energies

MDPI

Review

Data-Driven Approaches for Energy Theft Detection: A Comprehensive Review

Soohyun Kim [1], Youngghyu Sun [2], Seongwoo Lee [1], Joonho Seon [1], Byungsun Hwang [1], Jeongho Kim [1], Jinwook Kim [1], Kyounghun Kim [1] and Jinyoung Kim [1,*]

[1] Department of Electronic Convergence Engineering, Kwangwoon University, Seoul 01897, Republic of Korea; kimsoogus@kw.ac.kr (S.K.); swoo1467@kw.ac.kr (S.L.); dimlight13@kw.ac.kr (J.S.); hbsun1225@kw.ac.kr (B.H.); jh980828@kw.ac.kr (J.K.); yoonlight12@kw.ac.kr (J.K.); sdrudgnsdl@kw.ac.kr (K.K.)

[2] Research and Development Department, SMART EVER, Co., Ltd., Seoul 01886, Republic of Korea; ygsun@smartever.net

* Correspondence: jinyoung@kw.ac.kr; Tel.: +82-02-940-5567

Abstract: The transition to smart grids has served to transform traditional power systems into data-driven power systems. The purpose of this transition is to enable effective energy management and system reliability through an analysis that is centered on energy information. However, energy theft caused by vulnerabilities in the data collected from smart meters is emerging as a primary threat to the stability and profitability of power systems. Therefore, various methodologies have been proposed for energy theft detection (ETD), but many of them are challenging to use effectively due to the limitations of energy theft datasets. This paper provides a comprehensive review of ETD methods, highlighting the limitations of current datasets and technical approaches to improve training datasets and the ETD in smart grids. Furthermore, future research directions and open issues from the perspective of generative AI-based ETD are discussed, and the potential of generative AI in addressing dataset limitations and enhancing ETD robustness is emphasized.

Keywords: energy theft detection; data-driven approach; generative AI; supervised learning; semi-supervised learning; smart meter

check for updates

Citation: Kim, S.; Sun, Y.; Lee, S.; Seon, J.; Hwang, B.; Kim, J.; Kim, J.; Kim, K.; Kim, J. Data-Driven Approaches for Energy Theft Detection: A Comprehensive Review. *Energies* **2024**, *17*, 3057. https://doi.org/10.3390/en17123057

Academic Editors: Fernando Sánchez Lasheras and Daniel Pérez-López

Received: 15 May 2024
Revised: 15 June 2024
Accepted: 15 June 2024
Published: 20 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The development of the smart grid has revolutionized the operation and management of energy systems [1]. The smart grid has been recognized as a future solution for smart energy monitoring as it is aimed at stabilizing and optimizing complex energy networks. Among these digital transformations in the energy management system, the capability to analyze comprehensive energy information is essential. The collection and analysis of energy information plays an important role in improving energy management by enabling the understanding of energy usage patterns, optimizing energy distribution, and minimizing energy loss. Furthermore, it can enhance the reliability and stability of energy systems by detecting potential problems and implementing appropriate measures to prevent system failures.

The transformation of energy utilities into smart grids is being made possible by the introduction of metering systems due to the widespread adoption of smart meters. The large amount of energy information collected from smart meters serves to provide various data-driven services to individuals, enterprises, and utilities. These services include demand forecasting, flexibility forecasting, impedance estimation, phase grouping, remote switching, and hosting capacity [2]. However, several concerns regarding data vulnerability have surfaced as energy management systems increasingly evolve into data-driven systems [3]. Energy information is susceptible to hacking through unauthorized access, thus posing significant threats to consumer security and privacy [4]. One primary threat from data vulnerability is energy theft, which can lead to substantial financial losses for

distributed system operators (DSOs) and consumers [3,4]. The methods used in energy theft typically aim to reduce electricity bills or illicitly take financial benefits from the power grid, which damages the infrastructure of the energy system and results in significant revenue losses for utilities. According to the Northeast Group [5], energy theft incurs a massive global cost of USD 96 billion annually. In the USA, Progress Energy Incorporated reported a 5% increase in such incidents [6].

In order to build data-driven systems, big data techniques need to be considered. Big data characteristics can be summarized by the 5 Vs: velocity, volume, value, variety, and veracity [7]. These aspects are crucial in the analysis and processing of large datasets, such as those used in data-driven approaches. Therefore, integrating artificial intelligence (AI) with big data enables the extraction of valuable insights and the enhancement of predictive models, thereby improving applications such as data-driven energy theft detection (ETD).

In order to prevent energy theft from causing losses in the smart grid, various methodologies for detecting energy theft have been researched. ETD has mainly been classified into hardware-based and data-driven methods [8]. Hardware-based ETD has been proposed to detect the energy theft that occurs through the physical manipulation of mechanical meters in traditional power grids. Data-driven ETD has been proposed to detect, based on energy usage and power flow patterns, suspected energy theft. With the development of smart grids, the focus has shifted to data-based ETD because energy theft primarily occurs through data manipulation rather than the physical manipulation of meters [9].

In Figure 1, the technical elements in designing ETD using a data-driven approach are shown. The design of data-driven ETD involves several steps, such as data collection, data processing, malicious behavior modeling, and developing an intelligent algorithm that can detect energy theft. Most of the prior research has relied on various methodologies, including machine learning and deep learning, to achieve the guaranteed performance of data-driven ETD. A deep learning-based ETD model may perform better on datasets with a balanced distribution, but its performance may degrade on datasets with an imbalanced data distribution [10]. ETD with an imbalanced dataset can be biased, suitable only for specific cases, and pose challenges to model scalability.

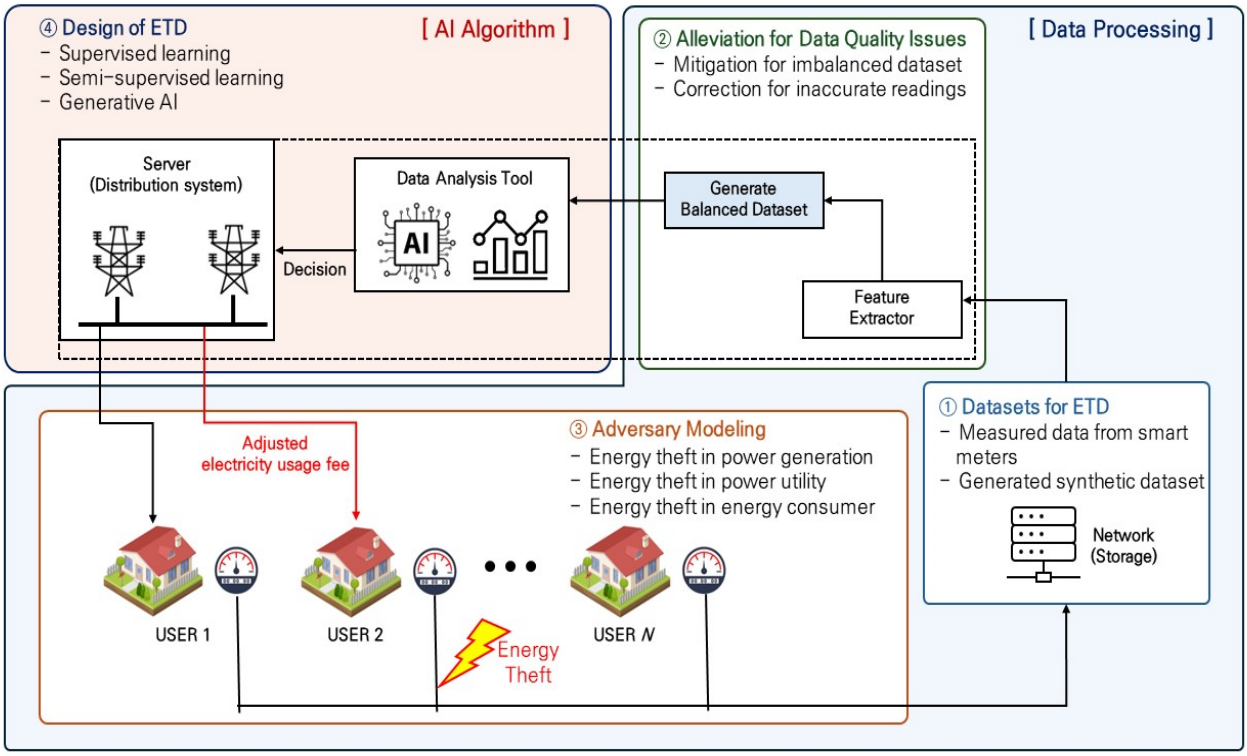


Figure 1. Overall design process of data-driven ETD.

Generative AI models are being introduced in ETD using augmentation with training datasets to handle the limitations of energy theft datasets [3,11–13]. Unlike traditional AI models, generative AI can be utilized to detect anomalies in energy consumption patterns by analyzing vast volumes of data. Moreover, generative AI can help to analyze data distribution and extract features across large spaces and high dimensions related to energy theft [14,15]. Generative AI-based ETD can help reconstruct an energy theft dataset in which a practical energy theft occurs. Even though generative AI can be expected to bring various improvements to ETD, there is still a lack of research.

In Table 1, a comparison of previous survey papers on data-driven ETD is presented. While existing review papers focus on specific aspects of data-driven ETD, such as methodologies [16–20], smart grid components [21], consumer privacy [22], and the scale of energy usage [23], there has not been an in-depth analysis of data-driven ETD that emphasizes the limitations of energy theft datasets. This survey addresses these limitations and categorizes AI modeling based on the types of energy theft datasets. Additionally, this paper provides insights into designing data-driven ETD using state-of-the-art generative AI, offering a comprehensive perspective for future research.

Table 1. Comparison of data-driven approaches for ETD with related survey papers.

Review	Year	Data Issues	Categorizations of ETD	Main Contribution
Viegas et al. [16]	2017	- Lack of public data	- Classification - Estimation - Game theory	- ETD analysis in terms of theoretical studies, hardware, and non-hardware solutions
Althobaiti et al. [21]	2021	- Class imbalance - Adversary modeling	- Classification - Regression - Clustering	- Survey on ETD within the three functions of demand, supply, and generation
Xia et al. [17]	2022	- Not mentioned	- Machine learning-based method - Measurement mismatch-based method	- ETD analysis in terms of machine learning- and measurement mismatch-based methods
Ahmed et al. [18]	2022	- Not mentioned	- Data mining - State and network - Game theory	- ETD analysis in terms of data mining, state and network, and game theory
Badr et al. [22]	2023	- Class imbalance - Adversary modeling	- Supervised learning - Data mining - Clustering - Generative AI (GAN)	- Analysis of consumer privacy protection methods and ETD design based on adversary modeling
Guarda et al. [19]	2023	- Class imbalance	- Data-oriented method - Network-oriented method - Hybrid method	- ETD analysis in terms of data-, network-, and hybrid-oriented methods
Stracqualursi et al. [23]	2024	- High dimension - Class imbalance	- Neural network - Regression - Clustering	- ETD analysis of users connected to low-voltage and medium-voltage networks
Kgaphola et al. [20]	2024	- Not mentioned	- Classification - Regression - Clustering	- A comprehensive survey of articles covering technology-based ETD
Our Contributions	2024	- High dimension - Class imbalance - Inaccurate reading - Absence of labels - Adversary modeling	- Supervised learning - Semi-supervised learning - Generative AI (VAE, GAN, and diffusion model)	- An analysis of data-driven ETD from the perspective of the problems with energy theft datasets and how to design ETD from a deeper applicability perspective, including the latest generation of AI

In this paper, a comprehensive analysis of more than 80 research papers was carried out. The distribution of these publications is summarized in Figure 2a. Figure 2b shows the yearly distribution of the analyzed papers, highlighting the increasing interest in data-driven approaches for ETD and various AI algorithms in recent years. The investigated

papers outlined definitions of energy theft type, limitations of datasets, and methodological perspectives for designing data-driven ETD.

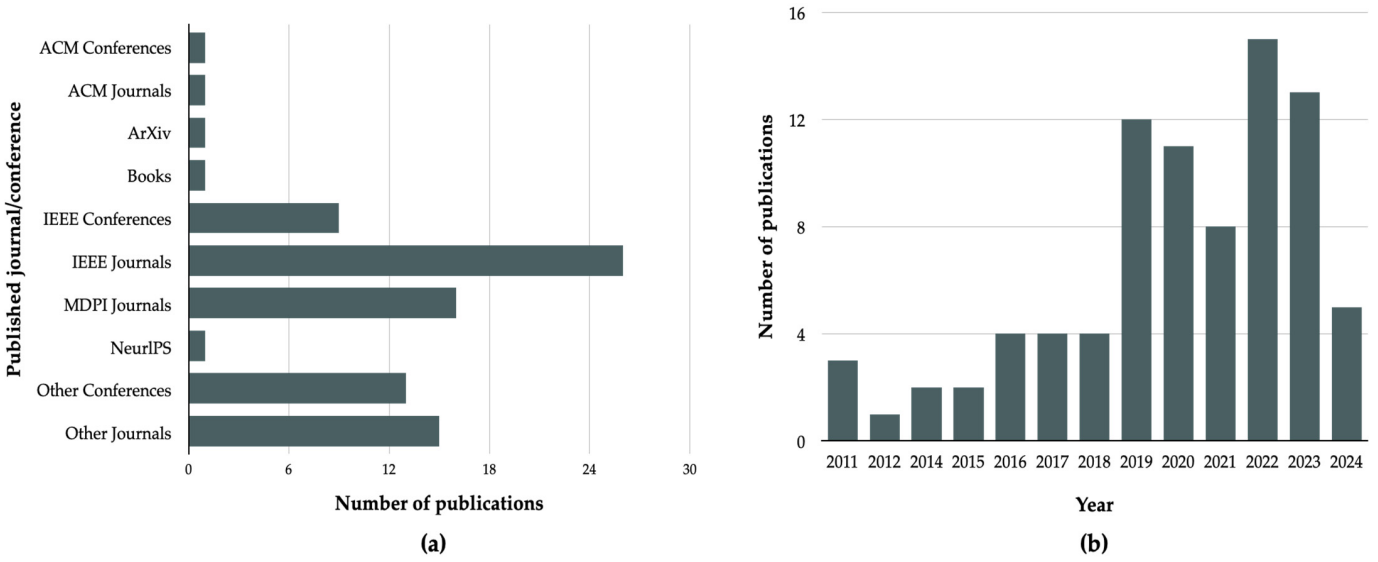


Figure 2. Number of publications per (**a**) journal/conference and (**b**) year.

In Figure 3, a comprehensive categorization of the proposed data-driven approaches for ETD is shown to address the challenges and to refine the methodologies. The diverse aspects of ETD are represented, including types of energy theft such as meter tampering, meter malfunctioning, cyber-attacks, feeder tapping, and billing irregularities. Additionally, various dataset issues are highlighted, such as high-dimensional data, imbalances, inaccurate readings, and missing labels. To address these issues, data-driven methodologies are categorized into supervised learning, semi-supervised learning, and generative AI.

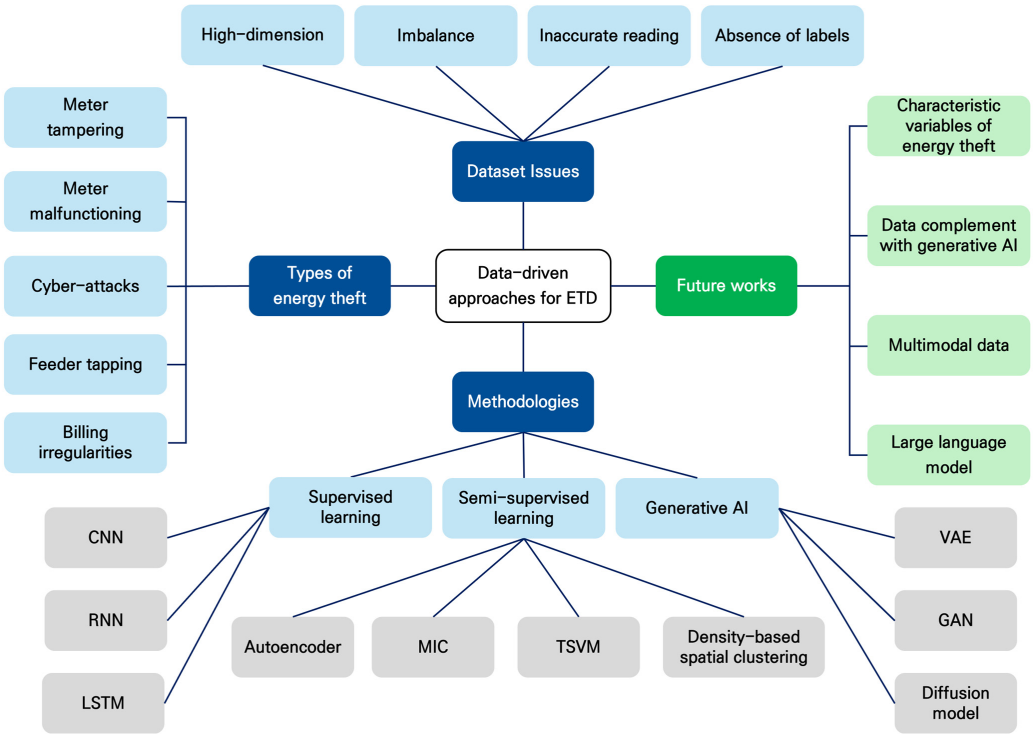


Figure 3. Overview and categorization of data-driven approaches for ETD.

The structure of this paper is as follows. Section 2 offers a review of energy theft methods used in energy management systems. Section 3 discusses the limitations of the datasets used in the design of data-driven ETD and technical methods. Section 4 reviews conventional AI and generative AI methods for detecting energy theft. Section 5 discusses the challenges and opportunities of ETD using generative AI, and the conclusions are presented in Section 6.

2. Energy Theft in Energy Management System

In an energy management system, losses are categorized into technical and non-technical losses [23]. Technical losses are caused by the energy dissipated in the conductors used in transmission, sub-transmission, and distribution lines. These losses are inherent to transmitting electricity over long distances and through various network components. Reducing these losses is crucial for improving grid efficiency and sustainability, thus requiring investments in efficient network equipment, advancements in the design and planning of transmission and distribution networks, and deploying intelligent technologies to monitor and manage electricity flow effectively. Non-technical losses are caused by external factors in the energy system, including administrative inefficiencies, theft, and billing errors, which are rampant with malicious and illegal activities. One of the significant contributors to non-technical losses, energy theft, is characterized by the unauthorized and illegal consumption of electricity. Energy theft leads to economic losses that are incurred by distributors and providers along with potential hazards, deterioration of service, and safety issues [24,25].

Energy theft is extensively passed off as the focus is on the various malicious practices employed by individuals to consume or reduce their billed energy consumption illegally. These methods are broadly categorized based on whether they involve tampering with metering devices or bypassing metering systems [23]. For users connected to a medium-voltage network, meter tampering methods are carried out by shunting the secondary winding of measuring current transformers and equipping electronic energy meters with external boards to alter measurements. For low-voltage network users, energy thefts are performed by bypassing the energy meter through illicit connections, directly tapping energy from the distribution network, and tampering with energy meters using external magnets to interfere with meter operations.

2.1. Types of Energy Theft

The key types of energy theft may encompass meter tampering and malfunctioning, feeder tapping, billing irregularities, and cyber-attacks [16,24–26]. The approaches to designing ETD by energy theft types are shown in Figure 4.

2.1.1. Meter Tampering and Malfunctioning to Achieve Energy Theft

Meter tampering and malfunctioning are characterized by deliberate alteration or damage being inflicted upon electricity meters to reduce recorded energy consumption [24]. The efficiency of meters is compromised by the insertion of strong magnets, leading to incomplete consumption recording. Various methods are employed to manipulate meter readings, including physical interference, such as attaching magnets to meters, slowing mechanical rotation, injecting substances to block the meter mechanism, and altering the internal wiring and components. The flow of all the current through the meter is prevented by shorting the ends, thereby failing to record the full energy consumption. Moreover, the amount of current measured is modified by altering the voltage wires of the electricity meter or the insulation on the secondary side wires.

2.1.2. Feeder Tapping for Energy Theft

Feeder tapping is commonly observed in areas where physical access to distribution lines is relatively uncomplicated and where monitoring is sparse [24]. Illegal or unauthorized connections are made by directly connecting to the transmission line. The illegal

connections to utility feeder lines are characterized by individuals bypassing the energy meter. Single-phase usage from a three-phase supply is employed to record zero voltage consumption and null energy consumption. In terms of safety issues, severe risks are posed by feeder tapping not only to the perpetrators themselves, but also to neighbors and utility workers. These risks are exacerbated by the unauthorized and unprotected wiring involved in such connections.

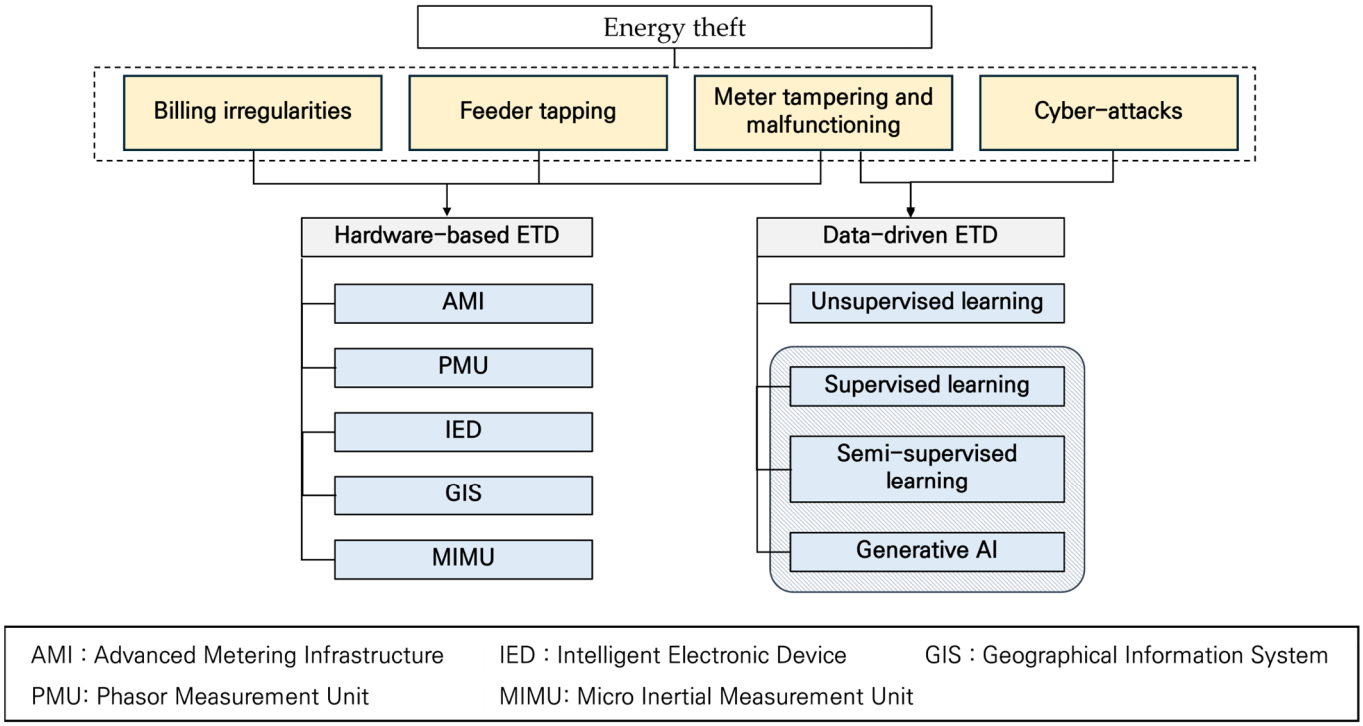


Figure 4. Approaches for designing ETD categorized by energy theft types.

2.1.3. Billing Irregularities for Energy Theft

Billing alterations are made through illicit payments to utility officials, leading to the recording of incorrect meter readings. Billing irregularities are characterized by various malicious activities associated with the manipulation of billing records [27]. These activities often involve collusion with utility employees, where meter readings or account records are altered to reduce charges. In addition, it may include hacking into the utility's billing software to change consumption data or the unauthorized use of another customer's identity to evade payment.

2.1.4. Cyber-Attacks for Energy Theft

With the rise of smart grids, cyber-attacks have become an increasingly prevalent method of energy theft [16,18,21]. These attacks typically involve hacking into the smart grid network to manipulate the data transmitted from smart meters to the utility provider. This includes injecting false data, which is achieved by manipulating the system to bypass existing detection methods, leading to erroneous data measurement. Such illegal activities result in economic losses and compromise the integrity and reliability of a power supply.

2.2. Hardware-Based Energy Theft Detection

Recent progress has been made in detecting energy theft by integrating advanced metering infrastructure (AMI) and machine learning algorithms within a smart grid [28,29]. The main goals of these methods are classified into prediction, monitoring, detection and localization, classification, and resolution, with algorithms being applied accordingly [23].

In a hardware-based approach, unusual or unauthorized energy access within an energy metering and distribution system can be monitored and detected using physical devices and sensors [8]. In order to monitor the real-time energy flow throughout a distribution network, the detection process is carried out by utilizing data from AMI, the phasor measurement unit (PMU), and an intelligent electronic device (IED) [30]. Additionally, energy theft is effectively identified and pinpointed by employing the A-star derivative algorithm and geographical information system (GIS) applications. In [31], advanced multi-sensor fusion technologies, which are based on micro-inertial measurement units (MIMUs) and intelligent pipeline inspection gauges (PIGs), were comprehensively reviewed to enhance the detection and localization of potential theft or leakage points. In [32], a novel method for hardware-based ETD, combining clustering and the local outlier factor (LOF), was proposed to effectively identify the energy theft within an AMI system.

However, the limitations of hardware-based methods include high implementation costs, frequent maintenance, and the inability to adapt quickly to new theft techniques [22]. These methods can be limited by their dependency on physical infrastructure. Therefore, data-driven ETD methods are regarded as promising solutions for their adaptability, scalability, and efficiency.

2.3. Data-Driven Energy Theft Detection

Data-driven ETD methods have been emphasized in the aim to enhance detection capabilities without the high costs associated with hardware-based approaches via increasing reliance on AI techniques [19].

Data-driven ETD methods can be divided into four categories: data mining, state- and network-based, game theory, and machine learning methods [18]. Within state- and network-based methods, AMI systems play a crucial role in real-time monitoring and detection, offering detailed data and patterns on consumption that can be examined for ETD. In data mining methods, support vector machine (SVM), k-nearest neighbor (KNN), neural networks, and clustering algorithms are predominantly utilized to analyze consumption patterns and to identify irregularities suggestive of energy theft [14]. Game theory methods are employed to model consumer behavior as game players, where their decisions impact the utility losses and gains. By analyzing these interactions, utilities can predict and detect malicious behaviors. In order to identify and mitigate cyberattacks, in which malicious users manipulate smart meter data to misreport higher energy production, machine learning-based ETD models have been reviewed [33]. These models utilize historical solar irradiance, temperature data, and smart meter readings to detect anomalies that indicate potential electricity theft. A pattern-based and context-aware approach for ETD has been proposed by combining dynamic time warping (DTW) and KNN [34]. This method provides a robust ETD framework to detect and manage electricity theft more effectively by considering variations in electricity usage that align with human activities and seasonal changes.

In [21], data-driven ETD methods were emphasized in the aim to prevent financial losses and ensure the reliability and safety of energy distribution networks. These methods encompass supervised and semi-supervised learning-based detection techniques. Moreover, a generative AI-based ETD method was proposed to obtain a balanced dataset for improving data generation and classification performances [11,35]. In [35], a time series generative model and a hybrid, multi-time scale neural network model were developed to capture and analyze consumption patterns at different time scales effectively. However, in cyber-attack-based energy theft, the areas of consumption and generation can be distinguished, with each targeting different functions, thus necessitating the design of tailored models [36,37]. Cyber-attack functions applied in the consumption domain aim to reduce the electric charges for malicious customers; meanwhile, in the generation domain, the goal is to supply more energy to the grid. To provide better detection performance against cyber energy theft, data-driven ETD methods that capture complex patterns and temporal

correlations within generation profiles, data characteristics, and supervisory control and data acquisition (SCADA) meter readings need to be developed.

3. Dataset Issues with Data-Driven ETD

The energy theft data used for ETD refers to energy consumption data, which contain data on normal and malicious users. The energy consumption data are time series data measured from smart meters. The energy theft data, measured at various time intervals depending on the certain requirements of the detection model, can be summarized by four main characteristics.

- High-dimensional data: Given that energy theft data may be influenced by the interaction of numerous variables, including time, weather, and consumer behavior, they can exhibit high-dimensional properties [38]. The high-dimensional data can increase the model complexity and the required computational resources. Therefore, research using energy theft datasets should explore approaches to reduce computational complexity while reflecting high-dimensional characteristics.
- Imbalanced dataset: The datasets used for ETD are often imbalanced, with instances of energy theft being significantly outnumbered by normal energy usage instances [11]. This imbalance can be ascribed, when compared to benign energy consumption patterns, to the challenge of obtaining empirical data on incidents of energy theft and the relatively brief duration during which energy theft typically occurs. An imbalanced dataset poses challenges such as biased learning and overfitting for numerous data-driven algorithms that operate under the assumption of an equitable distribution of classes. Therefore, adequate data preprocessing or advanced algorithms may be required to mitigate these challenges.
- Inaccurate readings: Energy theft data may include inaccuracies and errors from data collection procedures or malfunctioning meters. These discrepancies have the potential to impact the effectiveness of detection models. In order to mitigate the problems caused by inconsistencies, a preprocessing measure could be required before model training is performed.
- Absence of labels: The instances of energy theft may not be accurately labeled, resulting in a label deficiency. This absence of labels can be ascribed to various factors, such as the difficulties in obtaining accurate labels and the time-consuming, costly nature of the labeling process. Such challenges are significant in training supervised machine learning models, which depend on labeled data for learning and making predictions. Comprehension of these properties may be crucial for developing efficient models to detect energy theft.

In order to design effective ETD, several types of datasets can be used to design data-driven ETD. First, the State Grid Corporation of China (SGCC) dataset is the most popular dataset used in energy theft due to offering detailed electricity consumption data with labeled information on energy theft. The Irish Commission for Energy Regulation (CER) dataset, derived from smart metering trials, provides valuable insights into consumer energy behavior and its impact on energy usage. The Electricity Load Diagrams (ELD) dataset is a time-series dataset commonly used for forecasting tasks, encompassing consumer and prosumer information. The Ausgrid Solar Dataset, gathered by Ausgrid in Australia, focuses on electricity consumption and solar power production, offering valuable insights for energy management tasks. Finally, the Uruguayan Electric Company (UTE) dataset provides detailed household electricity consumption data, and it is useful for research on energy consumption patterns and intelligent energy utilization in smart cities. Table 2 shows a summary of the datasets utilized in the research on energy theft.

Table 2. Overview of the dataset used in the field of energy theft.

Dataset	Time Stamp	Duration	Country	Unit	Characteristics		
					Imbalanced	Missing Values	Absence of Labels
SGCC [8,39–43]	1 day	2014.1.~2016.10.	China	kW	○	○	X
CER [29,44–47]	30 min	2009.1.~2010.12.	Ireland	kW	○	○	○
Electricity-Theft [4,48]	15 min	31 days	NA	W	○	X	○
Ausgrid Solar Dataset [44]	30 min	2010.7.~2013.6.	Australia	kW	○	○	○
UTE [49]	1 h	2004.1.~2004.12.	Uruguay	kW	○	X	○
India dataset [50]	15 min	24 h	India	kW	○	X	X
Honduras electricity consumption dataset [51]	1 day	2014.10.~2016.12.	Honduras	kW	○	○	X
UMass smart homes dataset [52]	15 min	2014.5.~2016.2.	United States	kW	○	○	○

3.1. Mitigation for Data Imbalance Problem

In order to address the imbalance problem in energy theft data, several techniques have been developed. Resampling methods can be considered as creating a balanced distribution. Balance can be achieved through over-sampling, where instances from the minority class are replicated, and under-sampling, where instances from the majority class are removed. The random over-sampling (ROS) technique is a simple and effective method where minority class instances are randomly duplicated to balance the dataset [53]. In addition, the synthetic minority class oversampling technique (SMOTE) has been employed to enhance model generalizability by interpolating between the existing instances to generate more diverse samples [54–58].

The adaptive synthetic (ADASYN) sampling technique [59], which is a more advanced technique derived from the SMOTE, generates minority class samples to address data imbalance. Unlike the SMOTE, the ADASYN technique can dynamically adjust the sample creation ratio based on the proximity of each minority class sample to the boundary of majority class samples. In [59], the ADASYN technique was used to design an anomaly transformer (AT) model for ETD. In [60], ADASYN-SGWO was used, which is a method that is designed to alleviate the class imbalance problem by using the ADASYN method for oversampling and the stochastic universal sampling-based grey wolf optimizer (SSO-GWO) for under-sampling [61].

3.2. Correction for the Inaccurate Readings Problem

In order to resolve the problem of missing values in an energy theft dataset, various interpolation methods have been employed, and they can be categorized into two techniques: linear [62] and polynomial [63,64] methods. Based on the simple linear algorithm, zero and average values can be used to replace missing values to recover the energy consumption data over a period [62]. Polynomial interpolation may be effective when the derivatives between data points approach tend to follow certain polynomial expressions [63,64]. Interpolation can be beneficial when there are only a few missing values, but it may lead to substantial information losses in scenarios with abundant missing values. A Bayesian ridge regression-based iterative interpolation method, which estimates missing values using a probabilistic approach, has been proposed to mitigate information loss [65]. Furthermore,

the relationship between missing values and energy theft has been explored, with missing data being used as a feature to enhance detection performance [66]. However, these conventional methods are limited in scope and do not comprehensively solve all the issues arising from the inaccurate data that comes from malfunctioning smart meters. Therefore, advanced methodologies, such as machine learning and deep learning, may be required to address the problems arising from limited storage capacity, disconnected communication, and extreme weather conditions.

3.3. Adversary Modeling for Deficiency Problem

Malicious behavior in power grid systems can lead to operational instability, financial losses, safety risks, data integrity issues, and regulatory problems. The economic stability and functionality of energy distribution networks may be vulnerable to such behaviors, making their consideration crucial in modeling ETD. Although data-driven methods for ETD using datasets with malicious data have been proposed, some datasets still lack this type of data. In order to address the deficiency of malicious data, various studies have been proposed with synthetic data using attack models or functions that replicate malicious behavior. The robustness and accuracy of ETD systems can be significantly enhanced by training models with these synthetic datasets.

3.3.1. Energy Theft in Power Generation

In power generation, energy theft is characterized by manipulating the data relevant to power production facilities or distributed energy resources (DERs). Numerous studies have been conducted on methods for injecting theft into benign data or utilizing data injected into datasets. Several malicious behaviors have been proposed to manipulate photovoltaic systems for modeling photovoltaic electricity theft [67]. According to [67], photovoltaic electricity theft can include voltage boosting, current boosting, altering electrical supply connections, and using solar array simulators. In [68], the performance of ETD has been evaluated by integrating synthetic anomalies simulating energy theft into a real dataset [68]. Herein, these synthetic anomalies encompass full scaling theft, which elevates all data points, and partial scaling theft, which involves adjusting readings to a specific threshold. The authors in [36] developed cyber-attack functions that manipulate the benign data from distributed generation smart meters to simulate electricity theft by malicious customers. A benign dataset was constructed by simulating an IEEE 123-bus test system using practical load and irradiance data, and malicious data were developed by applying the proposed cyber-attack functions. Considering the expansion of renewable energy in the future, modeling new types of energy theft in power generation will be required.

3.3.2. Energy Theft in Power Utility

In power utilities, energy theft can involve manipulating the data associated with the transmission and distribution of electrical power. Investigations have focused on methods that involve the injecting of supply data-oriented theft into benign data or utilizing data that have been injected into datasets. In [69], false data injection attacks on the state estimation of power grids were proposed. The proposed false data injection attacks demonstrated their capability to generate malicious data in state estimation using a standard IEEE test. The authors in [69] performed model simulations using a synthetic dataset that included malicious data to detect stealthy false data injection attacks in state estimation.

Power utilities can exhibit substantial and variable energy usage that are attributed to their operation as large-scale energy systems. Due to the complexity of power utilities, there is a lack of research and available datasets to design ETD in power utilities. Nevertheless, it is essential to examine the modeling of energy theft from the perspective of power utilities due to the potential for significant financial losses compared to other types of energy theft.

3.3.3. Energy Theft in Energy Consumers

With respect to energy consumers, energy theft typically manifests when consumers or prosumers reportedly consume less energy than utilized to diminish their financial obligations. There have been studies on methods to inject demand data-oriented theft into benign datasets or to utilize data injected into datasets. Attack models have been proposed to generate malicious data by manipulating smart meter readings in AMI systems [70]. Energy consumption data can be manipulated with the proposed attack models by drastically reducing recorded consumption or changing load profiles. A different attack model has been proposed to simulate malicious behavior such as meter tampering, bypassing, and malfunctioning meters. According to [71], synthetic data can be employed as a dataset for testing an intermediate monitor meter (IMM)-based power distribution network model. In [72], six theft generation functions were proposed to generate real-time malicious data and to evaluate the performance of a gradient-boosting-based energy theft detector using these synthetic data.

The majority of ETD models have been examined from the viewpoint of energy consumers. Nonetheless, it is anticipated that different forms of intelligent energy theft will emerge in the future, making traditional energy theft modeling approaches inadequate for modeling precise ETD.

4. Methodologies for Implementing Data-Driven ETD

This section introduces several methods for implementing ETD, including supervised learning-based, semi-supervised learning-based, and generative AI-based ETD approaches to improve data efficiency and model performance.

4.1. Supervised Learning-Based Approaches for ETD

Various methodologies have been proposed for ETD based on supervised learning [8,29,39,43,73–76]. When implementing extreme learning machines, support vector machine (SVM)-based models have achieved 70% accuracy [29]. To overcome the problem of numerous false positives with SVM models, a novel method combining SVMs with a decision tree (DT) has been proposed [76]. While detection accuracy has been substantially enhanced by integrating these two methods, the DT is prone to overfit specific patterns, which may reduce its effectiveness in identifying previously unseen attacks.

Deep learning methods have been employed as supervised learning approaches to enhance performance against unseen attacks over artificial feature extraction [8,40,43,73–75]. In [8,43], a structure combining multiple layers of convolutional neural networks (CNNs) with fully connected layers was employed via extracting features from energy theft data, thereby achieving higher accuracy than traditional machine learning methods. Several models have been proposed by employing recurrent neural network (RNN) structures to extract temporal data and enhance classification performance [74,75]. Hyper-parameter optimization with RNNs has been demonstrated to offer superior classification performance compared to conventional SVM models [74]. Subsequent research [75] has shown that efficient time series feature extraction and ETD can be achieved using the gated recurrent unit (GRU).

Methods have been introduced that combine the structures of CNNs and RNNs to improve classification performance by combining their separate feature extractions [39,73]. A hybrid deep learning model was demonstrated to enhance, when compared to using CNNs and LSTM independently, feature extraction by cascading a CNN with an RNN-based long short-term memory (LSTM) [39]. However, this architecture may have an inherent limitation in feature extraction capabilities as it relies on transferring features from the CNN to LSTM. ConvLSTM architecture has been introduced to address this inherent limitation, which replaces the matrix multiplication of the traditional CNN-LSTM stack with a globally connected layer. This results in a model that efficiently captures cyclical patterns and extracts local features.

Figure 5 depicts the representative structures of supervised learning methods for ETD. Figure 5a shows the CNN process with two separate convolutional layers to extract spatial features. In Figure 5b, the capability of RNN for processing data sequences is described by incorporating the contextual information from previous inputs. The processing for CNN with LSTM to extract integrated features is described in Figure 5c.

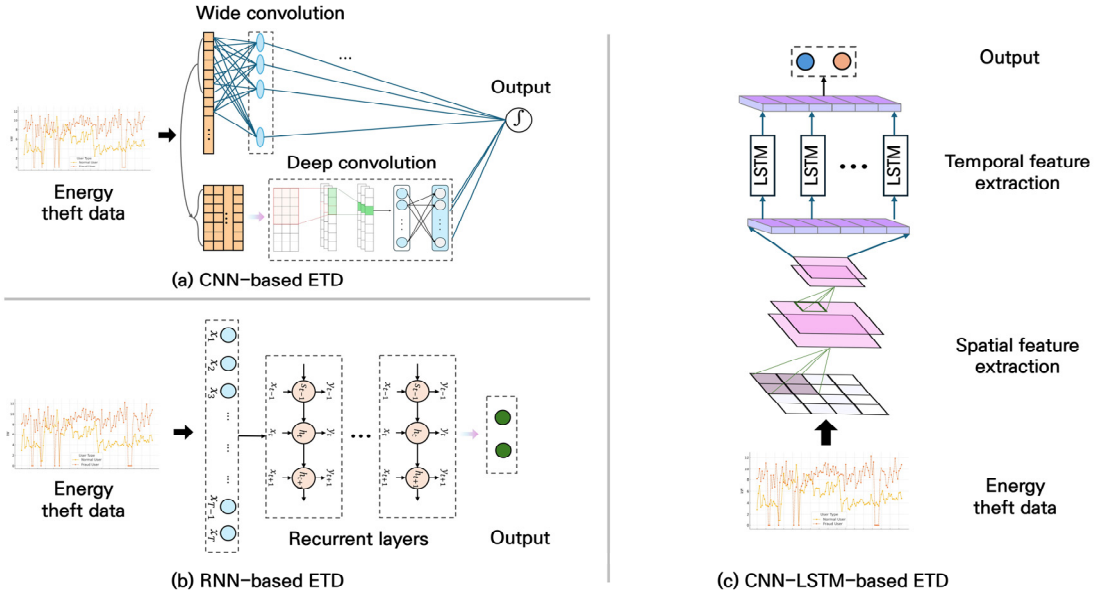


Figure 5. Architectures of representative supervised learning-based ETD.

4.2. Semi-Supervised Learning-Based Approaches for ETD

As mentioned in the previous subsection, several data-driven approaches for ETD have been proposed based on supervised learning. However, the uneven distribution of labeled data in ETD can lead to a decline in model performance due to biased learning effects. As alternatives, unsupervised learning methods have been proposed [77,78]. These conventional approaches can involve using clustering techniques in combination with the maximum information coefficient (MIC) [77] or using density-based spatial clustering [78] to identify abnormal behavior. However, the conventional unsupervised learning approaches may struggle with high-dimensional noisy data. Consequently, semi-supervised learning approaches have been developed to address the limitations of both supervised and unsupervised methods.

Semi-supervised learning methods have been developed to leverage the benefits of both supervised and unsupervised learning for efficient ETD. A transductive SVM (TSVM) method [79] has been utilized as a semi-supervised learning for ETD, but it may encounter challenges in scaling when confronted with large volumes of data. In deep learning-based semi-supervised learning for ETD, two primary approaches can be employed: (1) augmenting data by assigning pseudo-labels [80,81] and (2) integrating supervised and unsupervised learning [82,83]. In the first approach, a model can assign pseudo-labels to unlabeled data, reducing overfitting and improving model generalization [80,81]. In the second approach, networks trained on unsupervised and supervised learning tasks are used to reconstruct load profiles and differentiate between classes [82,83].

The structures of semi-supervised learning methods are described in Figure 6. Figure 6a depicts a framework that employs a trained teacher model to predict unlabeled data, generating pseudo-labels for augmenting data for ETD; the framework then proceeds to train the classifier from these augmented data. Figure 6b describes a model integration approach that combines an unsupervised learning framework using an autoencoder, thereby using similarity learning through a Siamese network and supervised learning using labeled data.

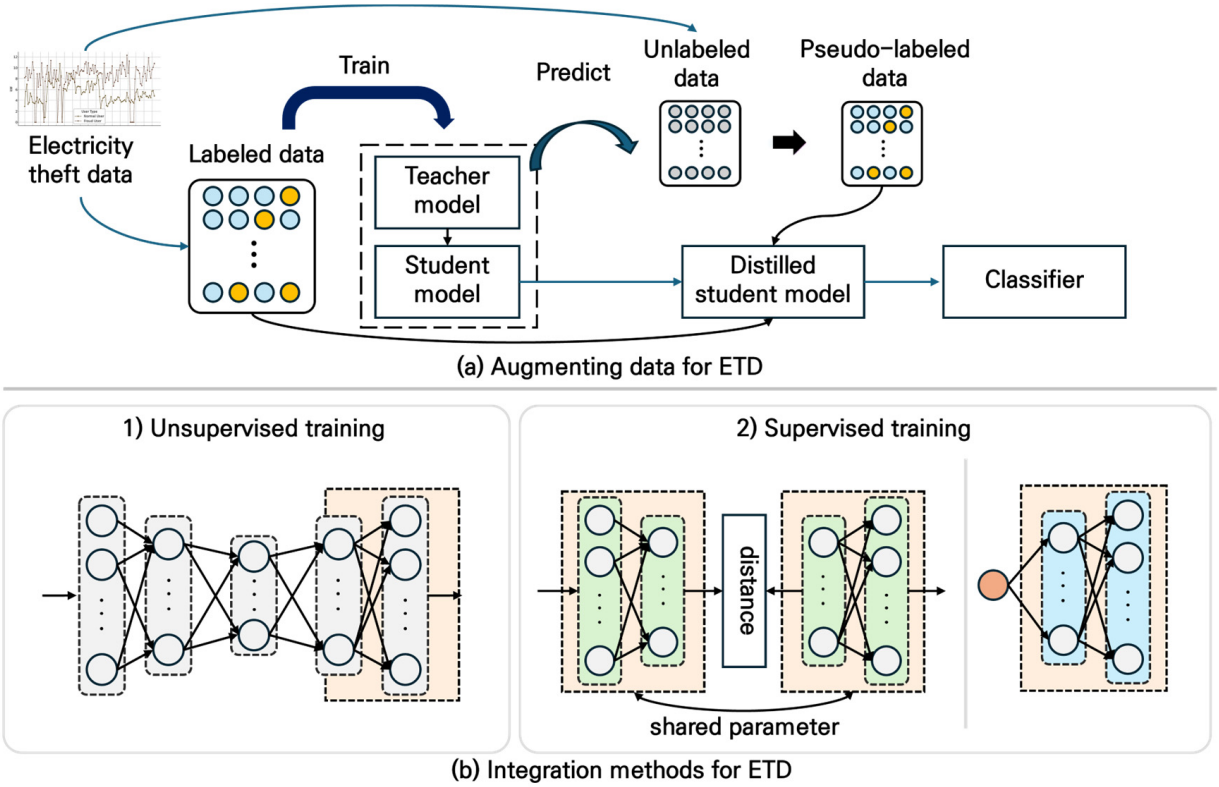


Figure 6. Architectures of representative semi-supervised learning-based ETD.

4.3. Generative AI-Based Approaches for ETD

While semi-supervised approaches have shown improvements in specific datasets, their effectiveness in practical systems may be compromised by the complexity of theft methods that are not represented in training data. In addition, when a specific label is extremely scarce, the correlation in the augmented data may become excessively high, thus requiring further adjustments to enhance the performance of the model. Generative AI-based approaches can be used to enhance the performance of model generalization. However, their application in ETD has been relatively limited [12]. These techniques can generate new data by analyzing patterns in the existing data, and they may effectively address unseen attacks when applied to ETD.

Generative AI methods for ETD can be classified into probabilistic, direct distribution approximation, and diffusion-based methods. Probabilistic methods, such as the variational autoencoder (VAE) [11,12], can extract information from a latent space within the data. VAEs have been utilized as a data augmentation method in ETD [11,12] since they offer more reliable training than GANs. In [12], a conditional VAE (CVAE) was applied, demonstrating the capability to generate samples resembling the original power curve through using only a few samples without assuming a probability distribution of the power curve. Furthermore, it has been confirmed that combining VAE and GANs [11] can improve classification accuracy compared to separately applying each VAE- and GAN-based method.

Unlike VAE, which assumes a prior data distribution such as a standard Gaussian distribution, direct distribution approximation techniques leverage generative adversarial networks (GANs) to produce new data. A cooperative training GAN (CT-GAN) method has been proposed to address the challenge of obtaining labeled data for ETD [13]. The CT-GAN method can enhance training stability by training two discriminators. This approach improves the generation of labeled sample data and increases the accuracy of semi-supervised classification. According to simulation results [13], CT-GAN substantially enhances generalization capabilities.

Diffusion-based methods [4] have focused on identifying complex data distributions and generating new data by progressively reducing noise. The diffusion method has been proposed to capture patterns for data sequences with intricate patterns by introducing noise to the data samples [4]. The inverse process is employed to gradually eliminate noise and generate new samples, permitting the neural network to acquire information regarding the data distribution. It has been demonstrated that diffusion-based detectors can outperform alternative methods, such as LSTM or AE-based methods, in identifying diverse user patterns.

In order to effectively address the 5 Vs of big data, generative AI-based ETD can be considered a promising solution, but it presents several challenges. Generative AI models can be trained on large volumes of datasets to detect specific patterns that are indicative of energy theft. However, managing and processing such vast amounts of data requires significant computational resources and efficient data-handling techniques to avoid performance bottlenecks. Additionally, if there are various types of data, such as time-series data, customer profiles, and transaction records, a generative AI model can comprehensively learn from this diverse information. Nonetheless, the integration and normalization of heterogeneous data sources can be challenging and may require advanced data fusion to ensure compatibility and reliability. The data from smart grids may be noisy or incomplete, posing a challenge to maintaining data integrity. Therefore, advanced data processing and validation techniques are necessary to mitigate the impact of poor data quality on detection performance. The goal of processing big data in generative AI-based ETD is to extract valuable insights that can lead to actionable outcomes. Generative AI-based ETD models with high-value data can uncover hidden patterns and anomalies that conventional methods might miss. In Table 3, the 5 Vs of big data in generative AI-based ETD are summarized.

Table 3. The 5 Vs of big data, and the issues of generative AI-based ETD.

5 Vs of Big Data	Description	Issues of Generative AI-Based ETD
Velocity	The speed at which data are generated and processed	- Necessity for real-time data processing - Rapid analysis for immediate detection of energy theft
Volume	The scale of the data	- Capability to handle large-scale datasets - Data management for training and analysis
Value	The usefulness of the data being analyzed	- High-quality data for discovering meaningful patterns - Maximizing data value for accurate theft detection
Variety	The different types of data available	- Learning from various data types - Integration of diverse information such as time series, profiles, and records
Veracity	The quality and accuracy of the data	- Maintaining data integrity - Need for advanced processing and validation techniques to address data quality issues

The structures of generative AI methods for ETD are detailed in Figure 7. According to Figure 7a, the data generated by the VAE model can be used to improve the learning performance of a separate ETD classifier module. Conversely, Figure 7b,c represent that the classifier is concurrently trained with the generative model in the structures of GANs and diffusion-based method. Additionally, LSTM algorithms are employed to extract temporal features from energy theft data and improve the performance of denoising diffusion probabilistic models (DDPM).

In Table 4, the background and limitations of the proposed model are summarized with a focus on methods for ETD model design.

Table 4. The main existing works using data-driven ETD approaches.

Ref.	Approach	Proposed Method	Motivation	Limitation
[8]	Supervised learning	CNN	- A proposal for an ETD based on 2D electricity consumption data	- A model optimized for only a specific dataset (SGCC)
[29]		SVM	- A proposal for an ETD based on user-centric energy usage data	- Insufficient diversity of collected energy usage data
[39,73]		ConvLSTM	- A proposal for an ETD that combines the strengths of CNN and LSTM models	- Preprocessing is required to convert the data into 2D form
[74]		RNN	- A proposal for an ETD tailored to individual customers	- Lack of verification of applicability in an actual environment
[75]		Deep RNN and GRU	- A proposal for a hyper-parameter tuning approach for the performance improvement of an ETD	- Fine-tuning tasks according to the dataset
[76]		Random forest and SVM	- A proposal for an ETD to minimize false alarms (false positives)	- Inefficiency in large-scale datasets due to model complexity
[37]	Semi-Supervised learning	GRU-RNN	- A proposal for an ETD to detect maliciously manipulated smart meter values	- Decreased performance in the model due to sensing errors in the smart meter
[41]		Autoencoder	- A proposal for an ETD to reconstruct normal patterns in the energy theft dataset	- Poor performance on the scarcity of the electricity theft dataset
[77]		MIC	- A proposal for an ETD based on a correlation analysis of energy theft	- Insufficient generalization of the proposed model
[78]		Density-based spatial clustering	- A proposal for an ETD with guaranteed generalization performance	- Lack of validation for large-scale datasets
[79]		TSVM	- A proposal for an ETD to overcome overfitting	- High complexity of the proposed model
[80]		Relational denoising autoencoder	- A proposal for an ETD to prevent loss of data relationships during feature extraction	- Computational complexity in data preprocessing
[82]		Multi-task feature extracting fraud detector	- A proposal for an ETD to handle a high-dimensional dataset	- High complexity of the proposed model
[83]		Semi-supervised Autoencoder	- A proposal for an ETD to overcome overfitting	- Difficulty analyzing results due to complex model structure
[84]		Autoencoder with LSTM	- A proposal for an ETD model based on semi-supervised learning using labeled and unlabeled data	- Lack of verification of applicability in an actual environment
[4]	Generative AI	Diffusion-based LSTM	- A proposal for a robust ETD model for an energy theft dataset with various variances	- Conducting tests within an experimental setting
[11]		CNN with VAE-GAN	- A proposal for an ETD model to mitigate the performance degradation of models caused by data imbalance issues	- Insufficient verification of the generalization performance
[12]		Conditional Autoencoder	- A proposal for a reliable ETD model with data augmentation	- Lack of diversity due to limited adversary modeling
[13]		CT (Cooperative Training)-GAN	- A proposal for an ETD model to overcome the limitations of a scarce labeled dataset	- The degradation of model performance due to the hallucination issue of the GAN

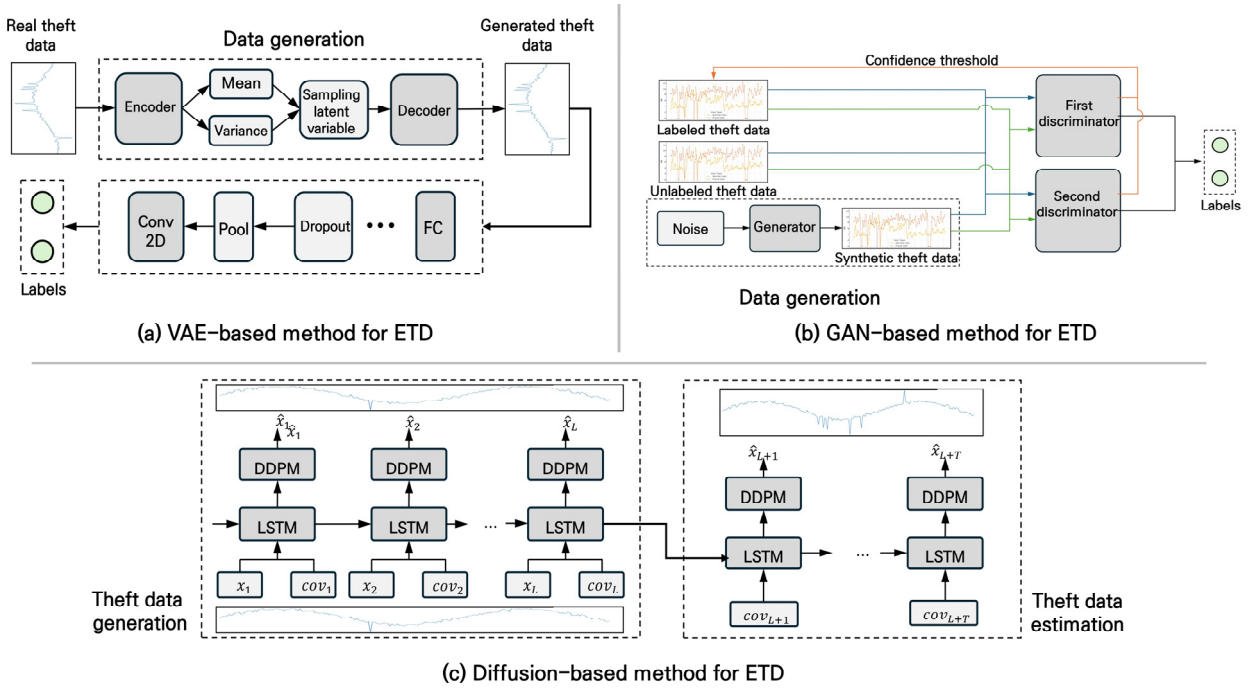


Figure 7. Architectures of representative generative AI-based ETD.

5. Open Issues and Future Research Directions

Given the complex nature of energy theft datasets, the current supervised and semi-supervised learning-based models face limitations in effectively analyzing such data characteristics. In order to address this challenge, recent research has explored the potential of generative AI-based ETD. Various significant advantages are expected through the implementation of generative AI-based ETD. Firstly, generative AI-based ETD possesses the capability to analyze large volumes of data. This ability allows generative AI to detect anomalies in energy consumption patterns and identify instances of energy theft or illegal energy acquisition. Secondly, the insights provided by generative AI into energy consumption patterns can enhance energy management systems, facilitating the optimization of energy usage and minimizing energy loss. These capabilities are instrumental in detecting and preventing energy theft effectively. Lastly, generative AI-based ETD contributes to improved system reliability by mitigating risks of system failures, reducing downtime, and ensuring a reliable power supply for customers. Despite these advantages, generative AI-based ETD is still in the early stages of development. Therefore, open issues and research directions focusing on generative AI in ETD will be discussed in this section.

5.1. Handling Imbalanced Data

By generating synthetic instances of the minority class, generative AI, such as GANs and VAEs, have shown promise in handling imbalanced datasets. However, challenges remain to be addressed since the generated instances cannot replace the real instances completely. For instance, ensuring the quality and diversity of the generated samples and avoiding overfitting to the minority class are critical issues. Furthermore, a combination of various proportions of real and synthetic samples may be advantageous to enhance the diversity of the training samples and the performance of the detectors. Therefore, future research could focus on developing novel generative AI or improving existing ones to generate synthetic samples and efficiently handle imbalanced data in ETD.

5.2. Incorporating Time-Series Analysis with Data Features

In order to capture temporal patterns in time-series data, generative AI has been employed in time-series analysis for the past few years. However, accurately modeling

and generating time-series data remains challenging due to its high dimensionality, noise, and complex temporal dependencies. Transformers with self-supervised representational learning ability can be a promising candidate model for challenging tasks. Future research could explore advanced generative AI using transformers for time-series data, as well as their application in ETD.

5.3. Dealing with High-Dimensional Data

Deep learning-based generative AI models have recently exhibited the capacity to handle high-dimensional data. However, training these models can be computationally expensive and may require large amounts of data. Future work could investigate efficient training methods for generative AI models in ETD, such as meta learning and transfer learning [85], and dimensionality reduction techniques, like singular value decomposition and diffusion maps [86].

5.4. Addressing Noise and Errors

In order to remove the noise from data and rectify errors, generative AI models can be utilized. The generative AI employing unsupervised or semi-supervised learning may denoise and correct errors of the energy theft data. However, ensuring that the denoising or error correction does not result in losing important information is challenging. Future research should aim to develop robust generative AI that can handle noise and errors in energy theft data by using unsupervised learning and semi-supervised learning approaches.

5.5. Exploiting Characteristic Variables

Generative AI models can potentially learn and generate characteristic variables of energy consumption records. However, ensuring that these generated variables are meaningful and useful for ETD is challenging. In order to overcome this challenge, one should look to information bottleneck-based approaches, which could be promising since they can enhance domain generalization and improve the performance of generative AI. Future work could focus on exploiting these variables, incorporating generative AI models with an information bottleneck-based approach for more accurate ETD.

5.6. Overcoming a Lack of Labels

Non-adversarial generative AI, particularly diffusion models, have recently gained significant interest. Diffusion models typically involve a forward process that gradually corrupts the input by an added noise and a reverse process that reconstructs them sequentially to learn the distribution of the latent representation of the input. The models can be more stable, and they can model small datasets more effectively. Diffusion models can be adequately employed to overcome the lack of labels in ETD. However, ensuring that these models can effectively learn from unlabeled data and make accurate predictions is a challenge. Research for ETD models that exploit the properties of non-adversarial generative AI could be an area of future work for generating reliable datasets in ETD.

5.7. Integration of Energy Consumption and Multimodal Data

There is a strong possibility that the future of generative AI models in detecting energy theft will involve an important shift toward using several types of data sources. Generative AI models can enhance the comprehension of complex patterns by integrating energy consumption data with textual, auditory, visual, and sensor-based information. Besides smart meter and/or energy consumption data, other types of data such as climatic and meta data may be exploited in detecting energy theft. Climatic data encompasses variables such as temperature, wind speed, humidity, etc., which can be provided in textual, audio, or visual formats. The meta data may drop a hint by indicating the device type, customer features, and region characteristics. By adopting a multimodal approach, ETD approaches can improve their ability to identify subtle anomalies and forecast instances accurately. This capability for ETD can be achieved by utilizing the combined capabilities of diverse

data sources, which also automates the validation process. Future research should aim to improve the performance of ETD by integrating energy consumption values with diverse data sources.

5.8. Large Language Models for ETD

Large language models (LLMs), an advancement of generative AI models, have been recently marked in a variety of fields due to their significant potential in analyzing comprehensive datasets to find patterns, forecast future events, and detect abnormal behavior across different domains [87]. For example, anomaly detection systems can discover uncommon access patterns that could indicate a cybersecurity security breach. Likewise, in ETD, LLMs may identify the energy thieves in energy systems. Future work could explore applying LLMs in ETD to achieve optimal performance.

6. Conclusions

In this paper, a comprehensive review of data-driven approaches for ETD is presented, focusing on methodologies such as datasets, preprocessing, adversary modeling, and detection algorithms. It has also underscored the need for data-driven ETD analysis in terms of the 5 Vs of big data: velocity, volume, value, variety, and veracity. In this regard, the limitations of the energy theft dataset and previous studies in overcoming limitations were analyzed and systematically organized. Then, various detection methods, including supervised learning, semi-supervised learning, and generative AI-based approaches, were analyzed to implement effective ETD models. The limitations of existing data-driven ETD, such as supervised learning and semi-supervised learning, were analyzed, and the potential of generative AI to revolutionize ETD was highlighted. Finally, this paper suggests future research directions in applicable generative AI for addressing imbalanced data, incorporating time-series analysis with data features, dealing with high-dimensional data, addressing noise and errors, exploiting characteristic variables, overcoming the lack of labels, utilizing LLMs for ETD, and integrating energy consumption data with multimodal data.

Author Contributions: Conceptualization, S.K., S.L. and J.S.; methodology, S.K., S.L. and J.S.; formal analysis, S.K., Y.S., B.H., J.K. (Jeongho Kim) and J.K. (Jinwook Kim); investigation, S.K., Y.S., B.H., J.K. (Jeongho Kim) and J.K. (Jinwook Kim); resources, S.K., Y.S., S.L. and J.S.; writing—original draft preparation, S.K., S.L. and J.S.; writing—review and editing, S.K., Y.S., S.L., J.S., B.H., J.K. (Jeongho Kim), J.K. (Jinwook Kim) and K.K.; visualization, S.K., S.L., J.S., B.H., J.K. (Jeongho Kim), J.K. (Jinwook Kim) and K.K.; supervision, J.K. (Jinyoung Kim); project administration, J.K. (Jinyoung Kim). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This work was supported by the Technology Development Program (RS-2023-00261975) funded by the Ministry of SMEs and Startups (MSS, Korea); in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-2020-0-01846), which is supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

Conflicts of Interest: Author Youngghyu Sun was employed by the company SMART EVER, Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Tan, S.; De, D.; Song, W.-Z.; Yang, J.; Das, S.K. Survey of security advances in smart grid: A data driven approach. *IEEE Commun. Surv. Tutor.* **2017**, *19*, 397–422. [CrossRef]
2. Athanasiadis, C.L.; Papadopoulos, T.A.; Kryonidis, D.I.; Doukas, D.I. A review of distribution network applications based on smart meter data analytics. *Renew. Sustain. Energy Rev.* **2024**, *191*, 114151. [CrossRef]
3. Carr, D.; Thomson, M. Non-technical electricity losses. *Energies* **2022**, *15*, 2218. [CrossRef]
4. Yuan, X.; Yang, Y.; Iqbal, A.; Gope, P.; Sikdar, B. A novel DDPM-based ensemble approach for energy theft detection in smart grids. *arXiv* **2024**, arXiv:2307.16149. Available online: https://arxiv.org/abs/2307.16149 (accessed on 1 May 2024).

5. Theron-Ord, A. Electricity Theft and Non-Technical Losses Total $96bn Annually—Report. Available online: https://www.smart-energy.com/regional-news/africa-middle-east/electricity-theft-96bn-annually/ (accessed on 7 May 2024).
6. Zulu, C.L.; Dzobo, O. Real-time power theft monitoring and detection system with double connected data capture system. *Electr. Eng.* **2023**, *105*, 3065–3083. [CrossRef]
7. Singh, N.; Singh, D.P.; Pant, B. A comprehensive study of big data machine learning approaches and challenges. In Proceedings of the 2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS), Jammu, India, 11–12 December 2017; pp. 80–85.
8. Zheng, Z.; Yang, Y.; Niu, X.; Dai, H.-N.; Zhou, Y. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. *IEEE Trans. Ind. Inf.* **2018**, *14*, 1606–1615. [CrossRef]
9. Gunduz, M.Z.; Das, R. Smart Grid Security: An effective hybrid CNN-based approach for detecting energy theft using consumption patterns. *Sensors* **2024**, *24*, 1148. [CrossRef]
10. Chen, W.; Yang, K.; Yu, Z.; Shi, Y.; Chen, C.L.P. A survey on imbalanced learning: Latest research, applications and future directions. *Artif. Intell. Rev.* **2024**, *57*, 137. [CrossRef]
11. Sun, Y.; Lee, J.; Kim, S.; Seon, J.; Lee, S.; Kyeong, C.; Kim, J. Energy theft detection model based on VAE-GAN for imbalanced dataset. *Energies* **2023**, *16*, 1109. [CrossRef]
12. Gong, X.; Tang, B.; Zhu, R.; Liao, W.; Song, L. Data augmentation for electricity theft detection using conditional variational auto-encoder. *Energies* **2020**, *13*, 4291. [CrossRef]
13. Xia, R.; Wang, J. A semi-supervised learning method for electricity theft detection based on CT-GAN. In Proceedings of the 2022 IEEE International Conference on Power Systems and Electrical Technology (PSET), Aalborg, Denmark, 13–15 October 2022; pp. 335–340.
14. Ohno, H. Training data augmentation using generative models with statistical guarantees for materials informatics. *Soft Comput.* **2022**, *26*, 1181–1196. [CrossRef]
15. Shivashankar, C.; Miller, S. Semantic data augmentation with generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Luxembourg, 11–15 September 2023; pp. 863–873.
16. Viegas, J.L.; Esteves, P.R.; Melício, R.; Mendes, V.M.F.; Vieira, S.M. Solutions for detection of non-technical losses in the electricity grid: A review. *Renew. Sustain. Energy Rev.* **2017**, *80*, 1256–1268. [CrossRef]
17. Xia, X.; Xiao, Y.; Liang, W.; Cui, J. Detection methods in smart meters for electricity thefts: A survey. *Proc. IEEE* **2022**, *110*, 273–319. [CrossRef]
18. Ahmed, M.; Khan, A.; Ahmed, M.; Tahir, M.; Jeon, G.; Fortino, G.; Piccialli, F. Energy theft detection in smart grids: Taxonomy, comparative analysis, challenges, and future research directions. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 578–600. [CrossRef]
19. Guarda, F.; Hammerschmitt, B.; Capeletti, M.; Neto, N.; Dos Santos, L.; Prade, L.; Abaide, A. Non-hardware-based non-technical losses detection methods: A review. *Energies* **2023**, *16*, 2054. [CrossRef]
20. Kgaphola, P.M.; Marebane, S.M.; Hans, R.T. Electricity theft detection and prevention using technology-based models: A systematic literature review. *Electricity* **2024**, *5*, 334–350. [CrossRef]
21. Althobaiti, A.; Jindal, A.; Marnerides, A.K.; Roedig, U. Energy theft in smart grids: A survey on data-driven attack strategies and detection methods. *IEEE Access* **2021**, *9*, 159291–159312. [CrossRef]
22. Badr, M.; Ibrahem, M.; Kholidy, H.; Fouda, M.; Ismail, M. Review of the data-driven methods for electricity fraud detection in smart metering systems. *Energies* **2023**, *16*, 2852. [CrossRef]
23. Stracqualursi, E.; Rosato, A.; Di Lorenzo, G.; Panella, M.; Araneo, R. Systematic review of energy theft practices and autonomous detection through artificial intelligence methods. *Renew. Sustain. Energy Rev.* **2023**, *184*, 113544. [CrossRef]
24. Depuru, S.S.S.R.; Wang, L.; Devabhaktuni, V. Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft. *Energy Policy* **2011**, *39*, 1007–1015. [CrossRef]
25. Lewis, F.B. Costly 'Throw-Ups': Electricity theft and power disruptions. *Electr. J.* **2015**, *28*, 118–135. [CrossRef]
26. Czechowski, R.; Kosek, A.M. The most frequent energy theft techniques and hazards in present power energy consumption. In Proceedings of the 2016 Joint Workshop on Cyber-Physical Security and Resilience in Smart Grids (CPSR-SG), Vienna, Austria, 12 April 2016; pp. 1–7.
27. Grewal, R.; Sharma, T.; Mourya, R.; Kumar, A.; Kaur, K. Cost effective overload and theft detection for power distribution system. In Proceedings of the 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 18–19 May 2018; pp. 450–455.
28. Shokry, M.; Awad, A.I.; Abd-Ellah, M.K.; Khalaf, A.A.M. Systematic survey of advanced metering infrastructure security: Vulnerabilities, attacks, countermeasures, and future vision. *Future Gener. Comput. Syst.* **2022**, *136*, 358–377. [CrossRef]
29. Jokar, P.; Arianpoo, N.; Leung, V.C.M. Electricity theft detection in AMI using customers' consumption patterns. *IEEE Trans. Smart Grid* **2016**, *7*, 216–226. [CrossRef]
30. Leite, J.B.; Mantovani, J.R.S. Detecting and locating non-technical losses in modern distribution networks. *IEEE Trans. Smart Grid* **2018**, *9*, 1023–1032. [CrossRef]
31. Guan, L.; Cong, X.; Zhang, Q.; Liu, F.; Gao, Y.; An, W.; Noureldin, A. A comprehensive review of micro-inertial measurement unit based intelligent PIG multi-sensor fusion technologies for small-diameter pipeline surveying. *Micromachines* **2020**, *11*, 840. [CrossRef] [PubMed]

32. Peng, Y.; Yang, Y.; Xu, Y.; Xue, Y.; Song, R.; Kang, J.; Zhao, H. Electricity theft detection in AMI based on clustering and local outlier factor. *IEEE Access* **2021**, *9*, 107250–107259. [CrossRef]
33. Shaaban, M.; Tariq, U.; Ismail, M.; Almadani, N.; Ahmed, M. Data-driven detection of electricity theft cyberattacks in PV generation. *IEEE Syst. J.* **2022**, *16*, 3349–3359. [CrossRef]
34. Ahir, R.K.; Chakraborty, B. Pattern-based and context-aware electricity theft detection in smart grid. *Sustain. Energy Grids Netw.* **2022**, *32*, 100833. [CrossRef]
35. Sun, Y.; Sun, X.; Hu, T.; Zhu, L. Smart grid theft detection based on hybrid multi-time scale neural network. *Appl. Sci.* **2023**, *13*, 5710. [CrossRef]
36. Ismail, M.; Shaaban, M.F.; Naidu, M.; Serpedin, E. Deep learning detection of electricity theft cyber-attacks in renewable distributed generation. *IEEE Trans. Smart Grid* **2020**, *11*, 3428–3437. [CrossRef]
37. Ezeddin, M.; Albaseer, A.; Abdallah, M.; Bayhan, S.; Qaraqe, M.; Al-Kuwari, S. Efficient deep learning based detector for electricity theft generation system attacks in smart grid. In Proceedings of the 2022 3rd International Conference on Smart Grid and Renewable Energy (SGRE), Doha, Qatar, 20–22 March 2022; pp. 1–6.
38. Pan, H.; Yin, Z.; Jiang, X. High-dimensional energy consumption anomaly detection: A deep learning-based method for detecting anomalies. *Energies* **2022**, *15*, 6139. [CrossRef]
39. Hasan, M.N.; Toma, R.N.; Nahid, A.-A.; Islam, M.M.M.; Kim, J.-M. Electricity theft detection in smart grid systems: A CNN-LSTM based approach. *Energies* **2019**, *12*, 3310. [CrossRef]
40. Ullah, A.; Javaid, N.; Samuel, O.; Imran, M.; Shoaib, M. CNN and GRU based deep neural network for electricity theft detection to secure smart grid. In Proceedings of the 2020 International Wireless Communications and Mobile Computing (IWCMC), Limassol, Cyprus, 15–19 June 2020; pp. 1598–1602.
41. Takiddin, A.; Ismail, M.; Zafar, U.; Serpedin, E. Deep autoencoder-based anomaly detection of electricity theft cyberattacks in smart grids. *IEEE Syst. J.* **2022**, *16*, 4106–4117. [CrossRef]
42. Takiddin, A.; Ismail, M.; Nabil, M.; Mahmoud, M.M.E.A.; Serpedin, E. Detecting electricity theft cyber-attacks in AMI networks using deep vector embeddings. *IEEE Syst. J.* **2021**, *15*, 4189–4198. [CrossRef]
43. Yao, D.; Wen, M.; Liang, X.; Fu, Z.; Zhang, K.; Yang, B. Energy theft detection with energy privacy preservation in the smart grid. *IEEE Internet Things J.* **2019**, *6*, 7659–7669. [CrossRef]
44. Krishna, V.B.; Gunter, C.A.; Sanders, W.H. Evaluating detectors on optimal attack vectors that enable electricity theft and DER Fraud. *IEEE J. Sel. Top. Signal Process.* **2018**, *12*, 790–805. [CrossRef]
45. Krishna, V.B.; Lee, K.; Weaver, G.A.; Iyer, R.K.; Sanders, W.H. F-DETA: A framework for detecting electricity theft attacks in smart grids. In Proceedings of the 2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Toulouse, France, 8 June–1 July 2016; pp. 407–418.
46. Takiddin, A.; Ismail, M.; Zafar, U.; Serpedin, E. Robust electricity theft detection against data poisoning attacks in smart grids. *IEEE Trans. Smart Grid* **2021**, *12*, 2675–2684. [CrossRef]
47. Li, S.; Han, Y.; Yao, X.; Yingchen, S.; Wang, J.; Zhao, Q. Electricity theft detection in power grids with deep learning and random forests. *J. Electr. Comput. Eng.* **2019**, *2019*, 1–12. [CrossRef]
48. Alromih, A.; Clark, J.A.; Gope, P. Electricity theft detection in the presence of prosumers using a cluster-based multi-feature detection model. In Proceedings of the 2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Aachen, Germany, 25–28 October 2021; pp. 339–345.
49. Martino, M.D.; Decia, F.; Molinelli, J.; Fernández, A. Improving electric fraud detection using class imbalance strategies. In Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods (ICPRAM), Algarve, Portugal, 6–8 February 2012; pp. 135–141.
50. Depuru, S.S.S.R.; Wang, L.; Devabhaktuni, V. Support vector machine based data classification for detection of electricity theft. In Proceedings of the 2011 IEEE/PES Power Systems Conference and Exposition, Phoenix, AZ, USA, 20–23 March 2011; pp. 1–8.
51. Figueroa, G.; Chen, Y.-S.; Avila, N.; Chu, C.-C. Improved practices in machine learning algorithms for NTL detection with imbalanced data. In Proceedings of the 2017 IEEE Power & Energy Society General Meeting, Chicago, IL, USA, 16–20 July 2017; pp. 1–5.
52. Mujeeb, S.; Javaid, N.; Ahmed, A.; Gulfam, S.M.; Qasim, U.; Shafiq, M.; Choi, J.-G. Electricity theft detection with automatic labeling and enhanced RUSBoost classification using differential evolution and jaya algorithm. *IEEE Access* **2021**, *9*, 128521–128539. [CrossRef]
53. Yap, B.W.; Rani, K.A.; Rahman, H.A.A.; Fong, S.; Khairudin, Z.; Abdullah, N.N. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), Kuala Lumpur, Malaysia, 16–18 December 2013; pp. 13–22.
54. Khan, Z.A.; Adil, M.; Javaid, N.; Saqib, M.N.; Shafiq, M.; Choi, J.-G. Electricity theft detection using supervised learning techniques on smart meter data. *Sustainability* **2020**, *12*, 8023. [CrossRef]
55. Syed, D.; Abu-Rub, H.; Refaat, S.S.; Xie, L. Detection of energy theft in smart grids using electricity consumption patterns. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 4059–4064.

56. Maraden, Y.; Wibisono, G.; Nugraha, I.G.D.; Sudiarto, B.; Jufri, F.H.; Kazutaka, K.; Prabuwono, A.S. 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* **2023**, *16*, 5405. [CrossRef]
57. Qu, Z.; Li, H.; Wang, Y.; Zhang, J.; Abu-Siada, A.; Yao, Y. Detection of electricity theft behavior based on improved synthetic minority oversampling technique and random forest classifier. *Energies* **2020**, *13*, 2039. [CrossRef]
58. Wang, J.; Zhang, X. Electricity theft detection based on SMOTE oversampling and logistic regression classifier. In Proceedings of the 2023 IEEE 6th International Electrical and Energy Conference (CIEEC), Hefei, China, 2–14 May 2023; pp. 2571–2576.
59. Chen, S.; Yang, Y.; You, S.; Chen, W.; Li, Z. A study of electricity theft detection method based on anomaly transformer. In Proceedings of the Big Data, Sorrento, Italy, 15–18 December 2023; pp. 164–180.
60. Tripathi, A.K.; Pandey, A.C.; Sharma, N. A new electricity theft detection method using hybrid adaptive sampling and pipeline machine learning. *Multimed. Tools Appl.* **2023**, *83*, 54521–54544. [CrossRef]
61. Pereira, J.; Saraiva, F. A comparative analysis of unbalanced data handling techniques for machine learning algorithms to electricity theft detection. In Proceedings of the 2020 IEEE Congress on Evolutionary Computation (CEC), Glasgow, UK, 19–24 July 2020; pp. 1–8.
62. Petrlik, I.; Lezama, P.; Rodriguez, C.; Inquilla, R.; Reyna-González, J.E.; Esparza, R. Electricity theft detection using machine learning. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 420–425. [CrossRef]
63. Lepolesa, L.J.; Achari, S.; Cheng, L. Electricity theft detection in smart grids based on deep neural network. *IEEE Access* **2022**, *10*, 39638–39655. [CrossRef]
64. Rahimi, A.; Shahrestani, A.; Ramezani, S.; Zamani, P.; Tehrani, S.O.; Moghaddam, M.H.Y. Filter based time-series anomaly detection in AMI using AI approaches. In Proceedings of the 2021 5th International Conference on Internet of Things and Applications (IoT), Isfahan, Iran, 19–20 May 2021; pp. 1–6.
65. Huang, L.; Qin, H.; Pan, Z.; Yu, M. Electricity theft detection based on iterative interpolation and fusion convolutional neural network. In Proceedings of the 2022 7th International Conference on Power and Renewable Energy (ICPRE), Shanghai, China, 23–26 September 2022; pp. 567–571.
66. Fei, K.; Li, Q.; Zhu, C. Non-technical losses detection using missing values' pattern and neural architecture search. *Int. J. Electr. Power Energy Syst.* **2022**, *134*, 107410. [CrossRef]
67. Yuan, X.; Shi, M.; Sun, Z. Research status of electricity-stealing identification technology for distributed PV. In Proceedings of the 2015 5th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT), Changsha, China, 26–29 January 2015; pp. 2031–2034.
68. Althobaiti, A.; Jindal, A.; Marnerides, A.K. Data-driven energy theft detection in modern power grids. In Proceedings of the Twelfth ACM International Conference on Future Energy Systems, Virtual, 28 June–2 July 2021; pp. 39–48.
69. Esmalifalak, M.; Liu, L.; Nguyen, N.; Zheng, R.; Han, Z. Detecting stealthy false data injection using machine learning in smart grid. *IEEE Syst. J.* **2017**, *11*, 1644–1652. [CrossRef]
70. Singh, S.K.; Bose, R.; Joshi, A. Energy theft detection for AMI using principal component analysis based reconstructed data. *IET Cyber-Phys. Syst. Theory Appl.* **2019**, *4*, 179–185. [CrossRef]
71. Kim, J.Y.; Hwang, Y.M.; Sun, Y.G.; Sim, I.; Kim, D.I.; Wang, X. Detection for non-technical loss by smart energy theft with intermediate monitor meter in smart grid. *IEEE Access* **2019**, *7*, 129043–129053. [CrossRef]
72. Punmiya, R.; Choe, S. Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing. *IEEE Trans. Smart Grid* **2019**, *10*, 2326–2329. [CrossRef]
73. Gao, H.-X.; Kuenzel, S.; Zhang, X.-Y. A hybrid ConvLSTM-based anomaly detection approach for combating energy theft. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–10. [CrossRef]
74. Nabil, M.; Ismail, M.; Mahmoud, M.; Shahin, M.; Qaraqe, K.; Serpedin, E. Deep learning-based detection of electricity theft cyber-attacks in smart grid ami networks. In *Deep Learning Applications for Cyber Security*; Alazab, M., Tang, M., Eds.; Advanced Sciences and Technologies for Security Applications; Springer International Publishing: Cham, Switzerland, 2019; pp. 73–102, ISBN 978-3-030-13056-5.
75. Nabil, M.; Mahmoud, M.; Ismail, M.; Serpedin, E. Deep recurrent electricity theft detection in AMI networks with evolutionary hyper-parameter tuning. In Proceedings of the 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Atlanta, GA, USA, 14–17 July 2019; pp. 1002–1008.
76. Jindal, A.; Dua, A.; Kaur, K.; Singh, M.; Kumar, N.; Mishra, S. Decision tree and SVM-based data analytics for theft detection in smart grid. *IEEE Trans. Ind. Inf.* **2016**, *12*, 1005–1016. [CrossRef]
77. Zheng, K.; Chen, Q.; Wang, Y.; Kang, C.; Xia, Q. A novel combined data-driven approach for electricity theft detection. *IEEE Trans. Ind. Inf.* **2019**, *15*, 1809–1819. [CrossRef]
78. Zhang, W.; Dong, X.; Li, H.; Xu, J.; Wang, D. Unsupervised detection of abnormal electricity consumption behavior based on feature engineering. *IEEE Access* **2020**, *8*, 55483–55500. [CrossRef]
79. Tacón, J.; Melgarejo, D.; Rodríguez, F.; Lecumberry, F.; Fernández, A. Semisupervised approach to non technical losses detection. In Proceedings of the Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Puerto Vallarta, Mexico, 2–5 November 2014; pp. 698–705.

80. Aslam, Z.; Ahmed, F.; Almogren, A.; Shafiq, M.; Zuair, M.; Javaid, N. An attention guided semi-supervised learning mechanism to detect electricity frauds in the distribution systems. *IEEE Access* **2020**, *8*, 221767–221782. [CrossRef]
81. Li, J.; Wang, F. Non-technical loss detection in power grids with statistical profile images based on semi-supervised learning. *Sensors* **2020**, *20*, 236. [CrossRef]
82. Hu, T.; Guo, Q.; Shen, X.; Sun, H.; Wu, R.; Xi, H. Utilizing unlabeled data to detect electricity fraud in AMI: A semisupervised deep learning approach. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3287–3299. [CrossRef]
83. Lu, X.; Zhou, Y.; Wang, Z.; Yi, Y.; Feng, L.; Wang, F. Knowledge embedded semi-supervised deep learning for detecting non-technical losses in the smart grid. *Energies* **2019**, *12*, 3452. [CrossRef]
84. Qi, R.; Li, Q.; Luo, Z.; Zheng, J.; Shao, S. Deep semi-supervised electricity theft detection in AMI for sustainable and secure smart grids. *Sustain. Energy Grids Netw.* **2023**, *36*, 101219. [CrossRef]
85. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33, pp. 1877–1901.
86. Ashraf, M.; Anowar, F.; Setu, J.H.; Chowdhury, A.I.; Ahmed, E.; Islam, A.; Al-Mamun, A. A survey on dimensionality reduction techniques for time-series data. *IEEE Access* **2023**, *11*, 42909–42923. [CrossRef]
87. Patil, R.; Gudivada, V. A review of current trends, techniques, and challenges in large language models (LLMs). *Appl. Sci.* **2024**, *14*, 2074. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.