. FDI-5: In FDI-5, a constant value number is periodically subtracted from the total consumption. It can be hourly, daily, weekly, and monthly base. The constant number is not a fixed value and varies over time. Such variations hide the originality of the consumed energy. Equation 11 shows the mathematical representation of FDI-5.

$$FDI_5 = E - \gamma_i \tag{11}$$

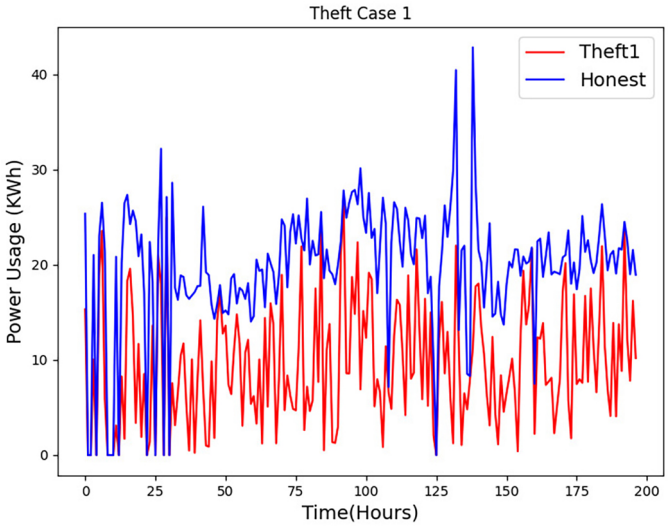

Figure 1: Theft Case and False Data Injection for Case 1

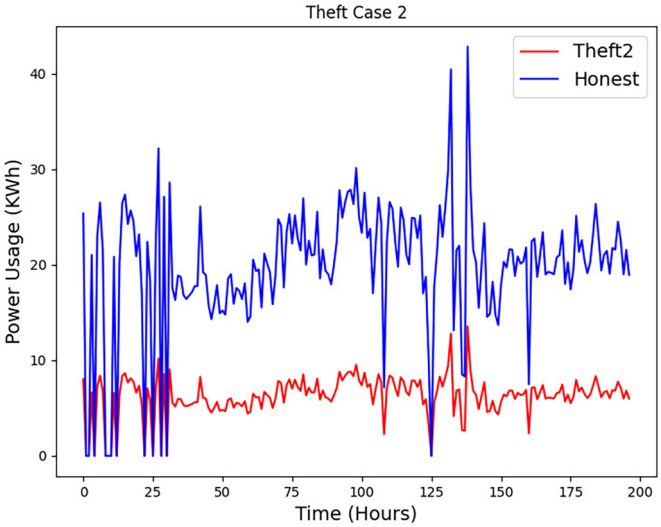

Figure 2: Theft Case and False Data Injection for Case 2

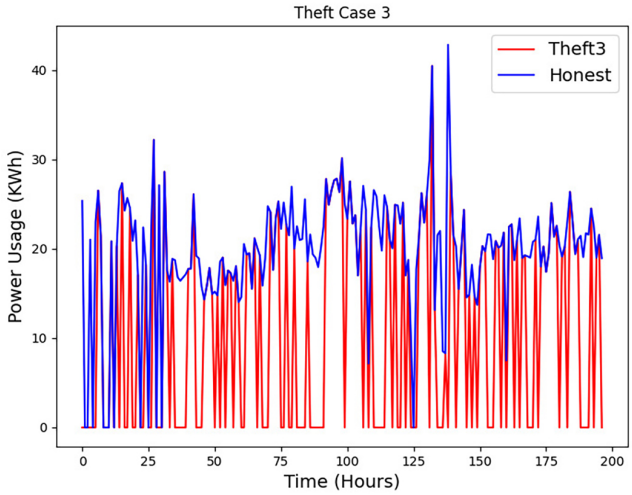

Figure 3: Theft Case and False Data Injection for Case 3

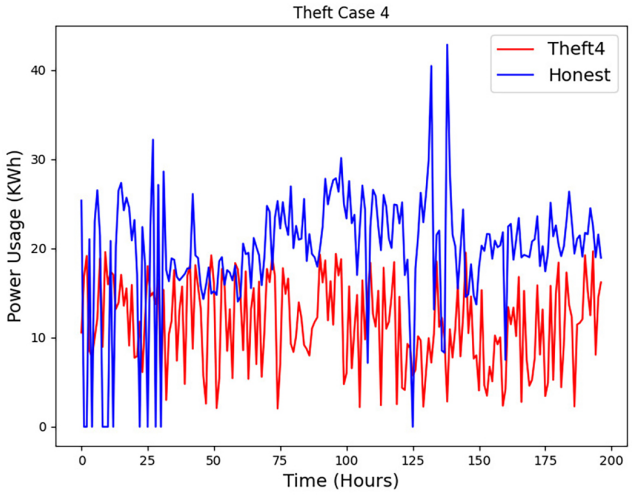

Figure 4: Theft Case and False Data Injection for Case 4

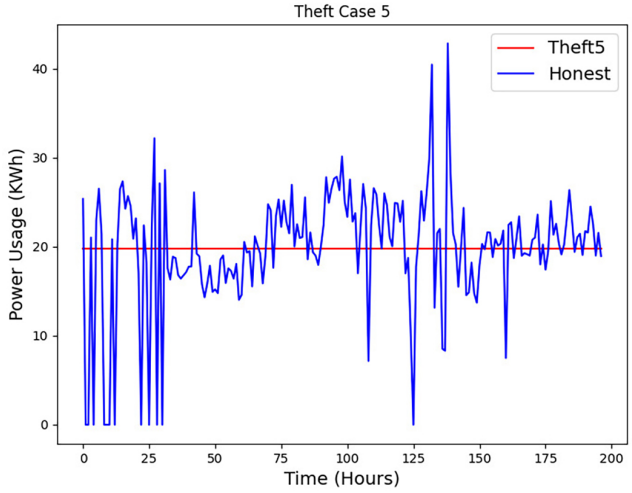

Figure 5: Theft Case and False Data Injection for Case 5

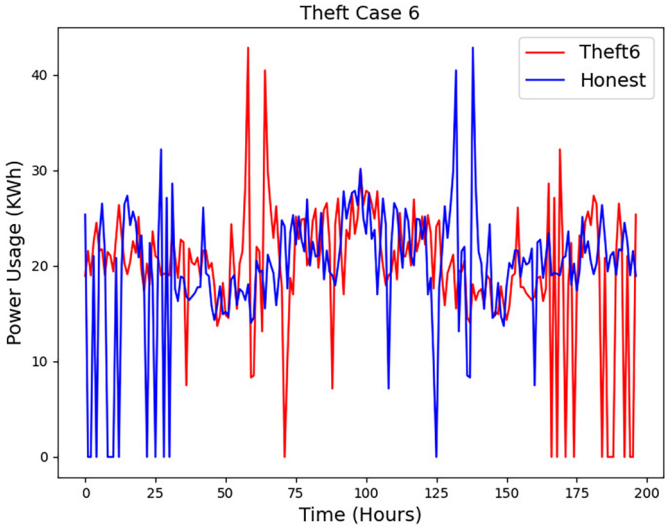

Figure 6: Theft Case and False Data Injection for Case 6

Where i= 0,...,E_{max}.
FDI-6: In FDI-6, The SM data is manipulated in such a way that the consumer is left with an open choice to target any of the time stamps. It may be time stamp-based or periodic. The time-based subtraction is observed as a sort of time swapping. It can report 0 consumption, which highlights that no load is connected or operated during that specific time period. Whereas it can show full load operation, however, time swapping and extra subtraction feature benefits the consumer during ON-peak and OFF-peak hours. Equation 12 shows the mathematical representation of FDI-6.

$$FDI_6 = E(t - d) = 0 \tag{12}$$

If $t < d$, FDI-6 is 0. In case, $t \geq d$ it is then 1. However, t and d is time and difference, respectively.

Table 1: Research Questions of the Review Article

Sr.No	**Research Questions (RQs)**
RQ1	What are the main objectives of classifiers to highlight for gaining reliable results?
RQ2	How to tackle data misclassification issue and what are the major approaches used in literature?
RQ3	What are the major procedures and techniques to extract prominent and abstract features of the data?
RQ4	How to minimize FPR and data misclassification issues?
RQ5	How to monitor the performance of various classifiers, which results in satisfactory outcomes?
RQ6	What are the major research issues and how to tackle them?
RQ7	How to avoid human interface and make the detectors reliable and automatic?
RQ8	What are the major causes of NTLs?
RQ9	What are the evaluating metrics used to monitor the performance of various algorithms?
RQ10	How to approach compact and detailed information used in literature for detection of NTLs?

Table 2: Research Questions and their Validations

Sr.No	**Validations (V)**
V1	Evaluation matrices used in graphs for each of the case study to show their statistical analysis validates RQ1.
V2	Each case study mentions the data balancing techniques and approaches used for data balancing validates the issue of data balancing?
V3	Most of the case study highlight the feature extraction techniques and are written properly in case study literature, which validates RQ3.
V4 and V5	Statistical analysis of each case study validates misclassification in term of FPR and performance parameters (AUC, F1-score).
V6	Research issues and their identified solutions are validated in Table 3 (Continue)
V7-V9	Analysis in Table 4 validates RQ7-RQ9 How to avoid human interface and make the detectors reliable and automatic?
V10	Analysis in Table 3, 3 (Continue) and 4 validates the access to many research articles in a single review.

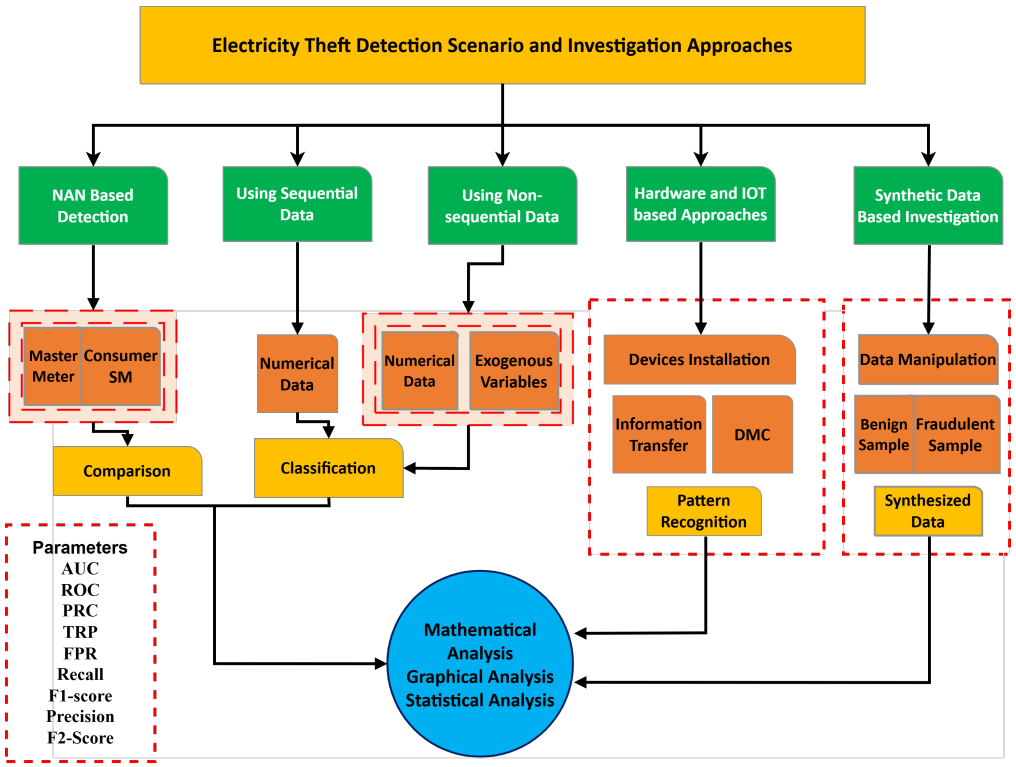

Figure 7: Architecture of Various Detection Scenarios and Approaches Used in Literature

It can be seen in Figs 1-6 that the SGCC dataset's data are manipulated using theft cases and FDIs. The original SGCC dataset is shown in black color. The theft cases and FDIs are applied over the temporal sequence. As a result, the data are manipulated and under-reported. Data representation in blue is the manipulated data using theft cases, however, the red color data are the manipulated data using FDIs. Both six theft cases and six FDIs are used and comparative sequences are analyzed. FDIs are fatal data manipulating techniques as compared to theft cases. It is observed that FDIs manipulate data quite severely, which burdensome the UPs. Henceforth, it is important to consider such data manipulating techniques in the NTL scenario. Literature in [20] uses a machine learning approach to combine many DTs, to form a string of DTs known as an ensemble decision tree to investigate NTLs. Similarly, literature in [21] uses maximal overlapped discrete wavelet-packet transform to tackle the higher data dimensionality issue and to extract the prominent features of the data. Furthermore, [22] utilizes the applications of SMOTE and binary classifiers to investigate NTLs. Moreover, [23] develops maximum information coefficient and clustering technique by fast search and find of density peaks based technique to investigate NTLs.

2. Detection Architecture of Various Approaches

This review study develops relationships between different approaches used for the detection of NTLs. The approaches are divided based on the architecture and mode of detection. Machine learning-based approaches consist of: approaches using sequential data, non-sequential data, and synthetic data whereas AMI-based NAN and hardware-based detection approaches are categorized into architecture-based topologies. Fig 7 shows the complete picture of the relationship between different detection approaches.

The approaches are discussed as follows. Synthesizing manipulated data subjects to manipulate the benign class data of the consumer.Thee benign class data is easily available and can be manipulated to synthesize the fraudulent class data. Literature represents many data manipulating approaches due to the unavailability of fraudulent class data. Datasets like SGCC represent few fraudulent consumers' data, however, novel approaches are considered by the researchers to identify the synthetic topologies for synthesizing realistic theft class data. Many researchers represent different types of data manipulating techniques and are named as theft classes, false

data injection (FDI) techniques, cyber-attacks, attack vectors, etc. Two of the scenarios: theft classes and FDI techniques are investigated in this study, however, the rest of the approaches are represented in mathematical forms. Theft classes and FDI techniques are also discussed in our published articles as well. Such techniques consider the sequential data of benign consumers and are manipulated through mathematical approaches in order to under-report the consumed energy. The under-reported consumed energy is represented by the consumers as their final consumption. Such consumption patterns can easily gain financial benefits if they are not properly investigated and reported. Synthetic techniques highlight the importance of such manipulating techniques in order to identify them and detect them if a consumer adopts them. In consideration of sequential data, the SMs' data are collected. It contains the readings of the consumed data. The readings are represented in numerical numbers and no excessive exogenous variables are added with such data. The exogenous variables are the demographic, geographical, and topographical factors, which may vary with the socio-demographic and weather factors. Such factors are named non-sequential factors and are rarely considered in the investigation of ETD scenarios. SGCC dataset contains sequential data and is mostly used by the researcher to detect anomalies in the patterns of the consumers.
In consideration of non-sequential information, the exogenous factors are considered with the sequential information. Few researchers claim that the consideration of non-sequential information leads to good classification results, however, such data are usually avoided due to their less involvement in the anomalies. The factors are the indirect factors and affect consumer behavior, not the consumers' energy consumption. Irish dataset and few more available datasets consider such parameters, however, literature pays less concern to such variables.

In the literature, it is listed that the ETD scenario is investigated through different approaches. AMI-based NAN topologies are developed to interconnect the observer meter and SMs installed on the consumer premises. Such an approach is used to identify the anomaly through a comparative analysis of the readings. Observer meter is instead on the utilization side of the transformer and SMs are interconnected to it, which forms a network. The observer meter works as a master meter and the SMs are considered as sub-meters. The master meter collects the consumption data of all the interconnected SMs in a network. The collective reading of the master meter should be equal to the sum of the individual SMs. If there is any anomaly or difference in the relationship the network is named as anomalous and the SM of each individual is verified against the consumed energy.

Hardware and IoT-based investigation is another aspect of detecting electricity theft. In this approach combination of hardware and software are integrated. The use of hardware devices is a complex method and excessive costs are required for its fabrication, installation, maintenance, and services. Moreover, such systems are vulnerable and can be easily interfaced, which leads to their malfunctioning and tampering. Literature reports many hardware-based approaches, however, no practical implementation has been reported. Furthermore, it is a complex network to fabricate and consumers can easily bypass it.
Keeping in view the total investigation mechanisms this review article considers all such approaches and concludes them into a single review article. Such a contribution is a novel approach and contributes to the literature. The aforementioned approaches are further divided into various case study scenarios along with their statistical representation. Different research questions are reported and validated in the study to cover the overall ETD scenarios from different aspects of the existing literature.
Furthermore, an aspect for identifying various shortcomings identified by various researchers and their proposed solutions is added to provide all the concerned and associated limitations of the literature in a single review. Such analysis can improve the investigation of ETD scenarios and novel approaches can easily be integrated to keep in view the aforementioned limitations of the literature. Moreover, different evaluation parameters are identified and highlighted in order to provide a pathway for the researcher to explore maximum evaluating factors for identifying electricity theft.

3. Research Contribution and Novelty

To conclude the discussion, the literature presents issues of data balancing, data augmentation, manipulated data synthesis, higher data reductionality, high false positive rate (FPR), issue of non-sequential data, exogenous factors, low detection rate, and so many other related factors. This study is a review study and highlights all the related issues of NTLs presented in the literature. The limitations along with their proposed solutions have been presented. The literature presents different detection approaches for NTLs and there is no specific relationship

among such relationships to present in a single hierarchy. In this article, all the approaches are categorized and presented in five different categories, which include (i) detection scenarios using the synthetic data (ii) consideration of sequential information (iii) consideration of non-sequential information (iv) Neighborhood area-based detection networks, and (v) IOT and hardware-based detection. Each category presents all the similar approaches under a single section. The technical papers are transformed into case studies for each of the category. Each case study is further explored and analyzed through mathematical, statistical, and graphical representations. The aforementioned four categories (i-iv) are statistically analyzed, however, the case studies based on IOT and hardware-based approaches are simply presented as there is no statistical representation of such studies in the literature. Moreover, all the identified limitations of the literature are properly tabulated and explored along with their proposed solutions, which is another novel aspect of this review article. The tabulated limitations are helpful for researchers to access the information in a single framework. Furthermore, all the evaluating matrices of the classification scenarios are tabulated and presented, which highlights the impacts of evaluating matrices and their use in various scenarios. such a comparative statement provides a pathway to explore novel evaluating parameters to improve the classification scenarios. Additionally, a comparative analysis based on the handcrafted F1-score is presented, which highlights the importance of an efficient detection scheme (detector) for the investigation of the aforementioned case studies. The F1-score is evaluated to develop a comparative statement among different detectors and investigate to investigate the best one.
To the authors' knowledge, no such review work has been presented before. This article presents a novel approach to collecting all the available approaches into a single review article along with their statistical, mathematical, and evaluating parameters. The motivation and aim of the study are to help readers, researchers, and developers identify efficient and reliable measuring parameters and solutions for the optimal classification scenario.

3.1. Research Motivation

Electricity theft detection (ETD) is one of the major research issues where NTLs disturb the financial graph of UPs. UPs seek many solutions along with the on-site inspections, however, still is uncontrollable. On-site inspection limits the continuous stealing though the a-periodic theft is much difficult to detect and investigate. Literature proposed AMI-based topologies with the help of NAN, which monitors both sides of consumption i-e the sending side and receiving side though the issue remains questionable. Demographical, topographical, geographical, and other exogenous factors affect the consumption pattern's morphology. Similarly, hardware-based approaches have been proposed with high operating efficiency, however, still flaws are present, and fair analysis is yet awaited. Furthermore, data-driven mechanisms based on machine learning in the form of supervised and unsupervised approaches have been implemented but the issue remains unresolved. Moreover, many other solutions have been presented using IOT-based schemes achieving higher efficiencies but in vain. This study summarizes all the aforementioned approaches in the form of case studies to summarize the literature in one article. It reviews many articles and their proposed solutions. The summary of all such approaches along with the identified problems and evaluating parameters is presented in this article. Readers, researchers, and R&D developers can seek compact and detailed knowledge of various technical studies, which are transformed into case studies with statistical analysis.

3.2. Data Collection

The data relating to this review article is searched from Google Scholar, Scopus, and other online research-related platforms. Twenty years of research articles (2003-2023) have been considered in this article, to provide a detailed survey of ETD and NTLs. The data collected are yearly based data and Fig 8 shows the detailed information of all the related data. Furthermore, Table 1 addresses the research questions for the reviewed study and 2 shows their validations. The questions are presented to capture the maximum deliverables of the literature regarding ETD and NTL detection scenarios. These questions are developed to cover maximum aspects of the literature from different perspectives. One of the aspects is to highlight the categorization of the approaches, their comparative analysis, and their statistical representation. Another aspect is to highlight the importance of evaluating parameters used in different case studies. Moreover, the review article presents identified limitation aspects of the literature along with their proposed solutions.

4. Literature Case Studies

NTLs are a serious issue and are of great attention. As it surges the financial burdensome over the UPs, which results in huge revenue loss every year. To tackle such issues literature presents many solutions, however,

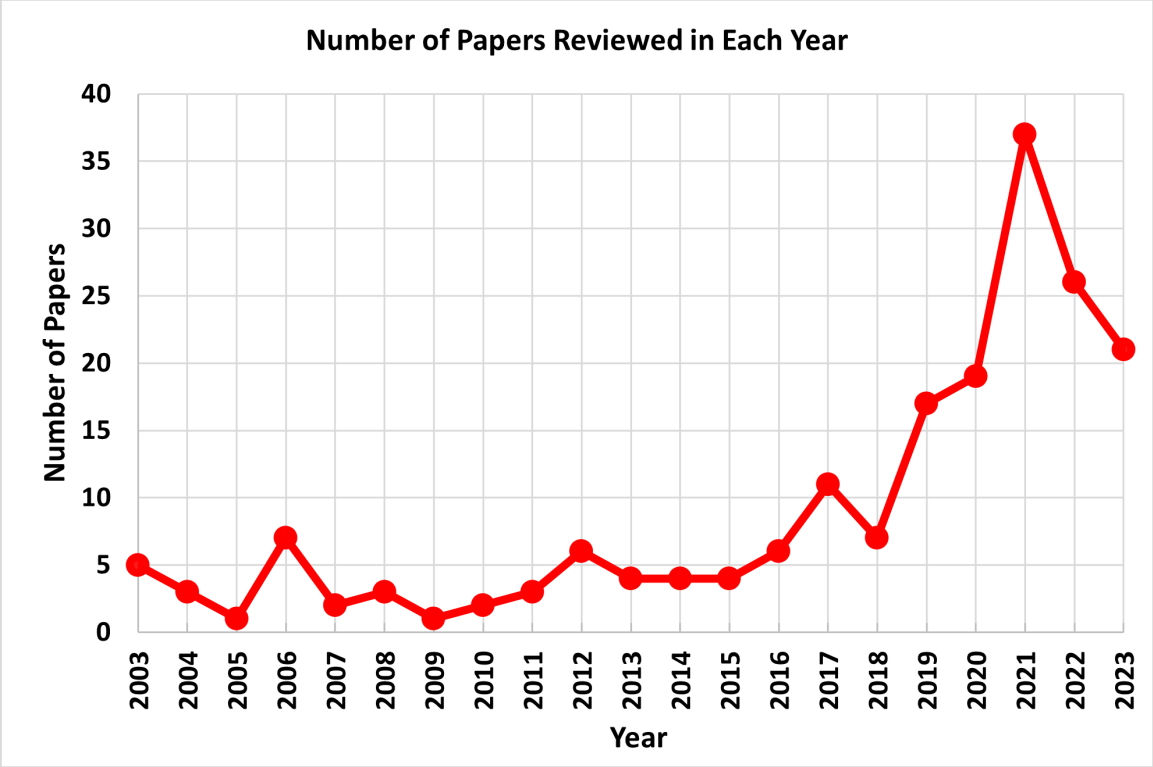

Figure 8: Number of Papers Reviewed Per Year

none of them is yet practically applied to operate reliably. Many data-driven approaches have presented efficient solutions though they are not still practically applicable. Fig 9 shows various methodologies used in this paper. It presents data-driven approaches, IOT-based schemes, and hardware-based designs for the identification and investigation of NTLs. All the detailed analyses of data-driven approaches are presented in Fig 9. The detailed analysis of data-driven approaches requires data preprocessing modules, data scaling, data augmentation, and data segmentation for training and testing of the algorithms. Furthermore, few IOT-based approaches are presented, however, such approaches are not practically yet common. Moreover, hardware-based designs are investigated using AMI and NAN-based protocols. Additionally, the whole review article is summarized in Fig 10, which provides a detailed study of the whole paper. To investigate NTLs statistical analysis of technical studies is targeted, which are transformed into case studies. the case studies are categorized into five different groups those are (i) detection schemes using synthetic data (ii) consideration of sequential data (iii) consideration of non-sequential information (iv) NAN-based topologies using AMI and (v) IOT and hardware-based schemes. Some of the case studies with prominent solutions are described in the next section.

5. Case studies Data Using Synthetic Data

These case studies use a synthetic data-based approach to investigate NTLs. The synthetic data is synthesized using the aforementioned theft and FDIs-based techniques. Usually, the synthesized data are in manipulated data form. Some of the case studies utilizing such approaches are as follows.

5.1. Case Study-1

In case study-1, the SM data of the State Grid Corporation of China (SGCC) dataset is considered [24]. SGCC is an openly available dataset, which was administered in 2014-2016. It consists of 1035 days with a total record of 42372. It is a 24-hour-based data of the consumers. The total number of benign consumers is 38756 and fraudulent is 3616. This study considers the samples of only benign consumers. The benign consumers' data are manipulated to synthesize manipulated data of fraudulent consumers. The data are manipulated using six theft cases. For each benign data sample, six theft variants are synthesized with 1:6. Initially, a random

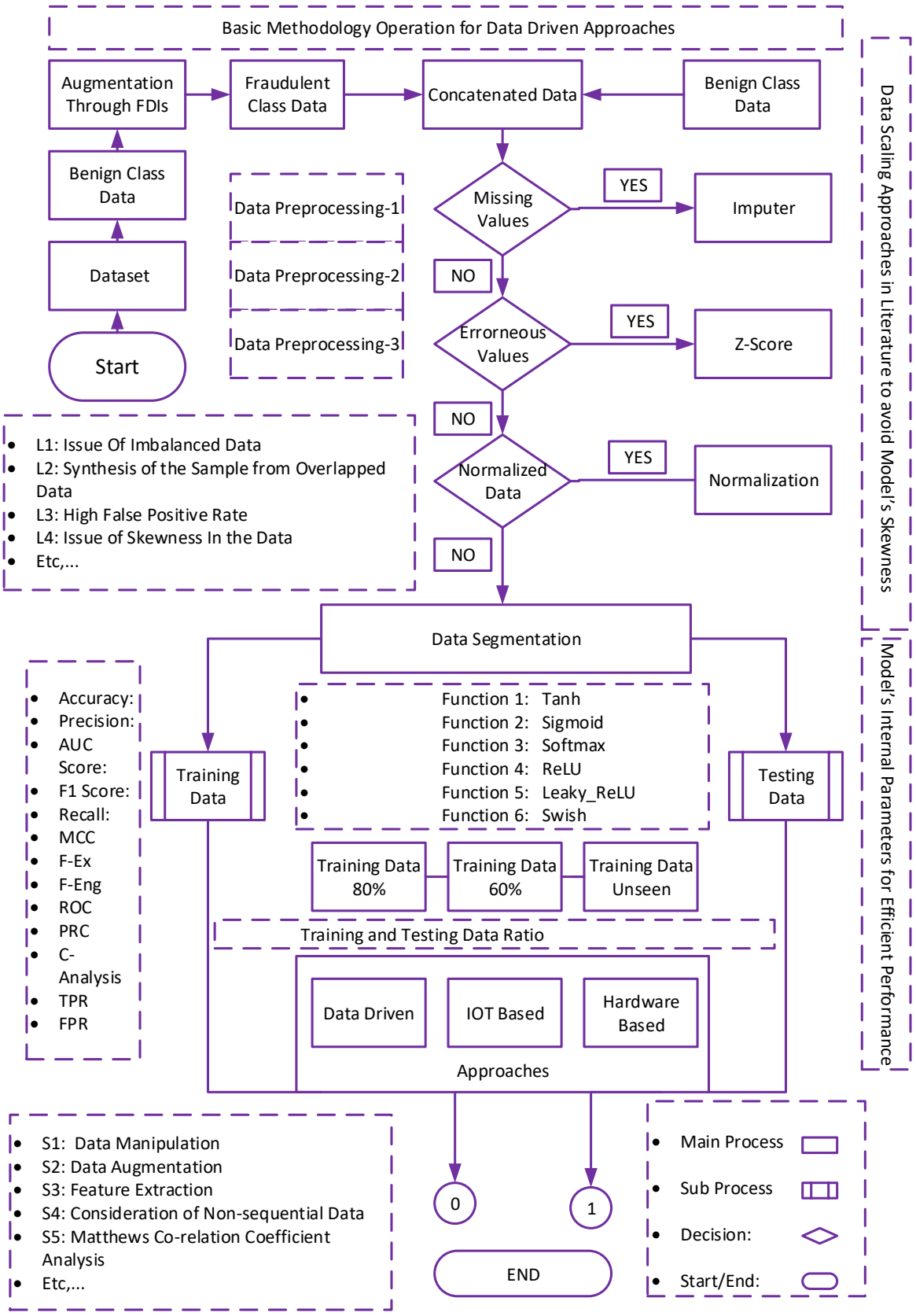

Figure 9: System Model for Data-Driven Approaches for Detection of NTLs

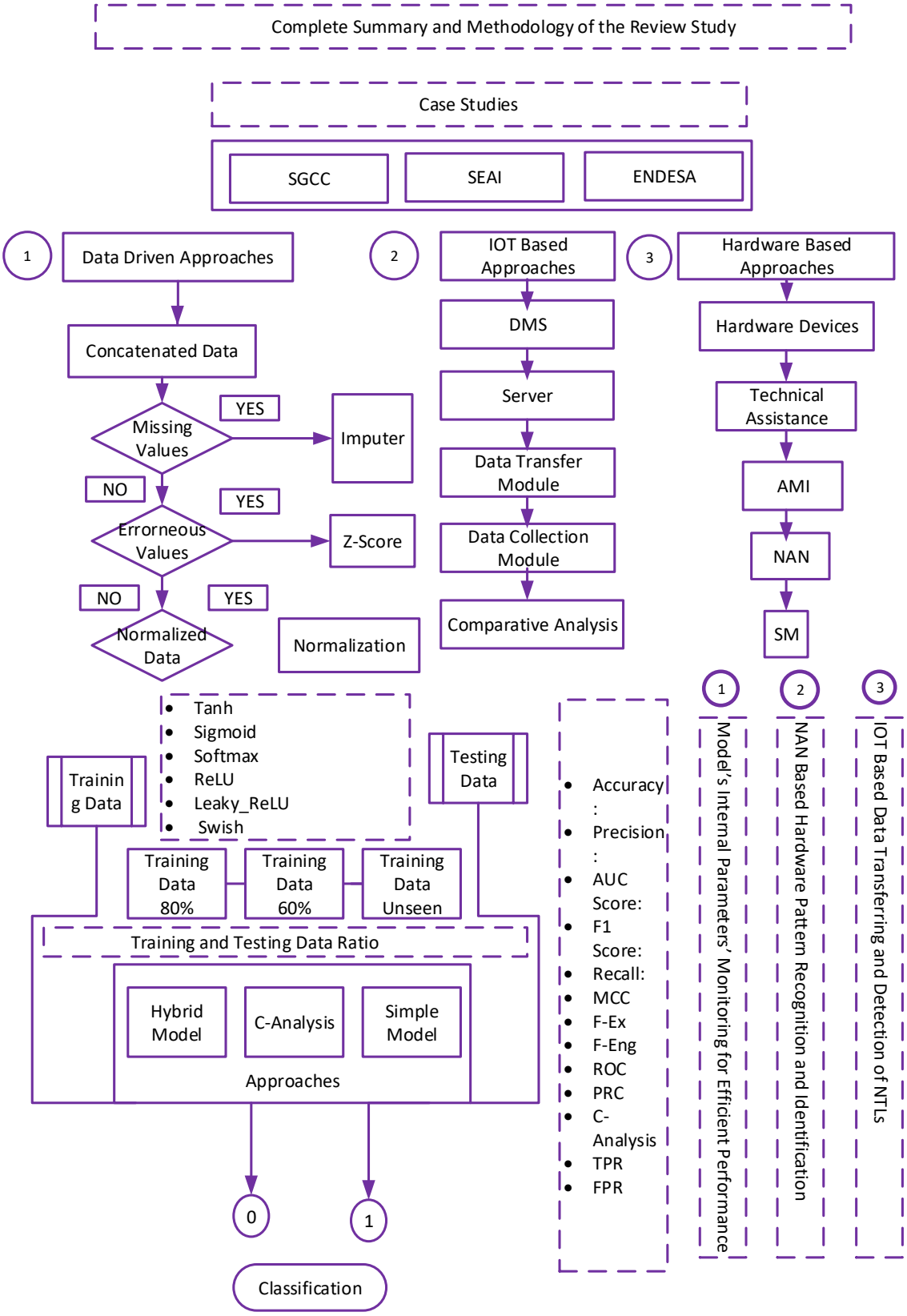

Figure 10: Summary of the Review Article and Different Detection Methodologies Reviewed

1500 benign consumers are considered. The 1500 benign consumers undergo six theft cases and synthesize 9000 manipulated consumers. The number of fraudulent consumers is increased as compared to benign consumers. So a data balancing technique is opted for the minority class data. Synthetic minority oversampling technique (SMOTE) is considered to balance the data [25]. The augmented and balanced data avoids skewness and biases of the model towards the majority class. After applying SMOTE to the minority class, data samples of both classes are balanced and data preprocessing is carried out [26]. A simple imputation technique is used for data preprocessing. The missing values are filled using a mean-based strategy and erroneous values are removed. Furthermore, an issue of cross-pair is highlighted and tackled. Cross pairs result in a high FPR and mislead the classifier. The cross pair is a combination of benign and fraudulent data samples, which resides beside the decision boundary. Cross pairs have traits of both classes and inrush the decision boundary. Such characteristics of the data mislead the classifier and the decision is not carried out in an affine and efficient mode. Afterward, the data are split into training and test slabs. Training data are considered 70% and testing data are taken as 30%. A novel model of Bi gated recurrent units and Bi long short term memory BiGRU-BiLSTM is integrated [27]. The integrated model is presented as a novel and proposed model. To evaluate the performance of the integrated model, FPR, the area under the curve (AUC) and precision are considered as evaluating metrics. Fig 11 shows the statistical investigation of case study 1 where AUC is used to monitor the performance of the hybrid model.

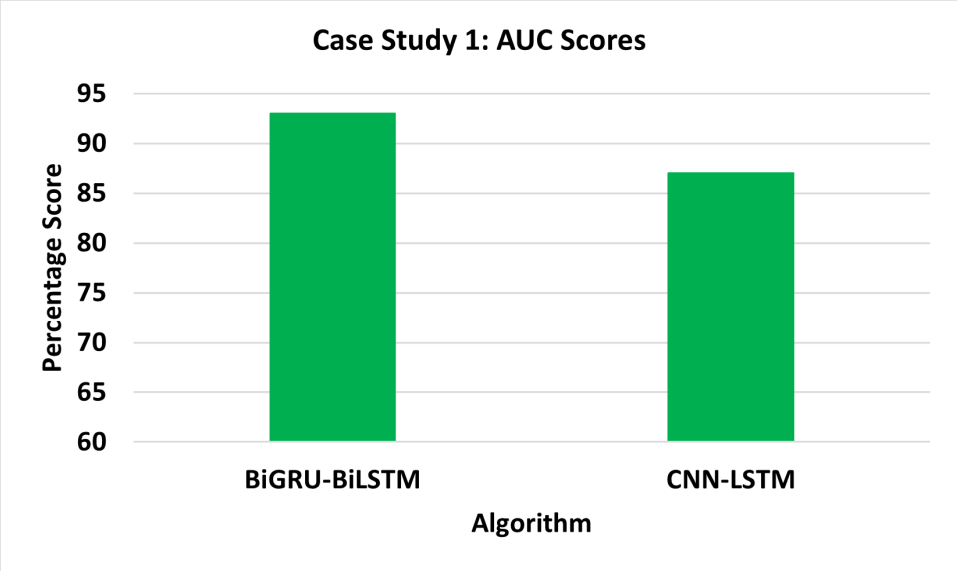

Figure 11: Statistical Analysis of Case Study 1

5.2. Case Study-2

In case study 2, the same dataset of SGCC is used [28]. It is an extension of case study 1. Few major contributions have been added to the extended work. It considers the same number of 1500 benign consumers and uses theft cases to synthesize theft class data. For each benign class sample, six theft versions are synthesized. The number of the theft class samples is increased as compared to the manipulated data samples. Henceforth, to tackle the bias issue in such scenarios data balancing technique SMOTE-SVM is used. Afterwards, the balanced data are preprocessed to eliminate the useless information and the null values are filled using mean mean-based strategy. The balanced data are transformed and extra features are synthesized through stochastical feature engineering. Stochastical features of min, max, mean and standard deviation are synthesized, which contribute as prominent features of the data. Synthesizing feature engineering is the opposite mechanism to data reductionality, where some part of the information is discarded. A Tomeklinks technique is used to remove the cross pairs across the decision boundary. The inrush cross pairs cause misclassification issues and result in high FPR [29]. Furthermore, it provides a comparative analysis with various base models LSTM [30], support vector machine (SVM), decision tree (DT) [31]-[32] and random forest (RF) [33]. BiGRU-BiLSTM is a hybridized novel model, which is used for classification scenarios. It has been observed that without feature engineering the

model's efficiency is 88% and it is increased to 95% when a stochastic feature engineering scheme is opted. Fig 12 shows the complete statistical analysis of case study 2. To evaluate the performance of the model, F1-score, precision, recall, AUC-score, and accuracy are used as evaluating metrics.

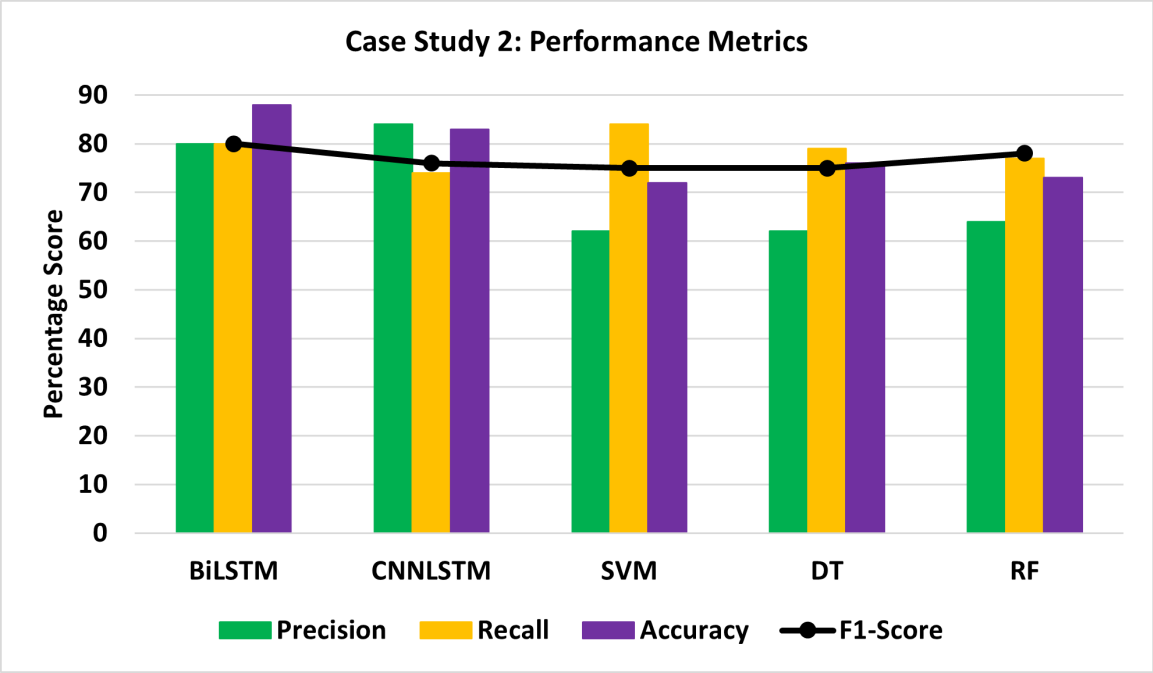

Figure 12: Statistical Analysis of Case Study 2 based F1-Score, precision and Precision

5.3. Case Study-3

In case study 3, the benign data of SM are considered and data manipulating techniques are applied to synthesize theft class data. Six variants of novel FDIs are introduced. The novel FDIs are introduced in comparison to six theft cases. For each benign sample, six variants are synthesized. The synthesized data samples are increased in number and data balancing is required. Initially, the data are preprocessed using a simple imputer [34]. Simple imputer uses a mean-based strategy to fill the missing values. The preprocessed data are transformed using min-max data normalization. The data are then balanced using a data augmentation mechanism. The balanced data are segmented into three portions i-e training data, testing data, and unseen data [35]. A novel integrated model is presented as a classifier for binary classification. The integrated model is combination of attention layer [36]-[38], LSTM layer and Inception module [39]-[41]. The integrated modules introduce a novel model of AttenLSTMInception. Afterward, the classification results are observed in three different stages i-e training@40%, 60%, and 80%, respectively. To evaluate the performance of the model, FPR, accuracy, precision, AUC [42]-[43] and recall are used as evaluating metrics. Furthermore, the complexity and variations that occurred in data while applying FDIs, are observed using kurtosis and skewness factors. Both theft classes and FDIs are compared and their comparative analysis is presented using the same parameters. Moreover, a novel sliding window concept is introduced to provide the data in slabs while training the model. The slabs are repeatedly slid over 10 data samples in forward propagation [44]-[46]. It carries 10 previous data samples as well so that the traits of previous information are preserved. The total window size is 20 samples. The statistical analysis of the case study 3 is shown in Fig 13.

5.4. Case Study-4

In case study-4, sequential data of the SGCC dataset is considered [47]. It is a temporal sequence data of benign consumers. The benign data are manipulated through FDIs and manipulated data are synthesized. A total of 1500 benign consumers are considered and six variants of FDIs are applied. Six variants of FDIs synthesize 9000 fraudulent consumers. The difference of 1:6 is observed in data samples and causes a data imbalance issue. Such issues cause model skewness and biases towards the majority class. To tackle such issue

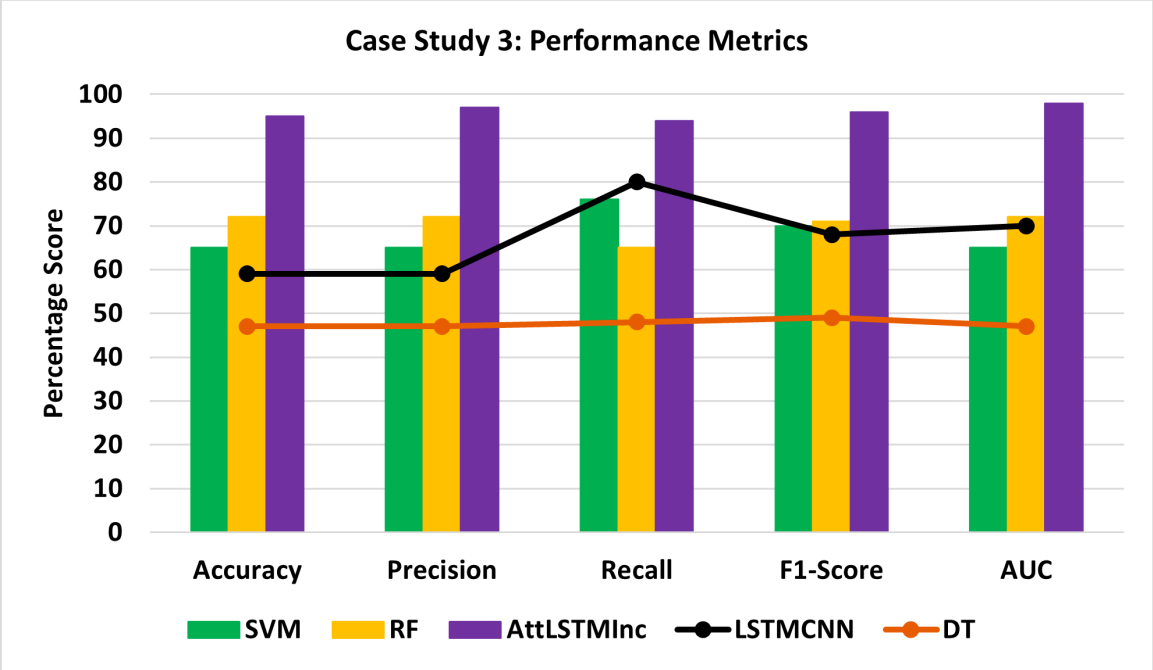

Figure 13: Statistical Analysis of Case Study 3 based on Recall and F1-Score

data augmentation is carried out using a borderline SMOTE. Borderline SMOTE oversamples the minority class samples and balances both classes' data [48]-[49]. The balanced data are preprocessed using a simple imputer. The imputed data are transformed using data normalization. The balanced data are then provided as input for the training purpose of the model. The training data is used for training and testing data are used for validation purposes of the model. The same model used in case study 2 i-e AttenLSTmInception is used as a classifier for binary classification. To introduce the novelty in the study various activation functions are used. The efficiency of the model is monitored using various activation functions. Activation functions of the sigmoid, softmax, Tanh, Leaky-RELU, RELU, and swish are used and shown in equations 13- 18.

$$Sigmoid\ f(z) = \frac{1}{1 + e^{(-z)}}. \tag{13}$$

$$Softmax\ (z_i) = \frac{exp\ (z_i)}{\sum_j exp\ (z_j)}. \tag{14}$$

$$f(z) = Tanh(z) = \frac{2}{1 + e^{-(2z)}} - 1. \tag{15}$$

For each activation function, an individual analysis is presented. The analysis is monitored using evaluating metrics of AUC, FPR, precision, recall, accuracy, and F1-score. Furthermore, to investigate the data relevancy before and after data augmentation [50] Mathews correlation coefficient (MCC) is used. Before data augmentation, MCC is zero for the proposed Attention LSTM Inception model.

$$f(z) = ReLU(z) = \begin{cases} max\ (0,\ z) & ,\quad z \leq 0 \\ 0 & ,\quad z < 0 \end{cases} \tag{16}$$

$$f(z) = \begin{cases} z_i & ,\quad if\ z_i > 0 \\ a_i\ z_i & ,\quad if\ z_i \leq 0 \end{cases} \tag{17}$$

After data augmentation, it is reported as MCC=0.88. MCC shows the reliability of the model [51]-[52]. The maximum value of MCC is 1 and the minimum is 0. MCC value of 0.88, shows that the proposed model

performs efficiently. The performance of the model is compared with the base models. The base models include convolutional neural network (CNN) [53]-[55], CNNLSTM, SVM [56], DT and RF [57]. Fig 14 shows the statistical performance analysis of the proposed study.

$$f(z) = (\beta * sigmoid(z)) \tag{18}$$

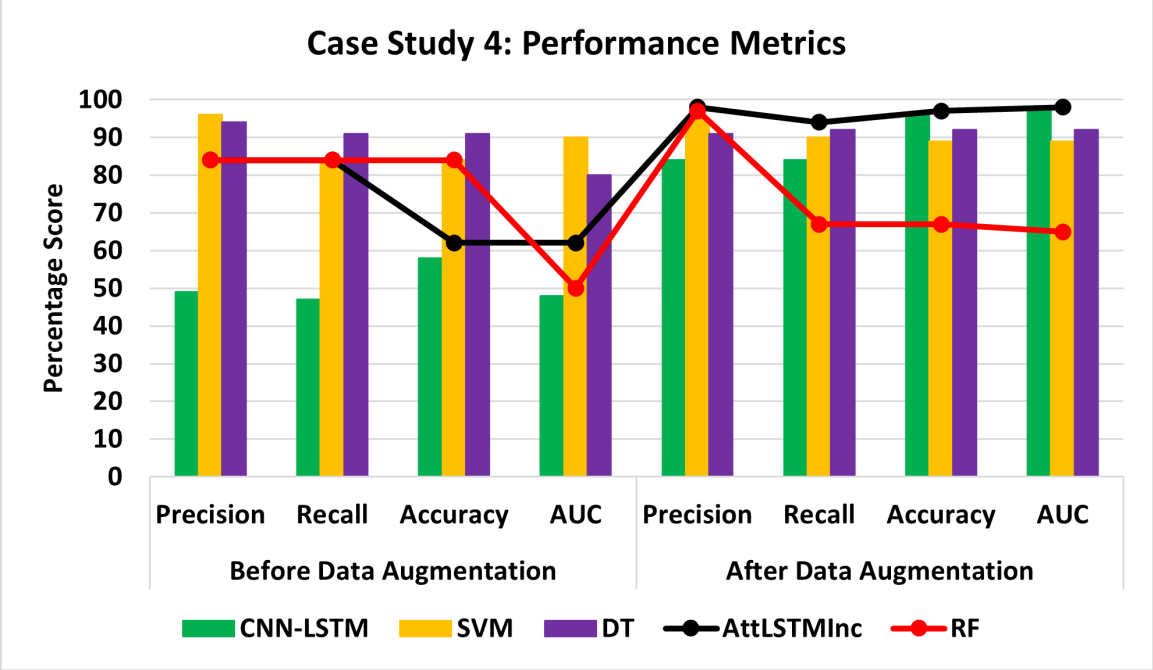

Figure 14: Statistical Analysis of Case Study 4 Before and after Data Augmentation

5.5. Case Study-5

Case study 5 targets electricity theft due to NTLs. It uses the Sustainable Energy Authority of Ireland (SEAI) dataset, which was administered in 2012 [58]-[60]. It has 5000 residential consumers and is half-hourly based. All the consumers of the dataset are benign consumers and their data are manipulated to synthesize the fraudulent consumers' data. Six data theft variants are used for data manipulation. In order to tackle the data imbalance issue, a simple data oversampling technique SMOTE is used. SEAI contains both sequential and non-sequential information of the consumers. Sequential data are SM recorded data of the consumed energy whereas non-sequential data are demographic and topographical based information. To preprocess the data, data imputation and transformation are carried out. The data are imputed and outliers are removed. For data transformation, data normalization is applied. Furthermore, the behavior of the consumer is analyzed and CNN is applied to extract the vital feature vectors. The feature vectors of the data represent consumer behavior and provide information about the usage of the consumed energy. Each day's energy consumption is analyzed and changes between the curves are observed to highlight the differences. Such a study of daily profiles helps to investigate the malicious consumption pattern. Moreover, the balanced data are proportioned into training and testing data. The model is trained using training data and testing data are used to investigate the efficiency of the proposed classifier. CNN-RF is used as a binary classifier. To compare the performance of the proposed model, RF, gradient boosting decision tree (GBDT), CNN, CNN-GBDT, CNN-SVM, SVM [61]-[62], RF and logistic regression (LR) are used as base models. LR uses the sigmoid activation function and SVM uses the kernel function for binary classification [63]. The proposed model outperforms the base models. Accuracy, precision, recall, F1-score, and AUC are used as performance metrics. Fig 15 shows the statistical analysis of case study 5 on SEAI and LCL datasets.

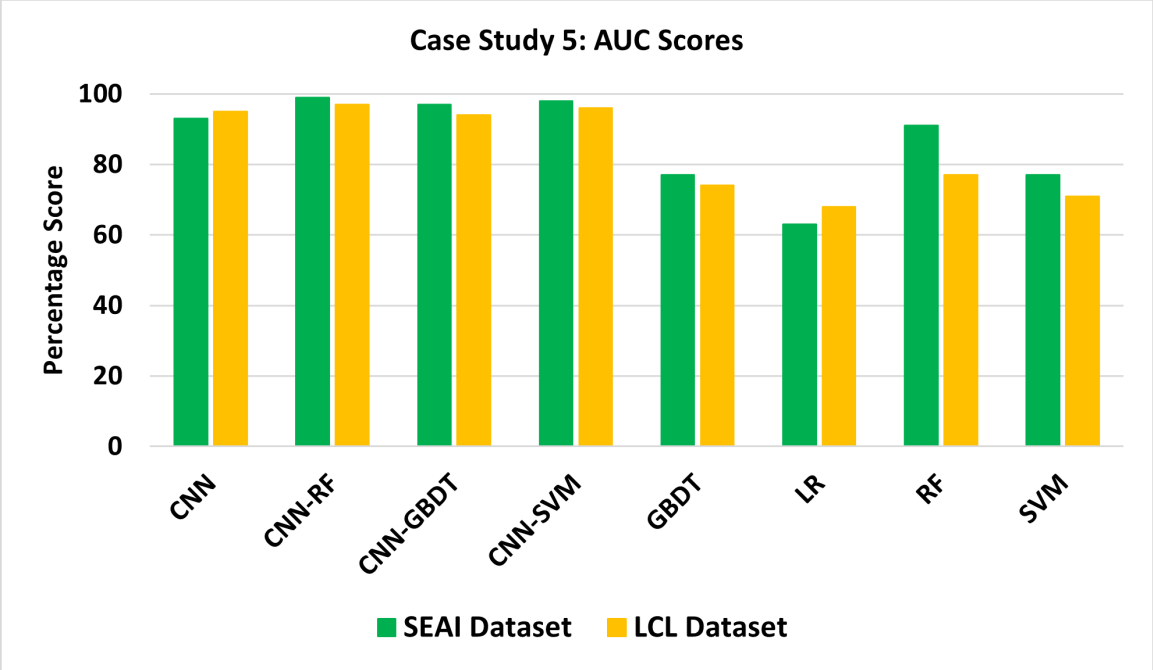

Figure 15: Statistical Analysis of Case Study 5 Using SEAI and LCL datasets Dataset

5.6. Case Study-6

In case study 6, a dataset of SM from London is used. The time instance for the recorded data is 30 minutes. For each day 48 features are presented to mimic the daily based energy consumption. Initially, the data are preprocessed using the same traditional approaches [64]. The preprocessed data are then normalized using min-max data normalization. The dataset provides benign data samples only. The benign data are manipulated using attack models. The attack models are of physical, communication, and data types. This study highlights the importance of data attack types. Five different attacks are used to manipulate the SM benign class data, which results in five different variants for each benign sample. The benign class data of the consumers is sequential. The manipulation of the data is carried out in random mode and no specific sequence of the data is observed. To avoid the class bias issue toward the majority class, the benign class data are augmented and balanced [65]-[66]. The traditional random oversampling method is used to synthesize the over-sampled data. Furthermore, to segregate the theft class data patterns, a conditional variational auto-encoder (CVAE) is proposed. Initially, the theft patterns are featured into low dimensional vectors using convolutional layers-based autoencoder. Later on, the same theft patterns are regenerated through the reconstruction strategy of autoencoders using deconvolutional autoencoders. To classify the consumers a CNN model is used as a classifier. The adaptability and performance of the model before and after data augmentation is monitored. To evaluate the case study, accuracy, macro F1-score, and G-mean are used as evaluating metrics. The results of the classifier are compared with MLP, SVM, and extreme gradient boosting machine (XGBoost) [67]. Fig 16 shows the statistical analysis of case study 6.

5.7. Case Study-7

In case study 7, the importance of theft detection is analyzed. The theft attacks are classified into two categories i.e. physical attacks and cyber attacks [68]-[70]. Reasons mentioned for the occurrence of the physical attacks are [71]-[72]: bypassing of the SM, illegal tapping [73], placing a strong magnet, reversing the meter, and non-intentional meter malfunctioning [74]. Cyber attacks are data attacks, which are occurred due to password extraction, erasing record events, interrupting data transmission, false data injection, and modifying the system software. This study highlights 12 theft cases in different scenarios. The theft cases are data-oriented approaches where the data of SM are manipulated. The benign data are initially considered and manipulated data are synthesized using the aforementioned 12 theft cases. For each benign sample, twelve variants of manipulated data are synthesized. The synthesized variants are in a ratio of 1:12. To avoid model skewness

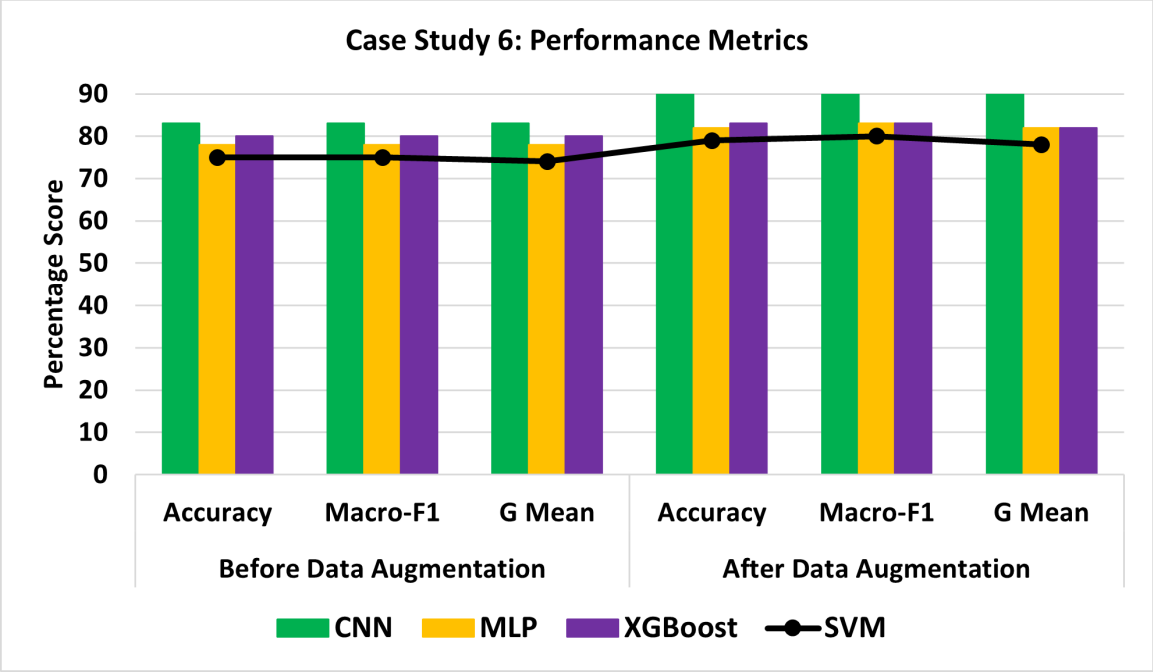

Figure 16: Statistical Analysis of Case Study 6 based on AUC

issues, the data are augmented and balanced using various data balancing schemes. The manipulated data have features like constant or varying reduction in energy consumption, void or zero consumption, irregularity in consumption patterns and abrupt decrease in energy consumption [75]. To monitor the behavior of such attacks various models are used for detection scenarios. The detection models are the binary classifiers and classify the classes based on their patterns. The patterns are later observed and reconsidered for classification. The benign consumer is labeled as 0 and the fraudulent consumer is labeled as 1. SVM, K-nearest neighbor (KNN) [76], Gradient boosting classifier (GBC), naive Bayes (NB) [77], RF, LR [78], linear discrimination analysis (LDA) [79] and AdaBoost classifier (ABC) are used. The performance of the case study is analyzed using accuracy, F1-score, receiver operating curve (ROC) [80]-[81], and AUC.

5.8. Case Study-8

In case study-8, a dataset of SEAI is used, which was administered in March 2012. SEAI contains honest consumers' data [82]-[84]. To synthesize the fraudulent consumers' data six theft case scenarios are considered. The scenarios are as follows:

- Cheating continuously $\alpha > 0$, β=0 and $\alpha + \beta > 0$
- Faulty constantly $\alpha < 0$, β=0, $\alpha + \beta < 0$
- ON-peak manipulation $\alpha = 0$, $\beta > 0$, $\alpha + \beta > 0$
- ON-peak faulty $\alpha = 0$, $\beta < 0$, $\alpha + \beta < 0$
- OFF-peak manipulation $\alpha > 0$, β=-α, $\alpha + \beta = 0$
- OFF-peak fault α=-β, $\beta > 0$, $\alpha + \beta = 0$

The benign class data are manipulated using the aforementioned data manipulation scenarios. The benign and theft data are concatenated and a dataset is synthesized. The data are initially preprocessed, where the anomaly coefficient, detecting coefficient, and corresponding p-value of each consumer is calculated. Two LR-based models are presented to investigate the scenario. In order to initiate the consumer's identification dummy variables are assigned. Variable x=0 is used for OFF-peak and x=1 is used for ON-peak scenarios. A linear regression-based scheme for the detection of energy theft and defective smart meters (LR-ETDM) is used to

detect malicious consumers and faulty meters. It detects the consumers who are manipulating their SM data inconsistently and constantly. However, it fails to detect the consumers who steal the energy during a certain period of time. It is able to detect five theft case scenarios. In order to detect the consumers with uniform and occasional manipulation a categorical variable-based (CVLR-ETDM) algorithm is proposed. CVLR-ETDM is the improved version of LR-ETDM, which tackles the issue of less detection efficiency of LR-ETDM. The improved version is able to detect the consumers with variable anomaly coefficients. The simulation of the case study is carried out in Matlab Simulink. To monitor the performance of the case study, 45 consumers were evaluated. A successful investigation is carried out to identify malicious consumers and faulty meters.

5.9. Case Study-9

In case study-9, NTLs due to prosumers are highlighted. Prosumers are those consumers who have two metering infrastructures. One is known as an import smart meter (ISM) and the other one is an export smart meter (ESM). The ISM is responsible for monitoring the consumed energy whereas ESM is a meter, which monitors the supplied energy from the prosumer's distributed energy resources (DER) [85]-[86]. DER may be solar panels or wind turbines. The consumers are divided into two categories i-e external adversary and internal adversary. External adversaries are those consumers who tamper with the data using physical approaches, however, internal adversaries change the reading of the central data management system where data are available. The study considers a new dataset, which is generated through GridLab-D. It is a synthesized dataset using taxonomy distribution feeder R1-12.47-2 [87]. It consists of 1594 residential users 49 of them are prosumers with solar panels. The data is 15 minutes long. The dataset is feature-based, which contains static, dynamic, and weather parameters. All the 1594 users are benign. To synthesize the theft class data various novel theft attacks known as 'balance attacks' are used. The balance attacks are eight in number and are as follows.

- In attack-1, 2 the ISM reading is decreased using a constant value, which is subtracted from the whole consumption. It may be prosumer or consumer-based. The equations 19 and 20 show the mathematical representation of attacks 1 and 2.

$$Attack - 1 : ISM_n = ISM_x - P \tag{19}$$

$$Attack - 2 : ISM_n = (ISM_x)(1 - \frac{Q}{100}) \tag{20}$$

- In attack-3, 4 the prosumer tries to increase the ESM reading using a constant factor P or constant percentage. The equations 21 and 22 show the mathematical representation of attacks 3 and 4.

$$Attack - 3 : ESM_n = ESM_x + P \tag{21}$$

$$Attack - 4 : ESM_n = (ESM_x)(1 + \frac{Q}{100}) \tag{22}$$

- In attack-5, 6 NAN based anomaly is reported [88]-[89]. The ISM reading is decreased using a constant value parameter P or percentage Q. However, the same decrease is adjusted in the neighbor's data to keep the data of the central data management system balanced. The equations 23 and 25 shows mathematical representation of attacks-5 and 6.

$$Attack - 5A : ISM_n = ISM_n - P \tag{23}$$

$$Attack - 5B : ISMm = ISM_m + P \tag{24}$$

$$Attack - 6 : ISM_n = (ISM_n)(1 - \frac{Q}{100}) \tag{25}$$

$$ISM_n = (ISM_m)(ISM_m \frac{Q}{100}) \tag{26}$$

$$Attack-7: ESM_n = ISM_m + P \tag{27}$$

$$ESM_m = ISM_m - P \tag{28}$$

$$Attack-8: ESM_n = ESM_n(1 + \frac{Q}{100}) \tag{29}$$

$$ESM_m = ESM_m - (ESM_m)(\frac{Q}{100}) \tag{30}$$

- In attack-7, 8 ESM reading is increased by the prosumer using a constant parameter P or percentage Q. The value of the same reading is adjusted in the neighborhood's data by decreasing their DER or ISM data. The equations 26- 30 shows mathematical representation of attacks-7 and 8.

The attacks are applied to the benign class data. The data are normalized using the standard-scalar mechanism [90]-[91]. The normalized data are augmented and presented for classification. Various base models, DT, KNN, LR, NB, NN, and SVM are used to detect the theft case scenarios. To evaluate the performance of the base models, accuracy, recall, and precision are used as evaluation metrics. Fig 17 shows the statistical performance analysis of the case study-9.

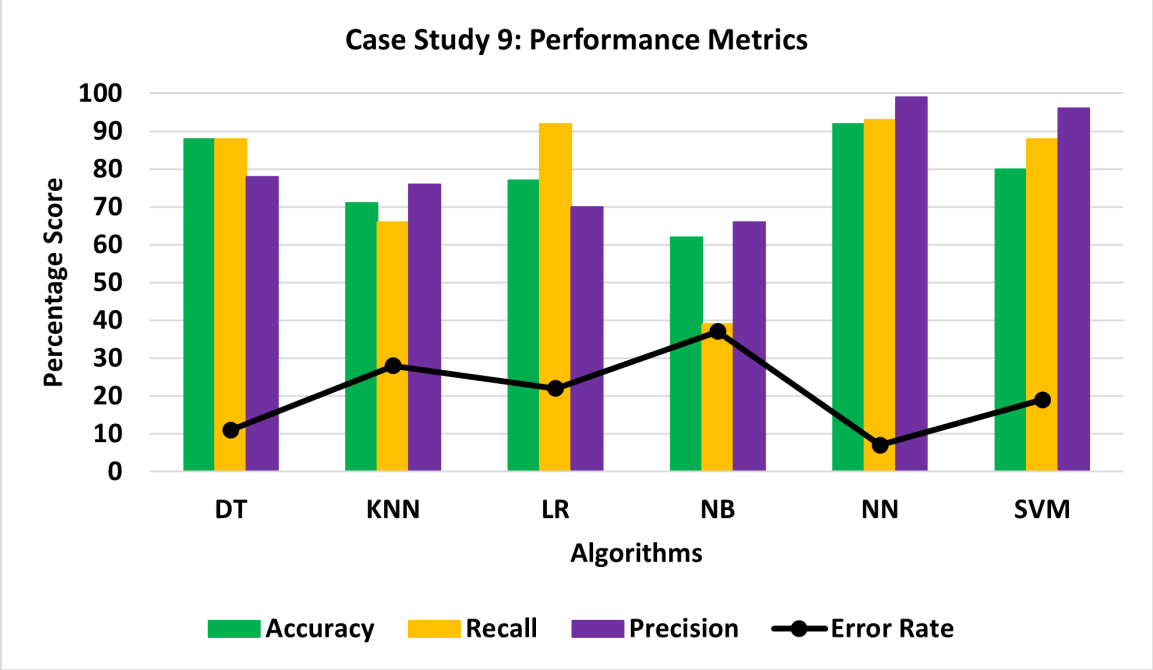

Figure 17: Statistical Analysis of Case Study 9 based on Precision, Recall, Accuracy, and Error Rate

6. Case Studies Using Non-sequential Data

Non-sequential data are the exogenous parameters of the data, which are used to enhance the detection scenario of NTLs. The non-sequential data are used along with the sequential data of the SMs in such detection schemes. Non-sequential data itself doesn't provide detailed information about the consumers that's why it is necessary to consider the sequential temporal sequence data. A few of the case studies investigated based on non-sequential information are as follows.

6.1. Case Study-10

In case study-10, dataset of ENDSEA is used [92]-[96]. It is a dataset of sequential and non-sequential information. Sequential data is temporal sequence data whereas non-sequential data are consumers' demographic information. It is a challenging task to read such information [97]. The dataset contains only benign consumers. To synthesize the theft class data, six theft variants are used. The theft variants manipulate the sequential data only. The ratio of benign and manipulated data is 1:6. To balance the data, the data augmentation technique SMOTE is used. The augmented data are then preprocessed using simple imputer [98]. Null values are filled using mean based strategy and erroneous values are removed. The affine and clean data are then subjected to training and testing purposes. Multilayer perceptron (MLP) is used as a proposed model to read the sequential and nonsequential information. MLP is used for binary classification in this study. The benign consumers are labeled with 0 and fraudulent consumers are labeled with 1. MLP outperforms the base models. SVM, DT [99], RF, and LSTM are evaluated as base models. The performance of the base models is very poor due to the presence of non-sequential information. To evaluate the performance of the model AUC, precision, recall, F1-score, and accuracy are used as evaluating metrics. Moreover, the main focus of the study is highlighting the issues of imbalanced data, non-availability of suitable models for non-sequential data, high FPR [100], and low TPR.

6.2. Case Study-11

In case study-13, a customer behavior trials (CBT) based dataset is used, which was administered in 2009-2010 by CER Ireland. It is considered an extensive dataset, which is monitored every 30 minutes. It contains sequential and non-sequential data [101]-[102]. Non-sequential information is socio-demographic based attributes, which include age, social class, employment status, and household covered area. It is a dataset of 4096 consumers. All the consumers are residential consumers and belong to the benign class. Machine learning mechanisms use both class data i-e., benign and fraudulent classes. Henceforth the benign class data are manipulated through six data theft cases and theft class data are synthesized. An algorithm of finite mixture models (FMMs) is used, which uses a probabilistic framework instead of deterministic and overlapping segments. The FMMs are considered soft assignments due to their probabilistic framework. FMMs operate over the training dataset and five soft segments of the consumers are derived. FMMs operation over the segmented data returns vectors of the standard deviation, mean, and weight of the cluster's center [103]-[104]. Furthermore, the genetic programming (GP) algorithm is used. The GP extracts extra prime features of the data along with the returned vectors of the segmented data [105]. About 200 optimized variables are returned. Finally, the optimized variables from GP and FMMs deliverables are presented in the form of training data, which is proceeded by classification scenario. The trained model is then expected to predict the clusters outside the training data. The association is then observed between the training data's clusters and the predicted clusters. A score value is assigned to each association. The score value provides information about the probability of fraud occurrence for the predicted clusters [106]-[107]. Afterwards, the binary classification is carried out. GBM is used as a binary classifier. The performance of the classifier is evaluated using AUC, accuracy, F1-score, and precision. The performance is monitored before and after feature engineering. A vibrant difference is seen in the performance of the classifier after feature engineering. Fig 18 shows the statistical analysis of the case study-11.

6.2.1. Case Study-12

In case study 12, a novel change and transit-based approach is presented. The change and transit (CAT) is an AMI-based approach, which monitors the transits over the tampered data and responds to it [108]-[110]. A dataset of residual energy disaggregation dataset (REDD) is used. It is a dataset of benign consumers. To synthesize the theft class data, five novel cyber attacks are generated shown in equations 31-35, which are as follows.

The attack-1 multiplies a gamma factor with the consumed data, which is reduced by the ratio of gamma value. It means that the consumption pattern is multiplied by a specific percentage. The specific percentage decreases the consumption and synthesizes a specific pattern for the consumed energy. The pattern is shown as a realistic one in order to confuse the detector. A comparative study is investigated by the detector relative to its initial reading and the malicious one. In attack-2, the previous reading data are sent to the model when the actual data are less than X_l^i otherwise it sends the malicious data $Q_d^n >$.

In attack-3 a threshold is multiplied and subtracted from the realistic recorded data. The threshold is percentage base i-e 10% or 20%.

In attack 4, the consumed data are scaled down by a ratio. The mean of the total energy over a day is calculated

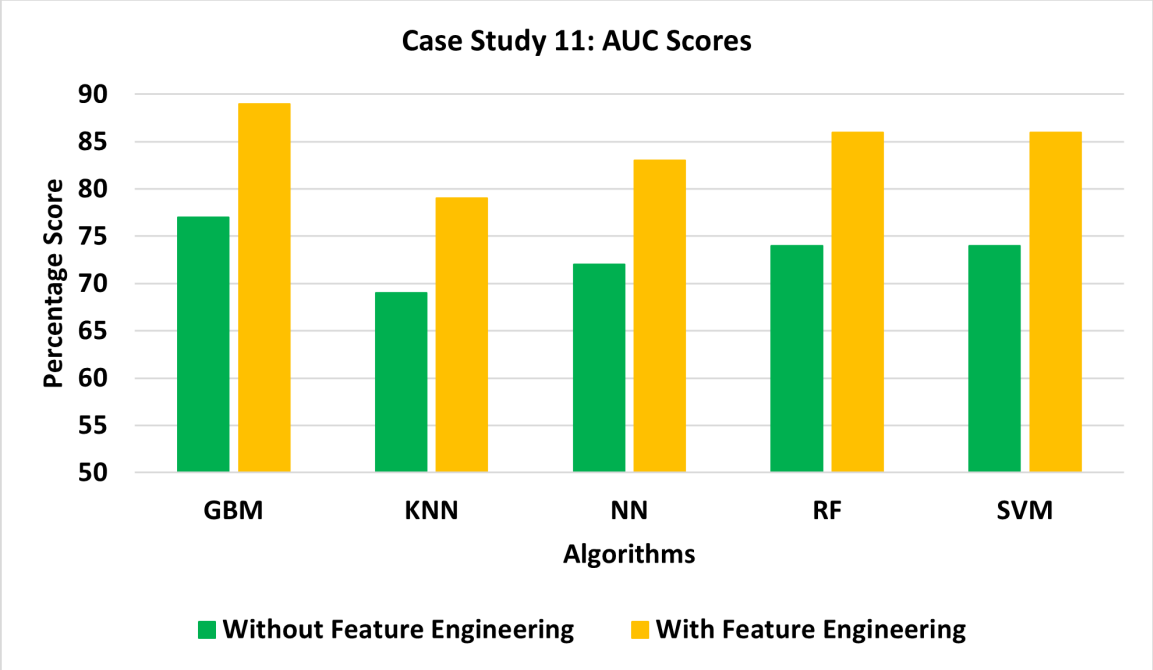

Figure 18: Statistical Analysis of Case Study 11

and multiplied by the ratio factor, which minimizes the readings.
Similarly, attack 5, reduces the consumed energy by a time-dependent ratio. All the cyber attacks follow the CAT approach and transmit the reading when the consumption change exceeds the predefined threshold. The theft data are concatenated with the benign class data. Both class data are augmented and balanced. Later on, the data are normalized. The normalization is opted so that no common feature or group of features dominates the classification scenario. Testing and training data are partitioned. To evaluate the performance of the study, benchmark detectors RF, SVM, MLP, CNN, and GRU are used. To monitor the performance of the benchmark models accuracy, precision, recall, ROC, AUC, TPR, and FPR metrics are used. Fig 19 shows the statistical analysis of case study 12 based on accuracy, detection rate, and false alarm.

$$A-1(p_c^i) = \begin{cases} (n)(p_c^i) & if\ first\ reading \\ (1+q)(p_l^i) & otherwise \end{cases} \tag{31}$$

$$A-2(p_c^i) = \begin{cases} (n)(p_c^i) & (p_c^i < (x_l^i) \\ (1+q)(p_l^i) & otherwise \end{cases} \tag{32}$$

$$A-3(p_c^i) = \begin{cases} Don't\ \ Transit & (p_l^i - p_l^i * Th_i \\ Transit\ \ (p_c^i) & otherwise \end{cases} \tag{33}$$

$$A-4(p_c^i) = C_g E[(p_d^i)] \tag{34}$$

$$A-5(p_c^i) = F_c * (p_d^i) \tag{35}$$

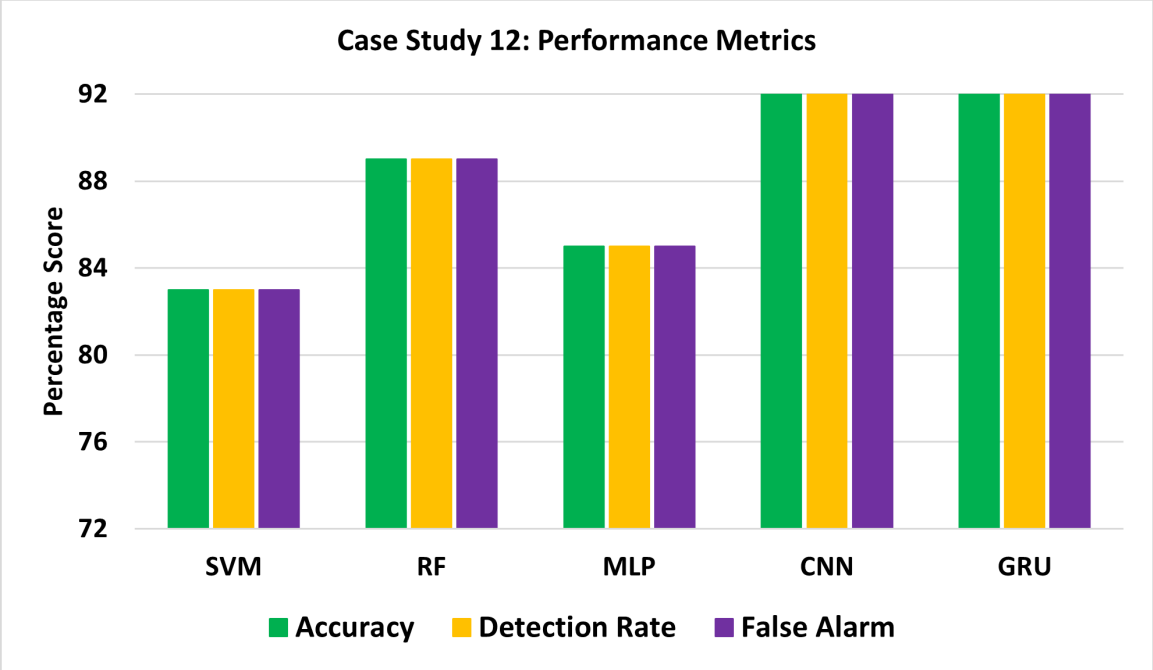

Figure 19: Statistical Analysis of Case Study 12

7. Case studies Using Sequential Data

Sequential data are the temporal sequence data of the SMs. It provides the consumers' consumption information for a specific time period, which can be hourly, daily, weekly, or monthly. Such data are vital information to investigate its originality. Their originality confirms the consumers' behavior and decides whether the consumer is benign or fraudulent. Case studies based on sequential information are as follows.

7.1. Case Study-13

Case study-13 uses two supervised learning-based models to investigate NTLs [111]. It uses the dataset of SGCC. The data are preprocessed using the linear interpolation method, which fills the missing values and removes the erroneous ones. After data preprocessing, the data are normalized using robust scalar normalization [112]. Robust scalar normalization is a similar approach to min-max, however, it uses quartile ranges for the normalization of the data. Afterward, the preprocessed data are augmented to tackle the class bias issue of the model towards the majority class. SMOTE edited nearest neighbor (SMOTEENN) is used for data augmentation [116]. In order to extract the main features of the data locally linear embedding (LLE) is used to extract the abstract features of the data. Furthermore, to classify the consumers' integrity self attention generative adversarial network (SAGAN) and CNN are used. The SAGANCNN is the supervised learning-based proposed model, which is used for binary classification. To evaluate the performance of the model, recall, accuracy, AUC, precision, and F1-score are used as evaluating metrics. In the second supervised learning mechanism, ERNET is used as a binary classifier [117]. Initially, the preprocessed data are augmented using adaptive synthetic edited nearest neighbor (ADASYNENN) [113]-[115]. The augmented data's abstract features are extracted using a spare autoencoder (SAE). Extraction mechanisms of abstract features highlight the main features of the data, which improves the classification scenario [118]-[119]. The classification model ERNET is an integrated model, which is the combination of efficient net, residual net, and GRU. To improve the efficiency of the model robust root mean square propagation (RMSprop) optimizer is used to improve the learning rate of the proposed classifier. Moreover, the proposed model is compared with CNN-RF, CNN, wide and deep CNN (WDCNN), and SAGAN. The proposed model outperforms the base models.

7.2. Case Study-14

Case study-14 is a high FPR-oriented study. It is a comparative study of the data oversampling techniques. A CNN model is presented as a classifier. It is trained on various data types to analyze its performance

over different data nature. To synthesize various data types, cost-sensitive learning, random oversampling, random undersampling, K-mediods based undersampling, SMOTE, and cluster-based data oversampling [120], techniques are opted. The dataset of SGCC is considered and various aforementioned techniques are applied to balance both classes i.e., benign and fraudulent class data. The oversampling techniques over-sample the minority class samples for data balancing, however, the undersampling reduces the majority class samples to balance both classes' data. Oversampling considers the overlapped data to synthesize the minority class data whereas the undersampling techniques discard useful information of the data. SGCC dataset contains the benign class in the majority and the fraudulent class in the minority. Oversampling techniques over-sample the fraudulent consumers' number (minority class) and under-sampling techniques under-samples the benign class number (majority class). To present the data as input to the classifier it should be balanced to avoid skewness and biases towards the majority class. A comparative analysis is presented using the aforementioned data balancing techniques [121]. The CNN model is considered a classifier. For each data type, the CNN is trained in an isolated fashion with different data nature. Various results are obtained after using different data types and the performance of the classifier is monitored over such data. The performance is treated as a comparative analysis and the best performance over the specified data type is chosen as the suitable data balancing technique in the case of the CNN model. To analyze the statistical scenario of the classifier AUC is considered as evaluating metric. AUC is a two-dimensional performance evaluating vector, which highlights true positive rate (TPR) and FPR. Fig 20 shows the statistical analysis of case study 14.

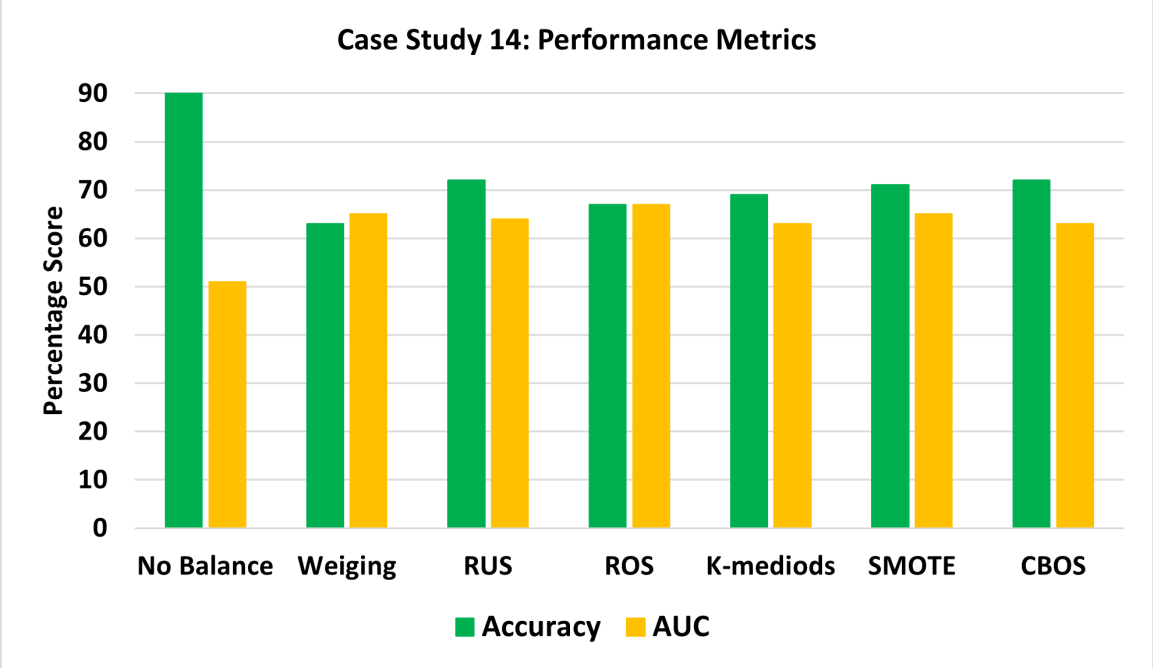

Figure 20: Statistical Analysis of Case Study 14

7.3. Case Study-15

In case study-15, a supervised learning approach is presented [122]. The dataset of SGCC is considered. The data of the SGCC dataset is 24 hours based on sequential data, which is taken as 1D data form [123]. The 1D data are converted into 2D data. The 2D data comprises weekly data on consumers' consumption. To analyze the initial investigation of 2D weekly data, it is plotted and drawn to read its pattern and behavior. Furthermore, the data are preprocessed and normalized using a min-max data normalization scheme [124]. The normalized data are affine and scalable data, however, it is imbalanced in shape. The imbalanced data are augmented using a minority-class data oversampling technique. Adaptive synthesis (ADASYN) technique is used for data augmentation and balancing [125]-[126]. The data augmentation is carried out in order to avoid the issue of class skewness and bias towards the majority class. To extract the abstract features of the study an integrated deep siamese network (DSN) and CNNLSTM are used [127]. The integrated model extracts abstract

features from the 2D data, which eliminates higher data dimensionality issues. Such techniques are applied to avoid excessive operations over the raw and less informative data. Considering such steps, it is reported that the efficiency of the model is increased. The DSN-based integrated model CNN-LSTM is used as a classifier. The CNN extracts vital features of the data using a max pooling strategy. However, LSTM is used to memorize the information for the long term to enhance classification efficiency. To evaluate the performance of the classifier, precision, recall, F1-score, mean average precision (MAP) and AUC are considered as evaluating metrics. The performance of the classifier is compared with LR, SVM, RF [128], WDCNN, and CNNLSTM. The proposed DSN-based CNNLSTM model outperforms the rest of the models. Fig 21 shows the statistical analysis of case study 15.

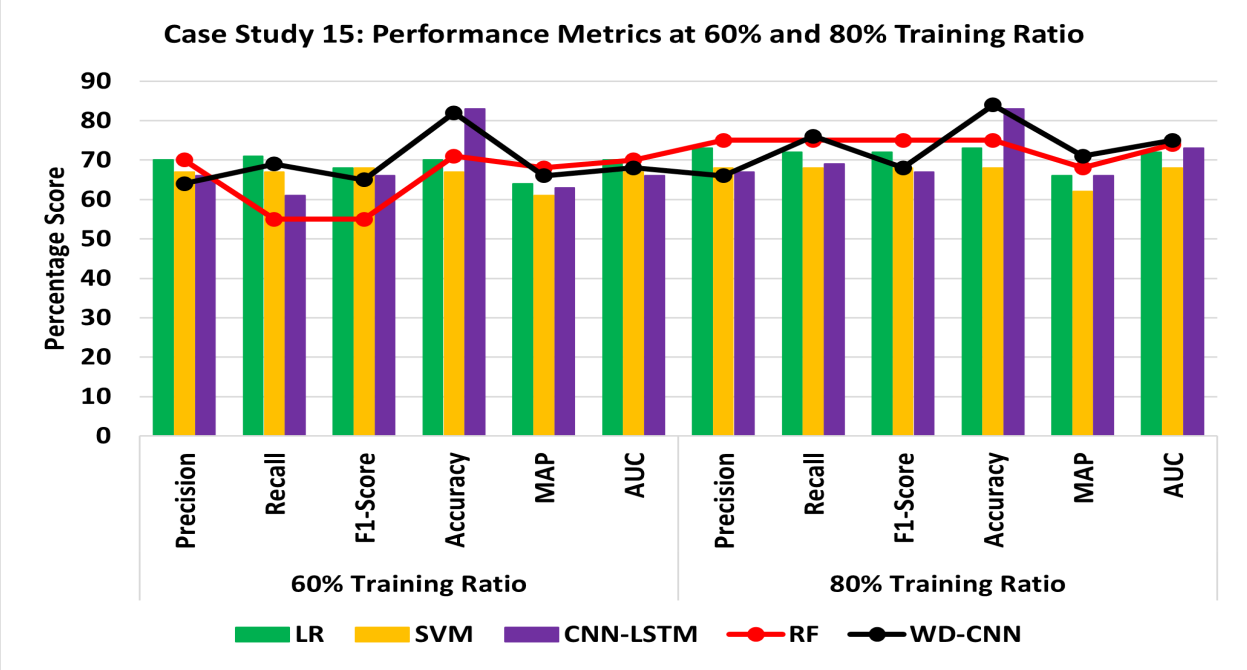

Figure 21: Statistical Analysis of Case Study 15 at different training ratios

7.4. Case Study-16

In case study 16, the SGCC dataset is considered [129]. Initially, the data are preprocessed. Missing values are filled and erroneous values are removed using data interpolation [130]-[131]. The preprocessed data are then normalized using min-max data normalization. Mathematically, the min-max normalization is represented in equation 36.

$$N_L = \frac{L - Min(x)}{Max(x) - Min(x)} \tag{36}$$

Where L is the original value of the electricity consumption and N_L is the normalized value. Afterward, the normalized data are oversampled using SMMOTE as a data oversampling technique. SMOTE considers the minority class data and oversamples it. The oversampling balances both classes' data. The considered data are too large and heavily distributed with a large number of consumers. In order to reduce the dimensions of the data principal component analysis (PCA) is used to extract the abstract information. PCA reduces the dimensionality of the data, which decreases the computational complexity of the model [132]. Furthermore, the AdaBoost-DNN ensemble model is used as a proposed binary classifier. To train the adaboost-DNN model the following steps are considered.

- Initially, the weights of the training samples are initialized.
- The current data with the initialized weights is learned by DNN.
- The classification error rate on the initialized data is calculated.

- Later on, the weights of the training data are updated.
- The weak classifiers are subjected to a weighted linear combination and a strong ensemble classifier is synthesized.

To evaluate the performance of the ensemble classifier, accuracy, TPR, and AUC are considered as evaluating metrics. The analysis is carried out on 60%, 70%, and 80% training data. To compare the performance of the proposed model SVM, artificial neural network (ANN) and RF [133] are considered as base models. Fig 22 shows the statistical analysis of case study 16 at 60% and 80% training ratio.

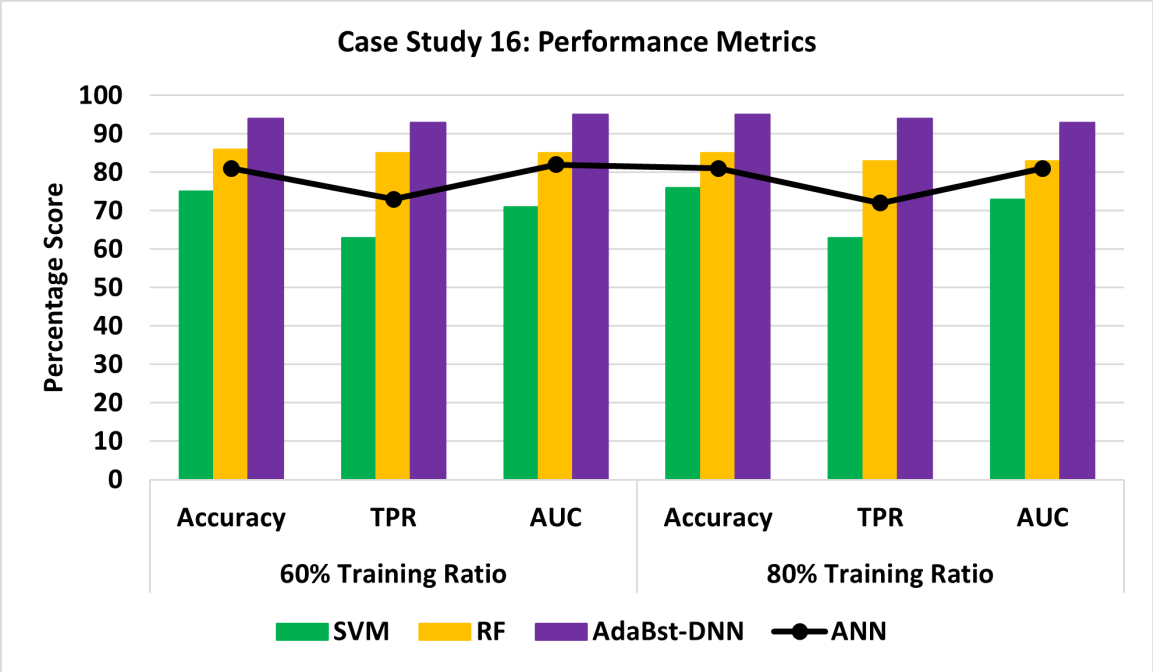

Figure 22: Statistical Analysis of Case Study 16 at different training ratios

7.5. Case Study-17

In case study 17, a novel supervised learning-based approach is presented to investigate NTLs. This study presents a novel approach to fill the null values of the dataset. Contrary to data interpolation mechanisms [134], KNN [135] is used to fill NaNs in consumers' records. To mitigate the class bias of the classifier towards the majority class, the SMOTETomek algorithm is used to balance the dataset. The balanced data are then normalized using min-max data standardization. To extract the most vital features, data reductionality mechanisms are opted. Feature extraction and scalable hypothesis (FRESH) are used to extract the abstract features [136]. The abstract data contains feature vectors of the whole data, which is presented as input data to boost the algorithm. The catboost algorithm is used as a binary classifier. Finally, to interpret the final decision of the classifier, the tree-SHAP algorithm is used as a predictor. To evaluate the performance of the model, precision, recall, F1-score, accuracy, kappa, and MCC are used as evaluating metrics. To monitor the performance of the proposed model boost is compared with XGBoost [137]-[138], lightBoost, ensemble bagging, adaboost, and RF models. Fig 23 shows the statistical analysis of the case study-17.

7.6. Case Study-18

Case study 18 highlights a communication channel-based network to identify fraudulent consumers. It is an integration of fraudulent consumers connected to a communication channel. The network is named fault-tolerant and privacy-preserving electricity theft detection (FPETD). The network is a part of four basic units i-e., data consumer, fog node, cloud server, and main grid. The basic protocol of the network is connection and disconnection-based. When the consumer is connected to the network the operating protocols highlight it as a normal consumer. In case the consumer is not connected to the network, the operating protocols highlight it as

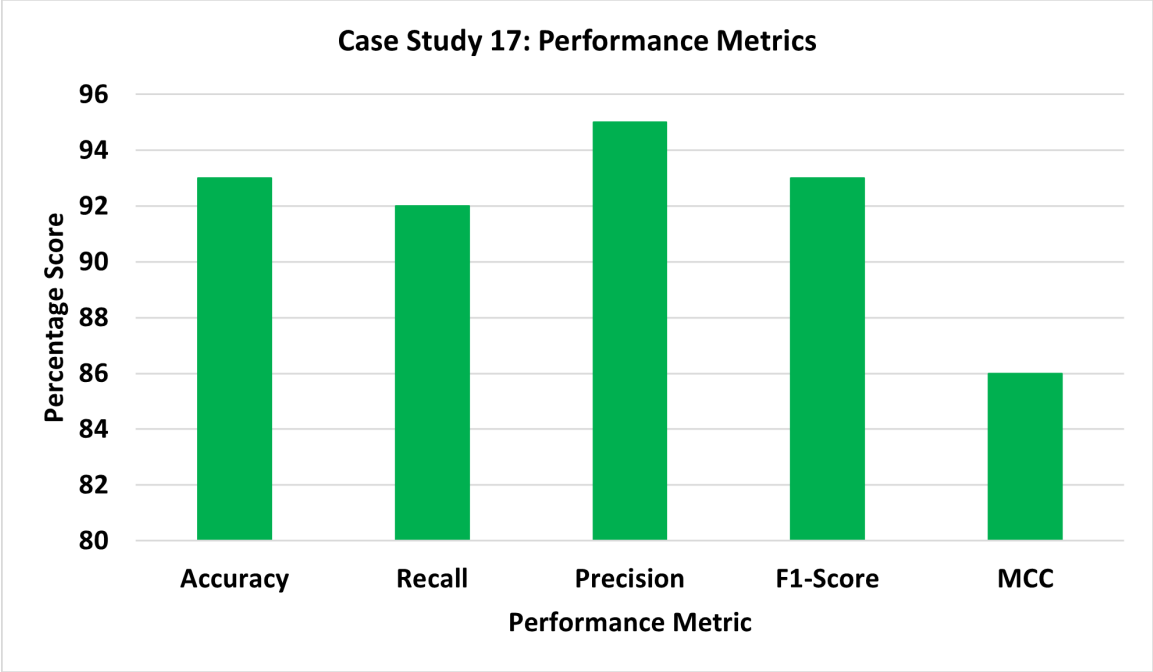

Figure 23: Statistical Analysis of Case Study 17

a malicious one due to its discontinuity. The data information is preserved and no such information is provided to the consumer side to know about their status of continuity and discontinuity. The scenario is time-based and operating protocols update the continuity information from time to time. It may be hourly, daily, weekly, and monthly based. The updating scheme highlights the consumer's continuity and discontinuity to the network. The continuity and discontinuity are then analyzed on the main grid side. A comparative note is generated, which decides the consumer's behavior. It utilizes a CNN model with 2D convolutional and pooling layers. The CNN structure is a 3-stage framework. Initially, the shape of the input data is analyzed. Later on, the max-pooling layers extract the prominent features of the data [139]. In the last stage, the 1D input data shape is featured in 2D input data. The model is trained in forward propagation. A softmax activation function is used to decide the output of the 2D MODEL. The normal participant is categorized as 0 whereas the fraudulent one is labeled as 1. In order to analyze the performance of the model, precision is used as an evaluation metric. The performance of the proposed model is compared with linear SVC, LR [140], RF, and normal CNN. Fig 24 shows the statistical analysis of the case study-18.

7.7. Case Study 19

In case study 19, the dataset of SGCC is considered [141]. It considers the data of 42372 consumers over 1035 days. Initially, a comparative statement is presented to highlight the difference between benign and malicious consumers. The data of the benign consumer are preprocessed where null and void data instances issue is resolved. The erroneous values are removed and null values are filled using mean based strategy. The sequential time series data are sliced into small data groups using a KNN-based approach [142]. The small data are presented in 1D and 2D data shapes. 1D data are a single feature vector of a single consumer over a specific time period. The specified time for 1D data is 24 hours. However, 2D data are data of group consumers. It is a weekly based data over a time period of 7 days. Similarly, monthly data are taken in the same pattern as the 2D data scenario, which is data over the time period of 30 days. The combination of 1D and 2D data over, daily, weekly, and monthly based is named as DWM. The CNN model is used for feature extraction over DWM data. Hence, the feature extractor is named DWMCNN. The CNN extracts daily, weekly, and monthly load features [143]. It utilizes convolutional layers, pooling layers, filters, and fully connected layers. To decide the output, the feature vector sigmoid activation function is used. Furthermore, to compare the performance of the DWMCNN extractor, PCA is used as a base extractor. Both DWMCNN and PCA extract prominent features of the DWM data. However, it is reported that DWMCNN extracts more prominent features as compared to base extractor PCA. Furthermore, to evaluate the binary classification scenario, an RF classifier is used. The

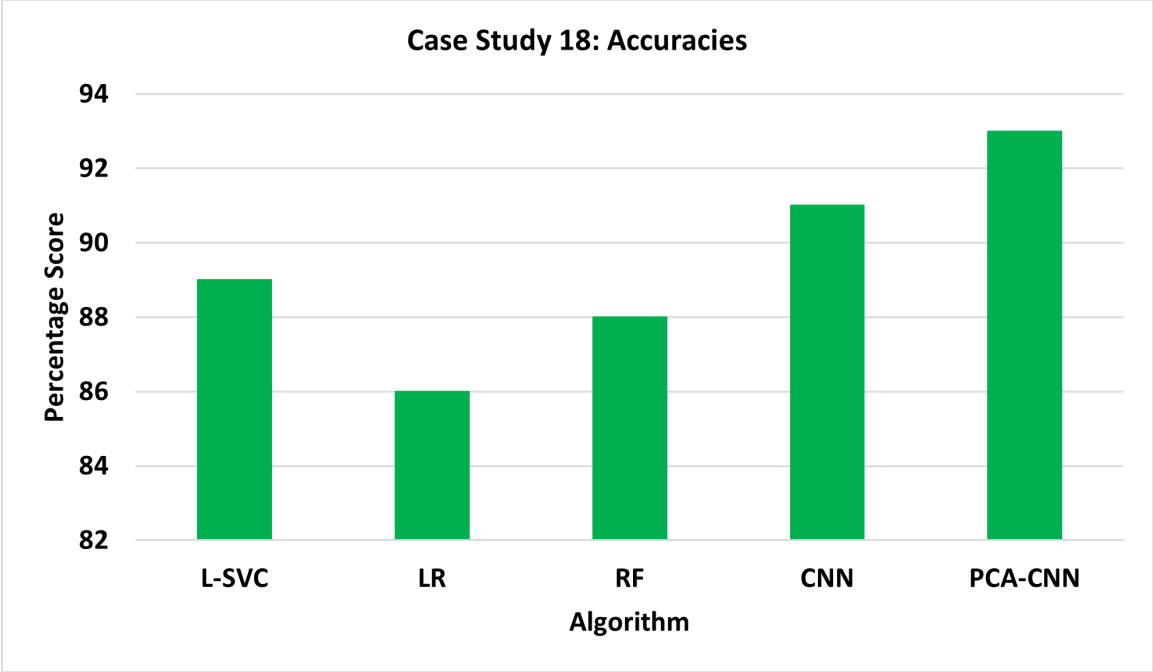

Figure 24: Statistical Analysis of Case Study 18

softmax activation function is used in the RF classifier. The performance of RF is compared with SVM [144], CNN, LR, DWMCNN, DWMCNN-SVM, gradient-enhanced decision tree (GDBT) and DWMCNN-GDBT. To evaluate the statistical performance of classifiers, precision, recall, F1-score, AUC curve, and ROC curve are used as evaluating metrics. Fig 25 shows the statistical analysis of case study 19.

7.8. Case Study-20

In case study 20, the importance of the data-driven approaches and their integrity is highlighted. As data-driven approaches use an intensive sampling rate to identify the nature of the consumed energy [145]. However, it causes many serious issues such as data security and privacy of the consumer. As a consumer, data integrity is the utmost priority basis for the Ups though the detection scenario is nearly impossible without accessing the consumer's data. To access the consumer's data, the granularity rate is the key approach. The data granularity rate is the monitored or recorded energy of SM over a specified period of time. It can be second, half-minute, one-minute, half-hour, hour, and 24-hour based. Exceeding such limits minimizes the chances of a higher detection rate. Analysis and monitoring through using approaches challenge the security issue of the consumer. To tackle such an issue, it is recommended to monitor a 24-hour pattern on the consumer's premises to keep the data integrity and privacy preserved. Monitoring over time period of 24 hours is much enough to analyze the irregularity and non-periodicity of the energy patterns. Basically, consumers follow a trend-based pattern of consumed energy due to their daily routine engagements, which makes it easy to understand the periodicity and non-periodicity in their patterns. To study such a relationship of the patterns [146]-[147], a correlations-based analysis is made. The study proposes a text convolutional-based neural network (text-CNN) to extract the prominent features of 2D data. A dataset of SGCC is used, which is initially preprocessed. The preprocessed data are then augmented and balanced. The features are then extracted using text-CNN. Features of both classes are then compared to highlight the differences [148]. The correlation-based differences highlight the continuity and discontinuity-based functions, which decide the class of the pattern. Later on, a binary classifier is used. To evaluate the performance of the study AUC and ROC are observed.

7.9. Case Study 21

In case study-21, the SGCC dataset is considered and an issue of NTLs is tackled. Initially, a dataset of benign and fraudulent consumers is considered. The fraudulent consumers are fewer in number than the benign consumers. To balance the number of both classes, a data augmentation mechanism is carried out

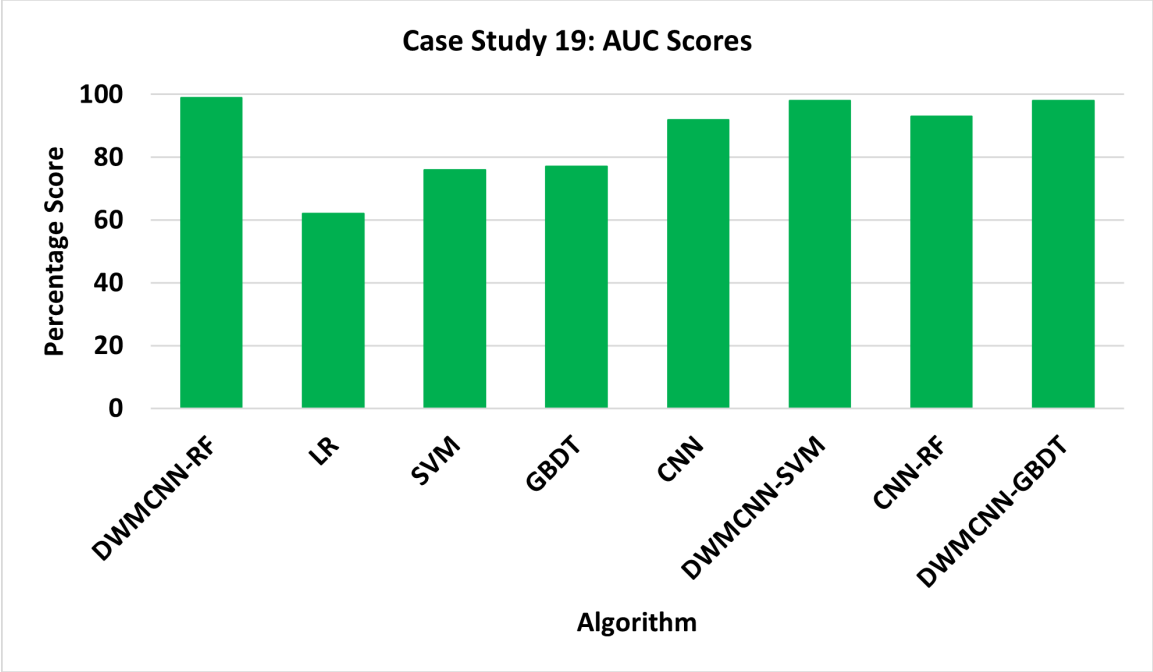

Figure 25: Statistical Analysis of Case Study 19 based on AUC score

using a data oversampling technique SMOTE. The balanced dataset has now a larger number of consumers and the available data are dense. To tackle the data's higher dimensionality issue kernel function-based principal component analysis (KPCA) is used. KPCA extracts the main prominent features of the data and reduces the data dimensions. The extracted data are the abstract data and can be presented for classification. To carry on the binary classification SVM is used as a classifier [149]. The binary classifier labels the fraudulent consumer with 1 and the benign one with 0. To monitor the performance of the proposed classifier, it is compared with the base models. The base models include RF, LSTM [150], CNN, DT, NB, and LR. To analyze the statistical performance of the model, AUC, ROC, precision, recall, and F1-score are used as evaluating metrics.

7.10. Case Study-22

In case study 22, a dataset of SGCC is used. To remove the missing values simple data imputation is opted. The imputed data is then normalized [151]-[152]. The normalized data are augmented using ADASYN. KNN is used to extract the prominent features of the data. The data are then presented for classification. Various features like mean, standard deviation, skewness, peak to peak, crest factor, average frequency, sparseness, waveform length ratio, median absolute deviation, log energy, total harmonic distortion, energy entropy, and spectral decrease are used. Accuracy is used as an evaluation metric to monitor the performance of the case study. The classifier is trained based on 13 13-fold cross-validation techniques. Afterward, fine KNN, medium KNN, coarse KNN, Cosine KNN, fine Tree, medium Tree, coarse Tree, LR, and LD models are used [153]-[154]. To evaluate their detection over the binary class, data fine KNN is reported to achieve the highest accuracy. However, the fine KNN performance is limited as 5% fraudulent and 13% normal consumers remain undetected. To evaluate the performance of the case study only an accuracy metric is used. Fig 26 shows the performance of the case study 22.

8. Case Studies Using NAN-based Detection Approaches

NAN is an integration of observer meters and consumer SMs. The observer meter works as a master meter and monitors the total consumption and is installed at the distribution side of the transformer. Such combinations are named AMI and are used to detect the anomalies. Anomalies are investigated by utilizing observer and consumer SMs data. The data are compared with each other and if any difference is marked the anomalies are then reported. A few of the case studies, which use AMI and NAN-based topology are as follows.

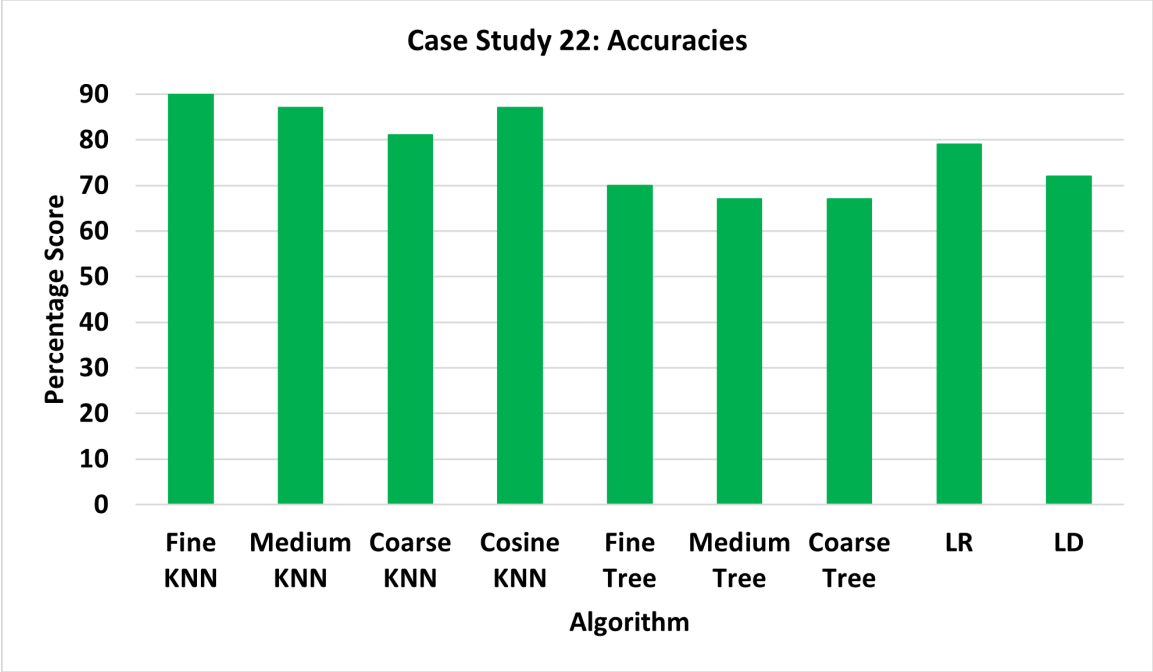

Figure 26: Statistical Analysis of Case Study 22 based on Accuracy

8.1. Case Study-23

In case study-23, a comparative scenario-based novel identification scheme is proposed. It concludes various detection schemes such as ETD identification systems based on HVDC, neural networks, SM topologies, advanced metering infrastructure (AMI) [155], power line communication approach (PLC), and intelligent system (IS). It is reported that the detection efficiency of the SM topologies, PLC, and IS is higher as compared to other aforementioned approaches. This case study introduced AMI-based infrastructure whose system reliability considered is the perfect one. The study develops a neighborhood area network (NAN), which consists of an observer or master meter, the consumer side SM, and operating protocol firmware [156]-[157]. The SM is installed on the consumer premises to record the consumed energy. The observer meter is a surveillance or check meter. Both SM and observer meter data are analyzed and investigated to highlight the traces of anomaly. A comparative statement of both meters is presented over a specific time. If the reading of both meters is the same it is reported as normal consumer. However, if there is any difference between the readings of both meters the consumer is reported as a theft. A threshold of 5% is set for the legitimate and anomalous consumer. The threshold is chosen as 5% due to certain reasons as the difference may be caused due some other issues such as exogenous factors. The exogenous factors are environmental factors, weather factors, and demographic, geographical, and topographical parameters. The cutoff threshold is set, which is based on the network firmware. It can be adjusted to any value. Exceeding the threshold of 5% the consumer is disconnected from the central meter and after the manual inspection, the connection is restored if the consumer is affected due to exogenous factors. However, the consumer is disconnected if there is no legitimate evidence. The disconnected consumer is considered a manipulated one and legal action is taken against such consumers. The proposed approach is a good contribution to the research, however, a physical interruption and investigation are required using such approaches. Moreover, excessive expense is required for the scheduled on-site inspection, which is financially burdensome for UPS.

8.2. Case Study-24

In case study 24, a machine learning-based approach is presented to detect power theft. Matlab simulink is used as a platform to investigate the study. Machine learning-based garra rufa fish (GRF) optimization is used to detect malicious activities. Real-time electrical energy is monitored using an intermediate monitor meter (IMM). Such a metering system avoids manual inspection, which results in excessive costs. The IMM architecture consists of load, SM, distribution station, UP, collector meter, and meter data management system

(MDMS). The MDMS is the basic comparing unit, which receives the SM data of the consumers. The data of the collector meter is represented as 37.

$$C_m = E_{u1} + E_{u2} + E_{u3} + ... + E_{u_N} = \sum_{n=1}^{N} E_{u_N} \tag{37}$$

C_m are the total collector meters' data and E_{u_N} is the consumed energy of the SMs. To accommodate the TLs, the following equation is used 38.

$$C_{mTL\&NTLs} = \sum_{n=1}^{N} E_{u_N} + P_u + Q_u + R_{uNTL} + T_{uTL} \tag{38}$$

Where P_u is the reduced value due to manipulation, n is the connected consumers, Qu is the value from the faulty meter, R_s is the theft value carried out through traditional approaches and Tu is the technical losses. The equation defines certain conditions as if $P_u + R_u > 0$ the data manipulation has been carried out. If $Q_u < 0$ the consumer is treated as a normal one. To develop a hardware-based model, the power on the consumer premises is injected in an analog-to-digital converter (ADC). The output from ADC is provided to MCU. The output power is displayed through the communication unit. GRF-based detection system monitors the consumed energy every minute. Three clusters of consumers are considered. The clusters are named as cluster-1, cluster-2 and cluster-3 [158]-[160]. The clusters have records of 205, 100, and 190, respectively. The cluster is the group profile of consumers. They are analyzed daily. Each cluster identifies ten profiles as bogus ones daily. The statistical standard deviation of all the clusters is monitored as 0.2. Evaluation metrics of accuracy, precision, sensitivity, and specificity are used to analyze the scenario. Fig 27 depicts the statistical performance of case study 24 based on AUC scores and Fig 28 shows the comparison between existing and GRF-based approaches in case study 24.

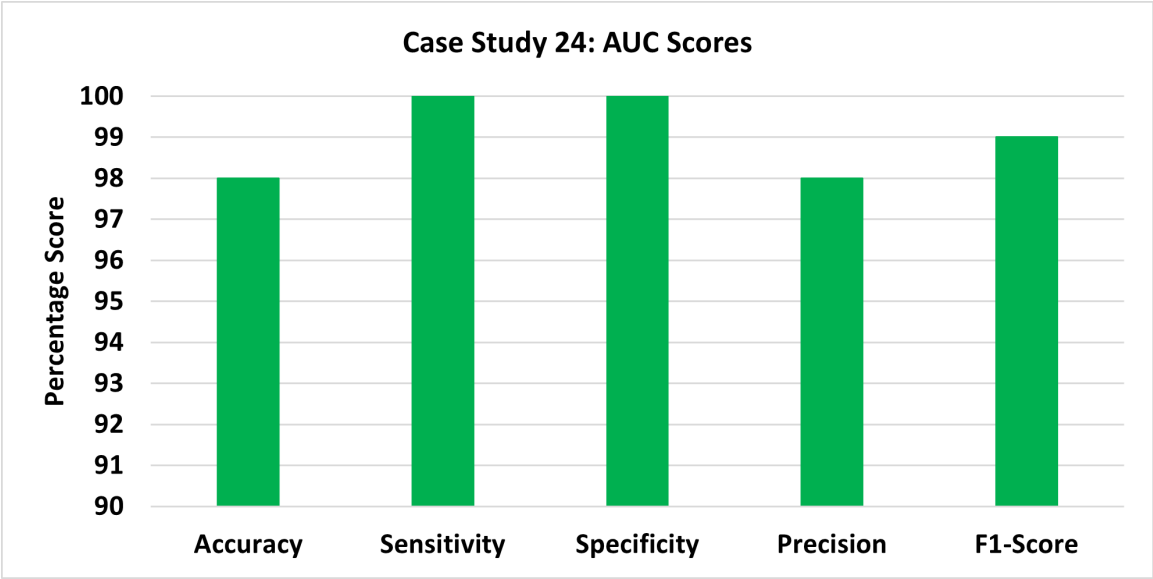

Figure 27: Statistical Analysis of Case Study 24 based on AUC Performance

9. Case Studies Using IoT and hardware-based Approaches

IoT and hardware-based detection schemes are expensive schemes due to their installation, service, and maintenance factors. Many approaches in the literature are presented to investigate NTLs through such schemes, however, they are not yet practical due to the aforementioned reasons. A few of the approaches, which are investigated in our review study are as follows.

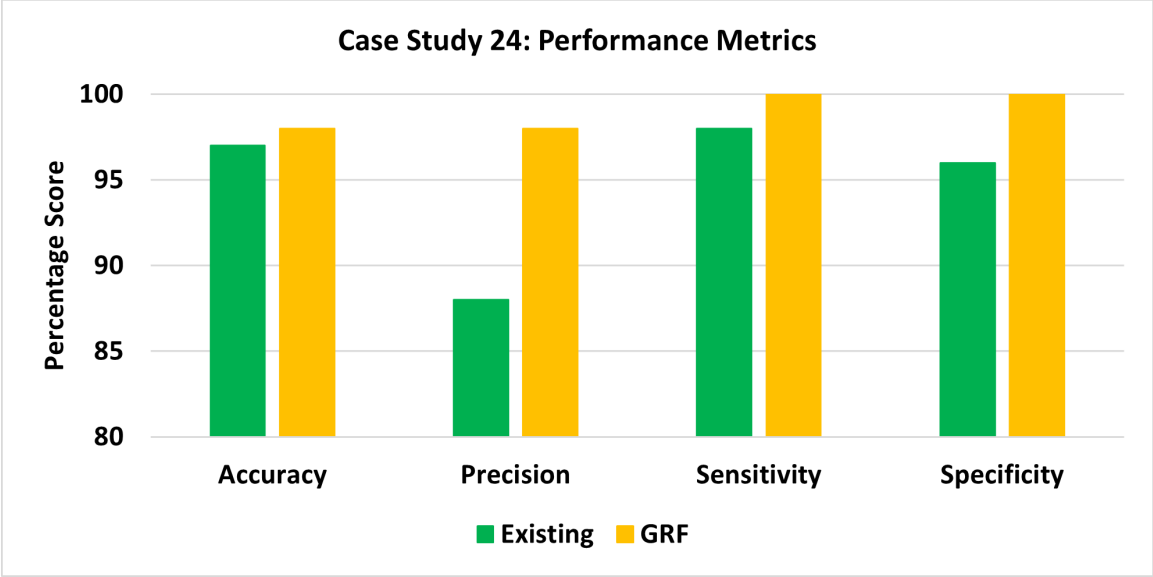

Figure 28: Statistical Analysis of Case Study 24 based on GRF

9.1. Case Study-25

In case study 25, a hardware-based approach is proposed to minimize the NTLs through an advanced billing mechanism. It is an intelligent energy meter (IEM) based hardware scheme, which consists of Arduino (UNO), sim900a GSM module, and other circuit auxiliaries. Initially, the single-phase AC signal is converted to a digital signal (ADC). For the purpose of ADC, output from the current transformer CT and potential transformer (PT) is provided at the full wave bridge rectifier. The rectifier converts the AC signal into a DC signal, which is provided as an input to the Arduino board. The Arduino board converts the DC signal into an 8-bit digital signal. The microcontroller (ATMEGA328P) interacts with the real-time consumption of the consumers. It calculates the consumed energy by monitoring the output signals from CT and PT. An average of 50 samples is received for each of the CT and PT signals at IEM. IEM analyzes the received CT, and PT data and applies two conditions over voltage check. If the voltage is greater than 240V or less than 220V the operating relay is switched ON, which turns ON the alarming buzzer. Operation of the relay from normally closed to normally open position restarts the sample re-collection, which initiates the process from the beginning. After monitoring the voltage level, if the voltage level is within limits it calculates the consumed energy and stores the value in its internal memory. A display screen is mounted, which provides the information to the consumer. A cycle of 24 hours is rotated and an SMS alert is generated. The generated SMS informs the consumer of their consumed units and their price after every 24 hours. The stored information of the consumed energy is transferred through GSM to the data monitoring center. The server monitors the energy consumption record. Furthermore, an advance billing mechanism is developed, which allows the consumer to pay in advance for their utilization [161]. If the consumer exceeds the policy limits and pays no bills for the consumed energy, the supply is automatically cut off. Moreover, it provides a check over meter bypassing [162] and shorting the phase line. Such mechanisms avoid NTLs and theft is reported when an excessive amount of energy is consumed. This case study is a hardware-based approach and no statistical analysis is presented in the proposed study.

9.2. Case Study 26

In case study 26, an IOT-based electricity theft detection system is developed. It is a combination of hardware and software, however, the hardware part is the major one. It consists of an energy meter, voltage sensor, current sensor, microcontroller, cloud interface, and communication channel. Initially, the AC signal is connected to current and voltage sensors. The sensors' signal is provided to the energy meter and the microcontroller is used to monitor the signal. Microcontroller ATMEGA328 is used to monitor the consumed energy. The load, SM, and sensors are considered as nodes. The data of the node is monitored by ATMEGA328 and is stored in the cloud server. The IOT-based approach is adopted. The data are transferred through the WIFI module. A data management center is created, which accesses the consumed data of the consumers. The

center accesses the data and keeps the records of the consumers. To opt for the software module, PROTEUS, and CCS compiler/Arduino are used. PROTEUS is used to interface the ATMEGA328 to various devices for sustainable operation. ATMEGA328 is boost-loaded using Arduino. Furthermore, to integrate the IOT module, the Thingspeak platform is used. It provides data storage space for time series data in the form of sequential data [163]-[165], which is stored in channels. The necessary API is provided by the ESP8266 module. Using such an IOT-based approach the illegal consumers are highlighted without interfacing and intervention of a human. Moreover, the usage pattern and statistical information can be accessed using the proposed topology. Similar to a few case studies, it doesn't provide any statistical analysis regarding NTLs.

9.3. Case Study 27

In case study 27, a hardware GSM-based approach is proposed for the identification of NTLs [166]. It detects theft occurring due to tampering of SMs [167], bypassing and data manipulation [168]. The hardware-based architecture contains two CTs, a microcontroller, an LCD, and a GSM module. PIC18F4520 microcontroller is used with other power operating auxiliaries. The two CTs are installed where one is used to provide the input signal to the microcontroller followed by the rectifiers and filters. PIC18F4520 functions as ADC. The digital signal is generated as output on the consumer's side. The microcontroller is responsible for keeping the balance between the input and output power. It calculates the power at each instance. A GSM-based module transmits the information of the consumed energy to the data center as well as to the consumer. The consumer can switch ON/OFF their consumption according to their needs and requirements. If there is any theft in the system the input and output of the microcontroller don't remain the same. Henceforth the theft is reported. It is able to detect seal tampering of the meter, and bypassing of the meter. The theft occurrence is reported through SMs to the consumer as well as to the data center. It has an automatic operating module to disconnect the theft consumer automatically. The module is operated through a relay where both the input and output data of the microcontroller are monitored. The difference between the input and output data of the microcontroller is treated as theft and relays cut off the supply to the consumers' premises. It is a reliable and accurate module using a wireless architecture to report theft cases. However, each consumer has different socio-demographical variables, which can interrupt such detection schemes.

10. Performance Metrics

To evaluate the performance of various algorithms reliable parameters are required to analyze the detection scenarios. The detection is based on the identification of benign and fraudulent consumers. Fraudulent consumer freaks the distribution networks through various data manipulating schemes. It is a requirement to have a reliable and efficient detector. The ETD classification is a binary scenario so binary classifiers are required for optimal performance. To evaluate the performance of various algorithms, different evaluating metrics are used such as precision, recall, accuracy, AUC, ROC, F1-Score, MCC, PRC, TPR, FPR, specificity, etc, All the metrics are based on major four parameters i-e FP, TP, FN, and TN. If a negative sample is considered as positive it refers to FP. Similarly, when a false positive sample is considered as negative it is said as FN. Contrary to FP and FN, TP and TN are the scenarios for true positive as positive and true negative as negative, respectively. To show the aforementioned metrics, mathematical representations are presented in equations 39-45.

$$DR = \frac{TP}{TP + FN} \tag{39}$$

$$FPR = \frac{TN}{TN + FN} \tag{40}$$

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{41}$$

$$Precision = \frac{TP}{TP + FP} \tag{42}$$

$$Recall = \frac{TP}{TP + FN} \tag{43}$$

$$F1 - Score = \frac{(Precision)(Recall)}{Precision + Recall} \tag{44}$$

$$Specificity = \frac{TN}{TN + FP} \tag{45}$$

11. Analysis and Investigation

Tables 3 and 3 (Continue) show the identified problems and their proposed solutions. Data augmentation, high FPR, and non-availability of the prominent features are the major problems identified in many cases. Data augmentation is an imbalanced data-related issue whereas high FPR is related to low detection and false identification rate of the classifier. Non-availability of the theft class data means that all the datasets majorly contain a large number of benign consumers as their data is obtained through willing consent. It is hard to obtain data on fraudulent consumers as no intentional involvement is subjected to such surveys. Theft class data is synthesized using various data manipulating or augmentation schemes. Furthermore, the high data dimensionality issue is another concern, which is tackled through feature engineering or extraction. Feature extraction is carried out using various algorithms. Similarly, non-consideration of exogenous factors is one of the major issues. It is based on the transitions in the data. The exogenous factors are those factors, which result in unusual patterns of consumption. Though exogenous factors are normal and natural factors, their impact on the consumption data raises the caution of maliciousness. Moreover, so many problems are identified and listed in Table 3, provides a gateway for researchers, readers, and developers to consider the aforementioned issues. Table 4 is another aspect of the research contribution of the presented study. It identifies all the evaluating metrics used in various case scenarios. The evaluating metrics monitor the performance of algorithms. All the metrics are of equal importance and none can be ignored as every parameter of these metrics has its unique contribution to the classification scenario. The symbolic representation is marked against each case study, which shows the use of the specific metrics in the perspective case study. The ✓ shows that the aforementioned metrics are used in that specific study and X shows that no such performance metrics are used. It is informative to use the maximum amount of metrics to examine the depth of the study, which provides the reliability and efficiency-based analysis of algorithms. The mentioned case studies information is enough for researchers, readers, and developers to evaluate any study based on these metrics.

Table 3: Review Analysis of the identified Problems and Proposed Solutions

References	Problems Identified	Solutions Proposed
[169]	• Cross Pairs Identification • High FPR • On-site Verification	• Tomek Links Algorithm • Hybrid BiGRU-BiLSTM • Reliable Classification
[170]	• Non Availabilty of Theft Class Data • Less Prominent Features • High FPR	• Theft Cases • Stochastical Feature Engineering • Hybrid Model
[171]	• Higher Data Dimensionality Issue • Lack of Theft Class Data • High FPR	• Feature Engineering • Novel FDIs • AttenLSTMInc
[172]	• Imbalanced Data • Overlapped Synthetic Data • High FPR • Skewness Factors	• Data Augmentation • FDIs • AttenLSTMInc • MCC
[173]	• Non-Sequential Data Issue • Imbalance Data • High FPR	• MLP • Data Augmentation • MLP-GRU
[174]	• Biasness Towards Majority Class • High Data Dimensionality • Non Availability of Prominent Feature	• SAGAN • ERNET • SAE
[175]	• Non-Sequential Data Issue • Imbalance Data • High FPR	• CNNGBDT • Data Augmentation • GBDT
[176]	• Class Biasness Issue • Lack of Abstract Features • Imbalance Data	• Oversampling • CVAE • Data Augmentation
[177]	• High FPR • Imbalance Data • Model's Skewness	• CNN • SMOTE and Clusters based Segmentation • Variational Data Approach
[178]	• Imbalance Data • Non-Availability of Prominent Features • Class Skewness	• Data Augmentation • Conversion into 2D data • DSN based CNNLSTM
[179]	• High Data Dimensionality Issue • Imbalance Data • High FPR	• PCA • SMOTE • ADABoost-DNN

Table 3 (Continue): Review Analysis of the identified Problems and Proposed Solutions

References	Problems Identified	Solutions Proposed
[180]	• Non-Availability of Theft Class Data • Model's Skewness • Low Detection Rate	• Cyber Attacks • Data Augmentation • AdaBoost Classifier
[181]	• Imbalance Data Issue • Traditional Data Interpolation • Non-Availability of Prominent Features	• SMOTE-TOMEK • KNN Cluster based Interpolation • FRESH
[182]	• High FPR in Soft Models • Excessive Operational Time • Data Manipulation Issue	• FPETD
[183]	• Non-Availability of Featured Data • High Data Dimensionality Issue • High FPR	• 2D Data Conversion • CNN • DWMCNN-GDBT
[184]	• High Data Dimensionality Issue • Pattern Recognition • High FPR	• text-CNN • Correlation Analysis • text-CNN
[185]	• On-Site Verification • Excessive Data Manipulation • Changes due to Exogenous	• Hardware based Design • Threshold to Tackle the Variations
[186]	• High Data Dimensionality Issue • High FPR • Imbalance Data	• KPCA • SVM • Data Augmentation
[187]	• Data Manipulation • Human Interface • High FPR	• Hardware based Design
[188]	• Non Consideration of Sequential Data • Physical Tampering • Occusional Manipulation	• SEAI Dataset • LR-ETDM • CV-LRETDM
[189]	• Data Manipulation • High FPR • Exogenous Factors	• Hardware based Design • IOT Based Data Transferring

Table 3 (Continue): Review Analysis of the identified Problems and Proposed Solutions

References	Problems Identified	Solutions Proposed
[190]	• Non-Availability of Abstract Features • Lack of Probabilistic Approaches • High FPR	• Genetic Programming • FMMs • Predictive Analysis using GBM
[191]	• High FPR • Human Interface • Axogenous Variables	• GSM based Hardware • Data Balancing • Real Time Data Analysis
[192]	• Non Reliable Models • Human Interface Error • Axogenuous Variable	• GSm based Hardware Detection Scheme • Real Time Data Analysis
[193]	• Data manipulation in DER Generation • Manipulation of Neighborhood's Data • High FPR	• ESM and ISM Interface • Data Balancing Attacks
[194]	• Imbalance Data Issue • Model's Skewness • High FPR	• ADASYN • Fine KNN • Data Augmentation
[195]	• Variational Attacks • Change and Transit in Continuity • Less Reliable Detectors	• AMI • CAT • Hybrid Model

Table 4: Review Analysis and Literature Comparison

Note: Case Study (CS), Feature Extractor (F-Exe), Feature Engineering (F-Eng)

Case Study (CS)	Preci.	Recall	Accuracy	AUC	ROC	PRC	MCC	F-Ext	F-Eng	TPR	FPR	F1-Score	C Analysis	Hard Scheme
CS 1	X	X	X	✓	X	✓	X	X	X	X	✓	X	X	X
CS 2	X	X	✓	✓	X	X	X	X	X	✓	✓	X	X	X
Case Study 3	✓	✓	✓	✓	X	X	X	X	X	X	✓	✓	X	X
CS 4	✓	✓	✓	✓	X	X	✓	X	✓	X	✓	✓	X	X
CS 5	✓	✓	✓	✓	X	X	X	X	X	X	✓	✓	X	X
CS 6	✓	X	✓	X	X	X	X	X	X	X	X	✓	X	X
CS 7	X	X	✓	X	X	X	X	X	X	X	X	X	✓	X
CS 8	X	X	X	X	X	X	X	X	X	X	X	X	✓	X
CS 9	✓	✓	✓	X	X	X	X	X	X	X	X	X	X	X
CS 10	✓	✓	✓	✓	X	X	X	X	X	X	✓	✓	X	X
CS 11	✓	X	✓	✓	X	✓	X	✓	✓	X	✓	✓	X	X
CS 12	✓	✓	✓	X	X	X	X	✓	X	X	✓	X	X	X
CS 13	✓	✓	✓	✓	X	X	X	✓	X	✓	✓	✓	X	X
CS 14	✓	✓	✓	✓	X	X	X	X	X	X	✓	X	X	X
CS 15	✓	✓	✓	✓	X	X	X	X	X	X	✓	✓	X	X
CS 16	X	X	✓	✓	X	X	X	✓	X	✓	✓	X	X	X
CS 17	✓	✓	✓	X	X	X	✓	X	X	X	X	✓	X	X
CS 18	X	X	✓	X	X	X	X	✓	X	X	X	X	X	X
CS 19	✓	✓	✓	X	X	X	X	X	X	✓	✓	✓	X	X
CS 20	X	X	X	✓	✓	X	X	X	X	X	X	X	X	X
CS 21	✓	✓	✓	X	X	X	X	✓	✓	X	✓	✓	X	X
CS 22	X	X	✓	X	X	X	X	X	X	X	X	X	X	X
CS 23	X	X	X	X	X	X	X	X	X	X	X	X	✓	X
CS 24	✓	✓	✓	X	X	X	X	✓	✓	X	X	✓	✓	X
CS 25	X	X	X	X	X	X	X	X	X	X	X	X	✓	✓
CS 26	X	X	X	X	X	X	X	X	X	X	X	X	✓	✓
CS 27	X	X	X	X	X	X	X	X	X	X	X	X	✓	✓
This Review	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

11.1. Statistical Investigation

In classification problems where one class is highly dominant, the accuracy metrics are sometimes misleading, hence a more reliable evaluation metric is necessary. From the open literature, it is concluded that the F1 score can serve as a reliable evaluation metric for such problems. Hence in this study, we have used the F1 score to reflect the true performance of selected case studies. The F1 score combines TP, FN, and FP to precisely define the model's performance and is given in eq. (44): A few case studies (CS7, CS8, CS10, CS13, CS20, CS21, CS23, CS25, CS26, and CS27) are not concluded due to a lack of related information for the aforementioned performance metric F1-score. The mentioned case studies are either comparative analysis, IOT-based approaches, or hardware designs for detection scenarios. All other case studies are mentioned and analyzed here. Fig. 29 shows the statistical analysis, where CS19 bears the highest F1 score. The reliable value of F1-score highlights the importance and efficiency of the classifiers in heavily imbalanced data scenarios. Furthermore, CS3, CS7, and CS15 show the second-highest efficient F1 score, which majorly highlights the importance of the F1 score in classification. In the case of ETD, the benign and fraudulent consumers are well versed in classified and no other such reliable approaches are yet available to compare. The hardware and IoT-based investigations are injected into literature streams for the detection of NTLs, however, no such optimal results are yet concluded.

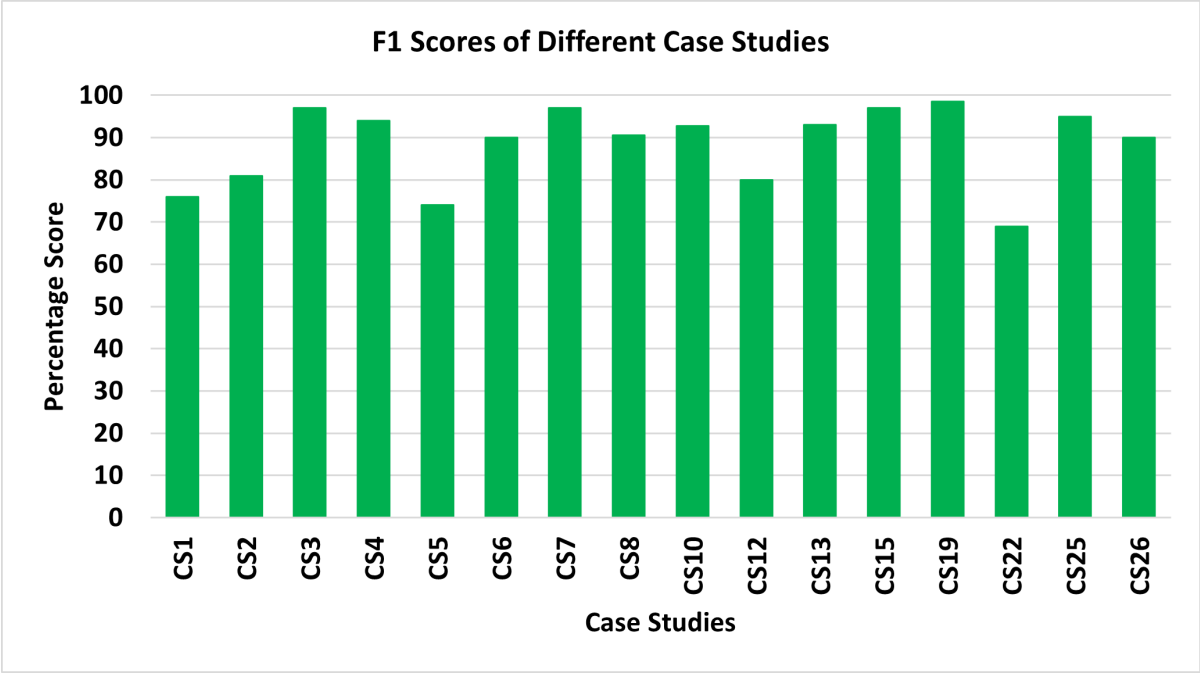

Figure 29: Statistical Investigation of All Studies Using Performance Metric F1-Score. *CS: (Case Study)

12. Conclusion

NTLs are a serious threat to developing countries due to the huge financial stress on UPs, which tends to defuse their optimal operations. This article presents various data-driven, IOT, and hardware-based approaches to investigate NTLs. Many case studies have been presented regarding the aforementioned approaches, however, data-driven approaches are the commonly used and prominent solutions due to their low cost and efficient results. The major focus of this review study is on data-driven approaches, which investigate many solutions for the detection of NTLs. All the identified problems and the proposed solutions of the case studies have been identified and listed properly, which contributes to the literature. Each case study's statistical analysis is presented to provide a platform for researchers and developers to investigate the impact of a detailed analysis of the literature regarding ETD. Additionally, the evaluating metrics of various cases are highlighted and listed for comparative investigation. It helps readers to provide a specific spectrum of evaluating metrics for the improved classification scenario. Furthermore, data manipulating approaches are identified and listed to provide a pathway

for researchers for provision of the synthetic fraudulent consumers' data. Moreover, the comparative analysis highlights the use of various mechanisms, methodologies, hybrid models, IoT, and hardware-based solutions, which is helpful for future research activities to synthesize optimal and effective solutions for the identification of NTLs.

13. Future Recommendations

Some probable future research directions are listed below.

- The study can be utilized to use its mentioned and identified approaches to investigate classification scenarios other than ETD.
- Theft cases and FDIs can be utilized and explored further to investigate novel data manipulating techniques.
- Researchers can opt for any of the highlighted problems and novel solutions can be proposed.
- One can investigate various hybrid models and can utilize them in applications other than ETD.
- Hardware and software approaches have been defined here to investigate ETD. The same can be hybridized and novel approaches can be identified and contributed.
- Identification of the attention layer provides other ways to adapt it for feature engineering and extraction purposes, which could be a good research contribution.
- Highlighting the memorization of LSTM and the sliding window concept can open other novel methodologies for the investigation of various analyses.
- The aforementioned algorithms can be utilized in medical applications to detect dangerous diseases like cancer etc,
- The same approaches can be utilized in the prediction of stock exchange status in real-time.
- Furthermore, it can be integrated with strategic studies for arm detection purposes.

ADABOOST	Adaptive Boosting Machine
ADASYN	Adaptive Synthetic
ADASYNENN	Adaptive Synthetic Edited Nearest Neighbor
ADC	Analogue to Digital Converter
ANN	Artificial neural Network
AMI	Advanced Metering Infrastructure
AUC	Area Under the Curve
CBT	Consumer Behavior Trials
CNN	Convolutional Neural Network
CS	Cloud Server
CVAE	Conditional Variational Auto Encoder
CV-ETDM	Categorical Variable Energy Theft & Defective SM
DER	Distributed Energy Resources
DSN	Deep Siemese Network
DT	Decision Tree
ESM	Export Smart Meter
FDI	False Data Injection
FMMs	Finite Mixture Models
FPR	False Positive Rate
FRESH	Feature Extraction and Scalable Hypothesis
FPETD	Privacy Preserving Electricity Theft Detection
GRU	Gated Recurrent Units
GBDT	Gradient Boosting Decision Tree
ISM	Import Smart Meter
KNN	K-Nearest Neighbor
KPCA	Kernal Function Principal Component Analysis
LD	Linear Descriminent
LLE	Linear Embedding
LR	Linear Regression
LSTM	Long Short Term Memory
MCC	Matthews Coefficient Correlation
MDMS	Meter Data Management System
MLP	Multi Layer Perceptron
NAN	Neighborhood Area Network
NB	Naive Bayes
NN	Neural Network
NTLs	Non Technical Losses
PCA	Principal Component Analysis
RF	Random Forest
REDD	Residual Energy Disaggregation Dataset
RMSProp	Root Mean Square Propagation
ROC	Receiver Operating Curve
SAGAN	Self Attention Generative Adverserial Network
SEAI	Sustainable Energy Authority Ireland
SGCC	State Grid Corporation of China
SM	Smart Meter
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TLs	Technical Losses
TPR	True Positive Rate
UP	Utility Provider
WDCNN	Wide and Deep Convolutional network
1-D	One Dimensional
2-D	Two Dimensional

References

[1] Bin-Halabi, A., Nouh, A. and Abouelela, M., 2019. Remote detection and identification of illegal consumers in power grids. IEEE Access, 7, pp.71529-71540.

[2] Polgári, B. and Raisz, D., 2012, September. Application of smart meters especially for the detection of illegal electricity usage. In Proceedings of 7th international conference on deregulated electricity market issues in South-Eastern Europe (DEMSEE 2012), Nicosia (pp. 1-5).

[3] Runnian, W., Zhishang, D., Xiang, G. and Jianwen, Z., 2022, September. Application Research of Computer Machine Learning Algorithm in Anti-theft Analysis under Smart Grid. In 2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE) (pp. 817-821). IEEE.

[4] Zheng, Z., Yang, Y., Niu, X., Dai, H.N. and Zhou, Y., 2017. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. IEEE Transactions on Industrial Informatics, 14(4), pp.1606-1615.

[5] Ahmad, T., Chen, H., Wang, J. and Guo, Y., 2018. Review of various modeling techniques for the detection of electricity theft in smart grid environment. Renewable and Sustainable Energy Reviews, 82, pp.2916-2933.

[6] Akram, R., Ayub, N., Khan, I., Albogamy, F.R., Rukh, G., Khan, S., Shiraz, M. and Rizwan, K., 2021. Towards big data electricity theft detection based on improved rusboost classifiers in smart grid. Energies, 14(23), p.8029.

[7] Haq, E.U., Pei, C., Zhang, R., Jianjun, H. and Ahmad, F., 2023. Electricity-theft detection for smart grid security using smart meter data: A deep-CNN based approach. Energy Reports, 9, pp.634-643.

[8] Smith, T.B., 2004. Electricity theft: a comparative analysis. Energy policy, 32(18), pp.2067-2076.

[9] Haq, E.U., Huang, J., Xu, H., Li, K. and Ahmad, F., 2021. A hybrid approach based on deep learning and support vector machine for the detection of electricity theft in power grids. Energy Reports, 7, pp.349-356.

[10] Huang, S.C., Lo, Y.L. and Lu, C.N., 2013. Non-technical loss detection using state estimation and analysis of variance. IEEE Transactions on Power Systems, 28(3), pp.2959-2966.

[11] Navani, J.P., Sharma, N.K. and Sapra, S., 2012. Technical and non-technical losses in power system and its economic consequence in Indian economy. International journal of electronics and computer science engineering, 1(2), pp.757-761.

[12] World Bank, 2003. World development report 2004: making services work for poor people. The World Bank.

[13] Gaur, V. and Gupta, E., 2016. The determinants of electricity theft: An empirical analysis of Indian states. Energy Policy, 93, pp.127-136.

[14] Bhatti, S.S., Lodhi, M.U.U., ul Haq, S., Gardezi, S.N.M., Javaid, E.M.A., Raza, M.Z. and Lodhi, M.I.U., 2015. Electric power transmission and distribution losses overview and minimization in Pakistan. International Journal of Scientific & Engineering Research, 6(4), pp.1108-1112.

[15] Buzau, M.M., Tejedor-Aguilera, J., Cruz-Romero, P. and Gomez-Exposito, A., 2019. Hybrid deep neural networks for detection of non-technical losses in electricity smart meters. IEEE Transactions on Power Systems, 35(2), pp.1254-1263.

[16] Jiang, R., Lu, R., Wang, Y., Luo, J., Shen, C. and Shen, X., 2014. Energy-theft detection issues for advanced metering infrastructure in smart grid. Tsinghua Science and Technology, 19(2), pp.105-120.

[17] Dangar, B. and Joshi, S.K., 2015, March. Notice of Violation of IEEE Publication Principles: Electricity theft detection techniques for metered power consumer in GUVNL, GUJARAT, INDIA. In 2015 Clemson University Power Systems Conference (PSC) (pp. 1-6). IEEE.

[18] Rengaraju, P., Pandian, S.R. and Lung, C.H., 2014, May. Communication networks and non-technical energy loss control system for smart grid networks. In 2014 IEEE Innovative Smart Grid Technologies-Asia (ISGT ASIA) (pp. 418-423). IEEE.

[19] Depuru, S.S.S.R., Wang, L. and Devabhaktuni, V., 2011. Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft. Energy policy, 39(2), pp.1007-1015.

[20] Jokar, P., Arianpoo, N. and Leung, V.C., 2015. Electricity theft detection in AMI using customers' consumption patterns. IEEE Transactions on Smart Grid, 7(1), pp.216-226.

[21] Huang, Y. and Xu, Q., 2021. Electricity theft detection based on stacked sparse denoising autoencoder. International Journal of Electrical Power & Energy Systems, 125, p.106448.

[22] Park, C.H. and Kim, T., 2020. Energy theft detection in advanced metering infrastructure based on anomaly pattern detection. Energies, 13(15), p.3832.

[23] Fenza, G., Gallo, M. and Loia, V., 2019. Drift-aware methodology for anomaly detection in smart grid. IEEE Access, 7, pp.9645-9657.

[24] Bohani, F.A., Suliman, A., Saripuddin, M., Sameon, S.S., Md Salleh, N.S. and Nazeri, S., 2021. A comprehensive analysis of supervised learning techniques for electricity theft detection. Journal of Electrical and Computer Engineering, 2021.

[25] Qu, Z., Li, H., Wang, Y., Zhang, J., Abu-Siada, A. and Yao, Y., 2020. Detection of electricity theft behavior based on improved synthetic minority oversampling technique and random forest classifier. Energies, 13(8), p.2039.

[26] García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J.M. and Herrera, F., 2016. Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), pp.1-22.

[27] Duarte Soares, L., de Souza Queiroz, A., López, G.P., Carreño-Franco, E.M., López-Lezama, J.M. and Muñoz-Galeano, N., 2022. BiGRU-CNN neural network applied to electric energy theft detection. Electronics, 11(5), p.693.

[28] Yao, R., Wang, N., Ke, W., Chen, P. and Sheng, X., 2023. Electricity Theft Detection in unbalanced sample distribution: A novel approach including a mechanism of sample augmentation. Applied Intelligence, 53(9), pp.11162-11181.

[29] Singh, S.K., Bose, R. and Joshi, A., 2017, December. PCA based electricity theft detection in advanced metering infrastructure. In 2017 7th international conference on power systems (ICPS) (pp. 441-445). IEEE.

[30] Liang, Q., Zhao, S., Zhang, J. and Deng, H., 2023. Unsupervised BLSTM Based Electricity Theft Detection with Training Data Contaminated. ACM Transactions on Cyber-Physical Systems.

[31] Kong, X., Zhao, X., Liu, C., Li, Q., Dong, D. and Li, Y., 2021. Electricity theft detection in low-voltage stations based on similarity measure and DT-KSVM. International Journal of Electrical Power & Energy Systems, 125, p.106544.

[32] Hussain, S., Mustafa, M.W., Jumani, T.A., Baloch, S.K., Alotaibi, H., Khan, I. and Khan, A., 2021. A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection. Energy Reports, 7, pp.4425-4436.

[33] Mutupe, R.M., Osuri, S.O., Lencwe, M.J. and Chowdhury, S.D., 2017, June. Electricity theft detection system with RF communication between distribution and customer usage. In 2017 IEEE PES PowerAfrica (pp. 566-572). IEEE.

[34] Mishra, P., Biancolillo, A., Roger, J.M., Marini, F. and Rutledge, D.N., 2020. New data preprocessing trends based on ensemble of multiple preprocessing techniques. TrAC Trends in Analytical Chemistry, 132, p.116045.

[35] Li, J., Liao, W., Yang, R. and Chen, Z., 2021, October. A Data Augmentation Method for Distributed Photovoltaic Electricity Theft Using Wasserstein Generative Adversarial Network. In 2021 IEEE 5th Conference on Energy Internet and Energy System Integration (EI2) (pp. 3132-3137). IEEE.

[36] Niu, Z., Zhong, G. and Yu, H., 2021. A review on the attention mechanism of deep learning. Neurocomputing, 452, pp.48-62.

[37] Sun, Y. and Gu, L., 2021, March. Attention-based machine learning model for smart contract vulnerability detection. In Journal of physics: conference series (Vol. 1820, No. 1, p. 012004). IOP Publishing.

[38] Vakanski, A., Xian, M. and Freer, P.E., 2020. Attention-enriched deep learning model for breast tumor segmentation in ultrasound images. Ultrasound in medicine & biology, 46(10), pp.2819-2833.

[39] Cahall, D.E., Rasool, G., Bouaynaya, N.C. and Fathallah-Shaykh, H.M., 2019. Inception modules enhance brain tumor segmentation. Frontiers in computational neuroscience, 13, p.44.

[40] Chudzik, P., Majumdar, S., Caliva, F., Al-Diri, B. and Hunter, A., 2018, March. Exudate segmentation using fully convolutional neural networks and inception modules. In Medical Imaging 2018: Image Processing (Vol. 10574, pp. 785-792). SPIE.

[41] Lebailly, T., Kiciroglu, S., Salzmann, M., Fua, P. and Wang, W., 2020. Motion prediction using temporal inception module. In Proceedings of the Asian conference on computer vision.

[42] Jiménez-Valverde, A., 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecology and Biogeography, 21(4), pp.498-507.

[43] Bouckaert, R.R., 2006, December. Efficient AUC learning curve calculation. In Australasian joint conference on artificial intelligence (pp. 181-191). Berlin, Heidelberg: Springer Berlin Heidelberg.

[44] Helwan, A. and Uzun Ozsahin, D., 2017. Sliding window based machine learning system for the left ventricle localization in MR cardiac images. Applied Computational Intelligence and Soft Computing, 2017, pp.1-9.

[45] Tanaka, T., Nambu, I., Maruyama, Y. and Wada, Y., 2022. Sliding-Window Normalization to Improve the Performance of Machine-Learning Models for Real-Time Motion Prediction Using Electromyography. Sensors, 22(13), p.5005.

[46] Desale, K.S. and Shinde, S.V., 2023. Concept drift detection and adaption framework using optimized deep learning and adaptive sliding window approach. Expert Systems, 40(9), p.e13394.

[47] Khattak, A., Bukhsh, R., Aslam, S., Yafoz, A., Alghushairy, O. and Alsini, R., 2022. A Hybrid Deep Learning-Based Model for Detection of Electricity Losses Using Big Data in Power Systems. Sustainability, 14(20), p.13627.

[48] Chen, Y., Chang, R. and Guo, J., 2021. Effects of data augmentation method borderline-SMOTE on emotion recognition of EEG signals based on convolutional neural network. IEEE Access, 9, pp.47491-47502.

[49] Lee, T., Kim, M. and Kim, S.P., 2020, February. Data augmentation effects using borderline-SMOTE on classification of a P300-based BCI. In 2020 8th International Winter Conference on Brain-Computer Interface (BCI) (pp. 1-4). IEEE.

[50] Sun, X., Hu, J., Zhang, Z., Cao, D., Huang, Q., Chen, Z. and Hu, W., 2023. Electricity Theft Detection Method Based on Ensemble Learning and Prototype Learning. Journal of Modern Power Systems and Clean Energy.

[51] Banga, A., Ahuja, R. and Sharma, S.C., 2021. Accurate detection of electricity theft using classification algorithms and Internet of Things in smart grid. Arabian Journal for Science and Engineering, pp.1-17.

[52] Appiah, S.Y., Akowuah, E.K., Ikpo, V.C. and Dede, A., 2023. Extremely randomised trees machine learning model for electricity theft detection. Machine Learning with Applications, 12, p.100458.

[53] Shi, J., Gao, Y., Gu, D., Li, Y. and Chen, K., 2023. A novel approach to detect electricity theft based on conv-attentional Transformer Neural Network. International Journal of Electrical Power & Energy Systems, 145, p.108642.

[54] Wang, Y., Jin, S. and Cheng, M., 2023. A Convolution–Non-Convolution Parallel Deep Network for Electricity Theft Detection. Sustainability, 15(13), p.10127.

[55] Nawaz, A., Ali, T., Mustafa, G., Rehman, S.U. and Rashid, M.R., 2023. A novel technique for detecting electricity theft in secure smart grids using CNN and XG-boost. Intelligent Systems with Applications, 17, p.200168.

[56] Nagi, J., Yap, K.S., Tiong, S.K., Ahmed, S.K. and Mohammad, A.M., 2008, November. Detection of abnormalities and electricity theft using genetic support vector machines. In TENCON 2008-2008 IEEE region 10 conference (pp. 1-6). IEEE.

[57] Balakrishna, T., rao Kolavennu, S., Reddy, M.V.V., Goud, P.R. and Rohith, B., Electricity Theft Detection in Power Grids with Deep Learning & Random Forests.

[58] Uidhir, T.M., Rogan, F., Collins, M. and Curtis, J., 2020. Residential stock data and dataset on energy efficiency characteristics of residential building fabrics in Ireland. Data in Brief, 29, pp.105247-105247.

[59] Mac Uidhir, T., Rogan, F., Collins, M., Curtis, J. and Gallachóir, B.Ó., 2020. Residential stock data and dataset on energy efficiency characteristics of residential building fabrics in Ireland. Data in brief, 29, p.105247.

[60] Lyu, X. and Zhang, L., 2022, July. A Spatial Attention Guided Scene Classification Method for Multi-scale Remote Sensing Dataset. In IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium (pp. 2235-2238). IEEE.

[61] Depuru, S.S.S.R., Wang, L. and Devabhaktuni, V., 2011, March. Support vector machine based data classification for detection of electricity theft. In 2011 IEEE/PES Power Systems Conference and Exposition (pp. 1-8). IEEE.

[62] Toma, R.N., Hasan, M.N., Nahid, A.A. and Li, B., 2019, May. Electricity theft detection to reduce non-technical loss using support vector machine in smart grid. In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) (pp. 1-6). IEEE.

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

[64] Kamiran, F. and Calders, T., 2012. Data preprocessing techniques for classification without discrimination. Knowledge and information systems, 33(1), pp.1-33.

[65] Zhou, Y., Zhang, X., Tang, Y., Mu, Z., Shao, X., Li, Y. and Cai, Q., 2021, July. Convolutional Neural Network and Data Augmentation Method for Electricity Theft Detection. In 2021 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia) (pp. 1525-1530). IEEE.

[66] Liao, W., Yang, Z., Bak-Jensen, B., Pillai, J.R., Von Krannichfeldt, L., Wang, Y. and Yang, D., 2023. Simple Data Augmentation Tricks for Boosting Performance on Electricity Theft Detection Tasks. IEEE Transactions on Industry Applications.

[67] Chen, X., Qiu, X., Ma, Y., Wang, L. and Fang, L., 2022, June. Boruta-XGBoost Electricity Theft Detection Based on Features of Electric Energy Parameters. In Journal of Physics: Conference Series (Vol. 2290, No. 1, p. 012121). IOP Publishing.

[68] Shaaban, M., Tariq, U., Ismail, M., Almadani, N.A. and Mokhtar, M., 2021. Data-driven detection of electricity theft cyberattacks in PV generation. IEEE Systems Journal, 16(2), pp.3349-3359.

[69] Takiddin, A., Ismail, M., Nabil, M., Mahmoud, M.M. and Serpedin, E., 2020. Detecting electricity theft cyber-attacks in AMI networks using deep vector embeddings. IEEE Systems Journal, 15(3), pp.4189-4198.

[70] Soleymanzadeh, R. and Kashef, R., 2022, January. The future roadmap for cyber-attack detection. In 2022 6th international conference on cryptography, security and privacy (CSP) (pp. 66-70). IEEE.

[71] Xia, X., Xiao, Y., Liang, W. and Cui, J., 2022. Detection methods in smart meters for electricity thefts: A survey. Proceedings of the IEEE, 110(2), pp.273-319.

[72] Arif, A., Alghamdi, T.A., Khan, Z.A. and Javaid, N., 2022. Towards efficient energy utilization using big data analytics in smart cities for electricity theft detection. Big Data Research, 27, p.100285.

[73] Somefun, T.E., Awosope, C.O.A. and Chiagoro, A., 2019. Smart prepaid energy metering system to detect energy theft with facility for real time monitoring. International Journal of Electrical and Computer Engineering, 9(5), p.4184.

[74] Kim, J.Y., Hwang, Y.M., Sun, Y.G., Sim, I., Kim, D.I. and Wang, X., 2019. Detection for non-technical loss by smart energy theft with intermediate monitor meter in smart grid. IEEE Access, 7, pp.129043-129053.

[75] Dortolina, C.A. and Nadira, R., 2005. The loss that is unknown is no loss at all: A top-down/bottom-up approach for estimating distribution losses. IEEE Transactions on Power Systems, 20(2), pp.1119-1125.

[76] Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K., 2003. KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg.

[77] GLAUNER, P., Big Data-Driven Detection of False Data Injection Attacks in Smart Meters.

[78] Ullah, A., Javaid, N., Samuel, O., Imran, M. and Shoaib, M., 2020, June. CNN and GRU based deep neural network for electricity theft detection to secure smart grid. In 2020 International Wireless Communications and Mobile Computing (IWCMC) (pp. 1598-1602). IEEE.

[79] Tharwat, A., Gaber, T., Ibrahim, A. and Hassanien, A.E., 2017. Linear discriminant analysis: A detailed tutorial. AI communications, 30(2), pp.169-190.

[80] Park, S.H., Goo, J.M. and Jo, C.H., 2004. Receiver operating characteristic (ROC) curve: practical review for radiologists. Korean journal of radiology, 5(1), pp.11-18.

[81] Cook, N.R., 2008. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clinical chemistry, 54(1), pp.17-23.

[82] Coyne, B. and Denny, E., 2021. Retrofit effectiveness: Evidence from a nationwide residential energy efficiency programme. Energy Policy, 159, p.112576.

[83] Soleimani Nasab, F. and Ghaderi, F., 2022. On the effect of sampling frequency on the electricity theft detection performance. IET Signal Processing, 16(9), pp.1094-1105.

[84] Ghaedi, H., Tabbakh Farizani, S.R.K. and Ghaemi, R., 2021. Improving power theft detection using efficient clustering and ensemble classification. International Journal of Electrical & Computer Engineering (2088-8708), 11(5).

[85] Wang, X., Xie, H., Tang, L., Chen, C. and Bie, Z., 2023. Decentralized Privacy-Preserving Electricity Theft Detection for Distribution System Operators. IEEE Transactions on Smart Grid.

[86] Zia, M.F. and Ali, F., 2016. An efficient electricity theft and fault detection scheme in distribution system.

[87] Ahmed, M., Khan, A., Ahmed, M., Tahir, M., Jeon, G., Fortino, G. and Piccialli, F., 2022. Energy theft detection in smart grids: taxonomy, comparative analysis, challenges, and future research directions. IEEE/CAA Journal of Automatica Sinica, 9(4), pp.578-600.

[88] Muzumdar, A., Modi, C. and Vyjayanthi, C., 2022. Designing a blockchain-enabled privacy-preserving energy theft detection system for smart grid neighborhood area network. Electric Power Systems Research, 207, p.107884.

[89] McLaughlin, S., Holbert, B., Fawaz, A., Berthier, R. and Zonouz, S., 2013. A multi-sensor energy theft detection framework for advanced metering infrastructures. IEEE journal on selected areas in communications, 31(7), pp.1319-1330.

[90] Sola, J. and Sevilla, J., 1997. Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Transactions on nuclear science, 44(3), pp.1464-1468.

[91] Ogasawara, E., Martinez, L.C., De Oliveira, D., Zimbrão, G., Pappa, G.L. and Mattoso, M., 2010, July. Adaptive normalization: A novel data normalization approach for non-stationary time series. In The 2010 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[92] Guerrero-Lemus, R., Cañadillas-Ramallo, D., Reindl, T. and Valle-Feijóo, J.M., 2019. A simple big data methodology and analysis of the specific yield of all PV power plants in a power system over a long time period. Renewable and Sustainable Energy Reviews, 107, pp.123-132.

[93] Badr, M.M., Ibrahem, M.I., Mahmoud, M., Fouda, M.M., Alsolami, F. and Alasmary, W., 2021. Detection of false-reading attacks in smart grid net-metering system. IEEE Internet of Things Journal, 9(2), pp.1386-1401.

[94] Badr, M.M., Ibrahem, M.I., Mahmoud, M., Fouda, M.M. and Alasmary, W., 2020. Detection of false-reading attacks in the AMI net-metering system. arXiv preprint arXiv:2012.01983.

[95] Guerrero, J.I., Monedero, I., Biscarri, F., Biscarri, J., Millan, R. and Leon, C., 2017. Non-technical losses reduction by improving the inspections accuracy in a power utility. IEEE Transactions on Power Systems, 33(2), pp.1209-1218.

[96] León, C., Biscarri, F., Monedero, I., Guerrero, J.I., Biscarri, J. and Millán, R., 2011. Variability and trend-based generalized rule induction model to NTL detection in power companies. IEEE Transactions on Power Systems, 26(4), pp.1798-1807.

[97] Javaid, N., Qasim, U., Yahaya, A.S., Alkhammash, E.H. and Hadjouni, M., 2022. Non-technical losses detection using autoencoder and bidirectional gated recurrent unit to secure smart grids. IEEE Access, 10, pp.56863-56875.

[98] Kotsiantis, S.B., Kanellopoulos, D. and Pintelas, P.E., 2006. Data preprocessing for supervised leaning. International journal of computer science, 1(2), pp.111-117.

[99] Aldegheishem, A., Anwar, M., Javaid, N., Alrajeh, N., Shafiq, M. and Ahmed, H., 2021. Towards sustainable energy efficiency with intelligent electricity theft detection in smart grids emphasising enhanced neural networks. IEEE Access, 9, pp.25036-25061.

[100] Tehrani, S.O., Shahrestani, A. and Yaghmaee, M.H., 2022. Online electricity theft detection framework for large-scale smart grid data. Electric Power Systems Research, 208, p.107895.

[101] Takiddin, A., Ismail, M., Zafar, U. and Serpedin, E., 2020. Robust electricity theft detection against data poisoning attacks in smart grids. IEEE Transactions on Smart Grid, 12(3), pp.2675-2684.

[102] Zhu, Y., Zhang, Y., Liu, L., Liu, Y., Li, G., Mao, M. and Lin, L., 2022. Hybrid-order representation learning for electricity theft detection. IEEE Transactions on Industrial Informatics, 19(2), pp.1248-1259.

[103] Elgarhy, I., Badr, M.M., Mahmoud, M., Fouda, M.M., Alsabaan, M. and Kholidy, H.A., 2023. Clustering and Ensemble Based Approach For Securing Electricity Theft Detectors Against Evasion Attacks. IEEE Access.

[104] Zheng, K., Wang, Y., Chen, Q. and Li, Y., 2017, December. Electricity theft detecting based on density-clustering method. In 2017 IEEE Innovative Smart Grid Technologies-Asia (ISGT-Asia) (pp. 1-6). IEEE.

[105] Shehzad, F., Javaid, N., Aslam, S. and Javed, M.U., 2022. Electricity theft detection using big data and genetic algorithm in electric power systems. Electric Power Systems Research, 209, p.107975.

[106] Viegas, J.L. and Vieira, S.M., 2017, July. Clustering-based novelty detection to uncover electricity theft. In 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-6). IEEE.

[107] Hussain, S., Mustafa, M.W., Jumani, T.A., Baloch, S.K. and Saeed, M.S., 2020. A novel unsupervised feature-based approach for electricity theft detection using robust PCA and outlier removal clustering algorithm. International Transactions on Electrical Energy Systems, 30(11), p.e12572.

[108] Ibrahem, M.I., 2021. Privacy-preserving and efficient electricity theft detection and data collection for AMI using machine learning (Doctoral dissertation, Tennessee Technological University).

[109] Gu, D., Gao, Y., Chen, K., Shi, J., Li, Y. and Cao, Y., 2022. Electricity theft detection in AMI with low false positive rate based on deep learning and evolutionary algorithm. IEEE Transactions on Power Systems, 37(6), pp.4568-4578.

[110] Anas, M., Javaid, N., Mahmood, A., Raza, S.M., Qasim, U. and Khan, Z.A., 2012, November. Minimizing electricity theft using smart meters in AMI. In 2012 Seventh International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (pp. 176-182). IEEE.

[111] Dougherty, J., Kohavi, R. and Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. In Machine learning proceedings 1995 (pp. 194-202). Morgan Kaufmann.

[112] Khan, I.U., Javeid, N., Taylor, C.J., Gamage, K.A. and Ma, X., 2021. A stacked machine and deep learning-based approach for analysing electricity theft in smart grids. IEEE Transactions on Smart Grid, 13(2), pp.1633-1644.

[113] Kamakura, W.A., Wedel, M., De Rosa, F. and Mazzon, J.A., 2003. Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. International Journal of Research in marketing, 20(1), pp.45-65.

[114] Thoma, D.S., Benić, G.I., Zwahlen, M., Hämmerle, C.H. and Jung, R.E., 2009. A systematic review assessing soft tissue augmentation techniques. Clinical oral implants research, 20, pp.146-165.

[115] Aghaloo, T.L. and Moy, P.K., 2007. Which hard tissue augmentation techniques are the most successful in furnishing bony support for implant placement?. International Journal of Oral & Maxillofacial Implants, 22(7).

[116] Chawla, N.V., 2010. Data mining for imbalanced datasets: An overview. Data mining and knowledge discovery handbook, pp.875-886.

[117] Zhao, Z. and Liu, H., 2007, June. Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th international conference on Machine learning (pp. 1151-1157).

[118] Guyon, I. and Elisseeff, A., 2006. An introduction to feature extraction. In Feature extraction: foundations and applications (pp. 1-25). Berlin, Heidelberg: Springer Berlin Heidelberg.

[119] Kuo, B.C. and Landgrebe, D.A., 2004. Nonparametric weighted feature extraction for classification. IEEE Transactions on Geoscience and Remote Sensing, 42(5), pp.1096-1105.

[120] Maraden, Y., Wibisono, G., Nugraha, I.G.D., Sudiarto, B., Jufri, F.H. and Prabuwono, A.S., 2023. Enhancing Electricity Theft Detection through K-Nearest Neighbors and Logistic Regression Algorithms with Synthetic Minority Oversampling Technique: A Case Study on State Electricity Company (PLN) Customer Data. Energies, 16(14), p.5405.

[121] Pereira, J. and Saraiva, F., 2021. Convolutional neural network applied to detect electricity theft: A comparative study on unbalanced data handling techniques. International Journal of Electrical Power & Energy Systems, 131, p.107085.

[122] Berry, M.W., Mohamed, A. and Yap, B.W. eds., 2019. Supervised and unsupervised learning for data science. Springer Nature.

[123] Javaid, N., Almogren, A., Adil, M., Javed, M.U. and Zuair, M., 2022. RFE based feature selection and KNNOR based data balancing for electricity theft detection using BiLSTM-LogitBoost stacking ensemble model. IEEE Access, 10, pp.112948-112963.

[124] Ali, P.J.M., Faraj, R.H., Koya, E., Ali, P.J.M. and Faraj, R.H., 2014. Data normalization and standardization: a technical report. Mach Learn Tech Rep, 1(1), pp.1-6.

[125] Sun, Y., Sun, X., Hu, T. and Zhu, L., 2023. Smart Grid Theft Detection Based on Hybrid Multi-Time Scale Neural Network. Applied Sciences, 13(9), p.5710.

[126] Shafee, A., Fouda, M.M., Mahmoud, M., Alasmary, W., Aljohani, A.J. and Amsaad, F., 2020. Detection of Lying Electrical Vehicles in Charging Coordination Application Using Deep Learning. arXiv preprint arXiv:2005.13813.

[127] Xia, X., Lin, J., Jia, Q., Wang, X., Ma, C., Cui, J. and Liang, W., 2023. ETD-ConvLSTM: A Deep Learning Approach for Electricity Theft Detection in Smart Grids. IEEE Transactions on Information Forensics and Security.

[128] Lin, G., Feng, X., Guo, W., Cui, X., Liu, S., Jin, W., Lin, Z. and Ding, Y., 2021. Electricity theft detection based on stacked autoencoder and the undersampling and resampling based random forest algorithm. Ieee Access, 9, pp.124044-124058.

[129] Alameady, M.H., George, L.E. and Albermany, S., 2023. Energy Theft Detection with Determine Date Theft Period for State Grid Corporation of China Dataset. International Journal of Intelligent Systems and Applications in Engineering, 11(1s), pp.01-13.

[130] Franke, R. and Nielson, G.M., 1991. Scattered data interpolation and applications: A tutorial and survey. Geometric Modeling: Methods and Applications, pp.131-160.

[131] Kaur, H., Pham, N. and Fomel, S., 2019. Seismic data interpolation using CycleGAN. In SEG technical program expanded abstracts 2019 (pp. 2202-2206). Society of Exploration Geophysicists.

[132] Van De Weijer, J. and Schmid, C., 2006. Coloring local feature extraction. In Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part II 9 (pp. 334-348). Springer Berlin Heidelberg.

[133] Li, S., Han, Y., Yao, X., Yingchen, S., Wang, J. and Zhao, Q., 2019. Electricity theft detection in power grids with deep learning and random forests. Journal of Electrical and Computer Engineering, 2019, pp.1-12.

[134] Alfeld, P., 1989. Scattered data interpolation in three or more variables. In Mathematical methods in computer aided geometric design (pp. 1-33). Academic Press.

[135] Zhang, H., Berg, A.C., Maire, M. and Malik, J., 2006, June. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 2, pp. 2126-2136). IEEE.

[136] Torkkola, K., 2003. Feature extraction by non-parametric mutual information maximization. Journal of machine learning research, 3(Mar), pp.1415-1438.

[137] Bamane, R., Vinod, M., Shah, J., Ahuja, S. and Sahariya, A., 2020. Smart meter for power theft detection using machine learning. International Journal of Scientific Research and Engineering Development, 3(1), pp.526-528.

[138] Kawoosa, A.I., Prashar, D., Faheem, M., Jha, N. and Khan, A.A., 2023. Using machine learning ensemble method for detection of energy theft in smart meters. IET Generation, Transmission & Distribution.

[139] Li, J.B. and Gao, H., 2012. Sparse data-dependent kernel principal component analysis based on least squares support vector machine for feature extraction and recognition. Neural Computing and Applications, 21, pp.1971-1980.

[140] Ullah, A., Javaid, N., Asif, M., Javed, M.U. and Yahaya, A.S., 2022. Alexnet, adaboost and artificial bee colony based hybrid model for electricity theft detection in smart grids. Ieee Access, 10, pp.18681-18694.

[141] Kulkarni, Y., Hussain, S., Ramamritham, K. and Somu, N., 2021, December. EnsembleNTLDetect: an intelligent framework for electricity theft detection in smart grid. In 2021 International Conference on Data Mining Workshops (ICDMW) (pp. 527-536). IEEE.

[142] Zhang, S., Li, X., Zong, M., Zhu, X. and Wang, R., 2017. Efficient kNN classification with different numbers of nearest neighbors. IEEE transactions on neural networks and learning systems, 29(5), pp.1774-1785.

[143] Shahid, M.B., Shahid, M.O., Tariq, H. and Saleem, S., 2019, July. Design and development of an efficient power theft detection and prevention system through consumer load profiling. In 2019 international conference on electrical, communication, and computer engineering (ICECCE) (pp. 1-6). IEEE.

[144] Nguyen, T.H.T. and Phan, Q.B., 2021, December. Electricity theft detection in power grid with a hybrid convolutional neural network-support vector machine model. In The 5th international conference on future networks & distributed systems (pp. 24-30).

[145] Petrlik, I., Lezama, P., Rodriguez, C., Inquilla, R., Reyna-González, J.E. and Esparza, R., 2022. Electricity Theft Detection using Machine Learning. International Journal of Advanced Computer Science and Applications, 13(12).

[146] Zhang, Q., Zhang, M., Chen, T., Fan, J., Yang, Z. and Li, G., 2018, November. Electricity theft detection using generative models. In 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 270-274). IEEE.

[147] Zheng, K., Chen, Q., Wang, Y., Kang, C. and Xia, Q., 2018. A novel combined data-driven approach for electricity theft detection. IEEE Transactions on Industrial Informatics, 15(3), pp.1809-1819.

[148] Kong, X., Zhao, X., Liu, C., Li, Q., Dong, D. and Li, Y., 2021. Electricity theft detection in low-voltage stations based on similarity measure and DT-KSVM. International Journal of Electrical Power & Energy Systems, 125, p.106544.

[149] Pereira, D.R., Pazoti, M.A., Pereira, L.A., Rodrigues, D., Ramos, C.O., Souza, A.N. and Papa, J.P., 2016. Social-Spider Optimization-based Support Vector Machines applied for energy theft detection. Computers & Electrical Engineering, 49, pp.25-38

[150] Shen, Y., Shao, P., Chen, G., Gu, X., Wen, T., Zang, L. and Zhu, J., 2021. An identification method of anti-electricity theft load based on long and short-term memory network. Procedia Computer Science, 183, pp.440-447.

[151] Nayak, S.C., Misra, B.B. and Behera, H.S., 2014. Impact of data normalization on stock index forecasting. International Journal of Computer Information Systems and Industrial Management Applications, 6(2014), pp.257-269.

[152] Borkin, D., Némethová, A., Michaľčonok, G. and Maiorov, K., 2019. Impact of data normalization on classification model accuracy. Research Papers Faculty of Materials Science and Technology Slovak University of Technology, 27(45), pp.79-84.

[153] Ioffe, S., 2006. Probabilistic linear discriminant analysis. In Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9 (pp. 531-542). Springer Berlin Heidelberg.

[154] Cochran, W.G., 1964. On the performance of the linear discriminant function. Technometrics, 6(2), pp.179-190.

[155] Qi, R., Li, Q., Luo, Z., Zheng, J. and Shao, S., 2023. Deep semi-supervised electricity theft detection in AMI for sustainable and secure smart grids. Sustainable Energy, Grids and Networks, p.101219.

[156] Yip, S.C., Tan, W.N., Tan, C., Gan, M.T. and Wong, K., 2018. An anomaly detection framework for identifying energy theft and defective meters in smart grids. International Journal of Electrical Power & Energy Systems, 101, pp.189-203.

[157] Qi, R., Zheng, J., Luo, Z. and Li, Q., 2022. A novel unsupervised data-driven method for electricity theft detection in AMI using observer meters. IEEE Transactions on Instrumentation and Measurement, 71, pp.1-10.

[158] Jain, A.K., Murty, M.N. and Flynn, P.J., 1999. Data clustering: a review. ACM computing surveys (CSUR), 31(3), pp.264-323.

[159] Gan, G., Ma, C. and Wu, J., 2020. Data clustering: theory, algorithms, and applications. Society for Industrial and Applied Mathematics.

[160] Kao, Y.T., Zahara, E. and Kao, I.W., 2008. A hybridized approach to data clustering. Expert Systems with Applications, 34(3), pp.1754-1762.

[161] Mohammad, N., Barua, A. and Arafat, M.A., 2013, February. A smart prepaid energy metering system to control electricity theft. In 2013 international conference on power, energy and control (ICPEC) (pp. 562-565). IEEE.

[162] Mohammad, N., Barua, A. and Arafat, M.A., 2013, February. A smart prepaid energy metering system to control electricity theft. In 2013 international conference on power, energy and control (ICPEC) (pp. 562-565). IEEE.

[163] Asif, M., Ullah, A., Munawar, S., Kabir, B., Pamir, Khan, A. and Javaid, N., 2021. Alexnet-AdaBoost-ABC based hybrid neural network for Electricity Theft Detection in smart grids. In Complex, Intelligent and Software Intensive Systems: Proceedings of the 15th International Conference on Complex, Intelligent and Software Intensive Systems (CISIS-2021) (pp. 249-258). Springer International Publishing.

[164] Asif, Muhammad, Benish Kabir, Pamir, Ashraf Ullah, Shoaib Munawar, and Nadeem Javaid. "Towards energy efficient smart grids: Data augmentation through BiWGAN, feature extraction and classification using hybrid 2DCNN and BiLSTM." In Innovative Mobile and Internet Services in Ubiquitous Computing: Proceedings of the 15th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS-2021), pp. 108-119. Springer International Publishing, 2022.

[165] Asif, M., Kabir, B., Pamir, A.U., Munawar, S. and Javaid, N., A Hybrid Deep Learning Approach for Detecting Non Technical Losses in Smart Grids.

[166] Amhenrior, H.E., Oloma, I. and Ikharo, B.A., 2022, November. Design and Implementation of a GSM and Wi-Fi Based Low Cost Smart Prepaid Energy Meter. In 2022 5th Information Technology for Education and Development (ITED) (pp. 1-8). IEEE.

[167] Jeffin, M.J., Madhu, G.M., Rao, A., Singh, G. and Vyjayanthi, C., 2020, July. Internet of things enabled power theft detection and smart meter monitoring system. In 2020 International Conference on Communication and Signal Processing (ICCSP) (pp. 0262-0267). IEEE.

[168] Ganguly, P., Nasipuri, M. and Dutta, S., 2018. A novel approach for detecting and mitigating the energy theft issues in the smart metering infrastructure. Technology and Economics of Smart Grids and Sustainable Energy, 3, pp.1-11.

[169] Munawar, S., Asif, M., Kabir, B., Pamir, Ullah, A. and Javaid, N., 2021. Electricity theft detection in smart meters using a hybrid bi-directional GRU bi-directional LSTM model. In Complex, Intelligent and Software Intensive Systems: Proceedings of the 15th International Conference on Complex, Intelligent and Software Intensive Systems (CISIS-2021) (pp. 297-308). Springer International Publishing.

[170] Munawar, S., Javaid, N., Khan, Z.A., Chaudhary, N.I., Raja, M.A.Z., Milyani, A.H. and Ahmed Azhari, A., 2022. Electricity Theft Detection in Smart Grids Using a Hybrid BiGRU–BiLSTM Model with Feature Engineering-Based Preprocessing. Sensors, 22(20), p.7818.

[171] Munawar, S., Khan, Z.A., Chaudhary, N.I., Javaid, N., Raja, M.A.Z., Milyani, A.H. and Azhari, A.A., 2022. Novel FDIs-based data manipulation and its detection in smart meters' electricity theft scenarios. Frontiers in Energy Research, 10, p.1043593.

[172] Munawar, S., Khan, Z.A., Chaudhary, N.I., Javaid, N. and Raja, M.A.Z., 2023. Machine intelligence aware electricity theft detection for smart metering applications. Waves in Random and Complex Media, pp.1-21.

[173] Kabir, B., Pamir, Ullah, A., Munawar, S., Asif, M. and Javaid, N., 2021. Detection of non-technical losses using MLP-GRU based neural network to secure smart grids. In Complex, Intelligent and Software Intensive Systems: Proceedings of the 15th International Conference on Complex, Intelligent and Software Intensive Systems (CISIS-2021) (pp. 383-394). Springer International Publishing.

[174] Javaid, N., Gul, H., Baig, S., Shehzad, F., Xia, C., Guan, L. and Sultana, T., 2021. Using GANCNN and ERNET for detection of non technical losses to secure smart grids. IEEE Access, 9, pp.98679-98700.

[175] Harshini, C., Deepthi, G., Reddy, G.A., Laxmi, G.V. and Rajasree, G., 2023. ELECTRICITY THEFT DETECTION IN POWER GRIDS WITH DEEP LEARNING AND RANDOM FORESTS. International Journal of Management Research and Reviews, 13(3), pp.1-10.

[176] Gong, X., Tang, B., Zhu, R., Liao, W. and Song, L., 2020. Data augmentation for electricity theft detection using conditional variational auto-encoder. Energies, 13(17), p.4291.

[177] Pereira, J. and Saraiva, F., 2021. Convolutional neural network applied to detect electricity theft: A comparative study on unbalanced data handling techniques. International Journal of Electrical Power & Energy Systems, 131, p.107085.

[178] Javaid, N., Jan, N. and Javed, M.U., 2021. An adaptive synthesis to handle imbalanced big data with deep siamese network for electricity theft detection in smart grids. Journal of Parallel and Distributed Computing, 153, pp.44-52.

[179] Qu, Z., Liu, H., Wang, Z., Xu, J., Zhang, P. and Zeng, H., 2021. A combined genetic optimization with AdaBoost ensemble model for anomaly detection in buildings electricity consumption. Energy and Buildings, 248, p.111193.

[180] Chuwa, M.G. and Wang, F., 2021. A review of non-technical loss attack models and detection methods in the smart grid. Electric Power Systems Research, 199, p.107415.

[181] Jindal, A., Dua, A., Kaur, K., Singh, M., Kumar, N. and Mishra, S., 2016. Decision tree and SVM-based data analytics for theft detection in smart grid. IEEE Transactions on Industrial Informatics, 12(3), pp.1005-1016.

[182] Dong, S., Zeng, Z. and Liu, Y., 2021. FPETD: Fault-tolerant and privacy-preserving electricity theft detection. Wireless Communications and Mobile Computing, 2021, pp.1-11.

[183] Cheng, G., Zhang, Z., Li, Q., Li, Y. and Jin, W., 2021. Energy theft detection in an edge data center using deep learning. Mathematical Problems in Engineering, 2021, pp.1-12.

[184] Hu, J., Li, S., Hu, J. and Yang, G., 2018. A hierarchical feature extraction model for multi-label mechanical patent classification. Sustainability, 10(1), p.219.

[185] Anwar, M., Javaid, N., Khalid, A., Imran, M. and Shoaib, M., 2020, June. Electricity theft detection using pipeline in machine learning. In 2020 International Wireless Communications and Mobile Computing (IWCMC) (pp. 2138-2142). IEEE.

[186] Wang, W., Zhang, M., Wang, D., Jiang, Y., Li, Y. and Wu, H., 2019. Anomaly detection based on kernel principal component and principal component analysis. In Communications, Signal Processing, and Systems: Proceedings of the 2017 International Conference on Communications, Signal Processing, and Systems (pp. 2222-2228). Springer Singapore.

[187] Prakash, A., Shyam Joseph, A., Shanmugasundaram, R. and Ravichandran, C.S., 2023. A machine learning approach-based power theft detection using GRF optimization. Journal of Engineering, Design and Technology, 21(5), pp.1373-1388.

[188] Yip, S.C., Wong, K., Hew, W.P., Gan, M.T., Phan, R.C.W. and Tan, S.W., 2017. Detection of energy theft and defective smart meters in smart grids using linear regression. International Journal of Electrical Power & Energy Systems, 91, pp.230-240.

[189] Patil, N.V., Kanase, R.S., Bondar, D.R. and Bamane, P.D., 2017, February. Intelligent energy meter with advanced billing system and electricity theft detection. In 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI) (pp. 36-41). IEEE.

[190] Razavi, R., Gharipour, A., Fleury, M. and Akpan, I.J., 2019. A practical feature-engineering framework for electricity theft detection in smart grids. Applied energy, 238, pp.481-494.

[191] Shanthi, M., Koodalarasan, M., Varshan, P.A.V. and Visvabarathi, S., 2022. IoT based Electricity Theft Detection System. International Research Journal of Modernization in Engineering Technology and Science, 4(12).

[192] Mohite, N., Ranaware, R. and Kakade, P., 2016. GSM based electricity theft detection. International Journal of Scientific Engineering and Applied Science, 2.

[193] Alromih, A., Clark, J.A. and Gope, P., 2021, October. Electricity theft detection in the presence of prosumers using a cluster-based multi-feature detection model. In 2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm) (pp. 339-345). IEEE.

[194] Aziz, S., Naqvi, S.Z.H., Khan, M.U. and Aslam, T., 2020, March. Electricity theft detection using empirical mode decomposition and K-nearest neighbors. In 2020 International Conference on Emerging Trends in Smart Technologies (ICETST) (pp. 1-5). IEEE.

[195] Ibrahem, M.I., Mahmoud, M.M., Alsolami, F., Alasmary, W., AL-Ghamdi, A.S.A.M. and Shen, X., 2022. Electricity-theft detection for change-and-transmit advanced metering infrastructure. IEEE Internet of Things Journal, 9(24), pp.25565-25580.