energies

MDPI

Article

Towards Big Data Electricity Theft Detection Based on Improved RUSBoost Classifiers in Smart Grid

Rehan Akram [1,†], Nasir Ayub [1,2,†], Imran Khan [3,†], Fahad R. Albogamy [4,†], Gul Rukh [3,*,†], Sheraz Khan [3,†], Muhammad Shiraz [1,5] and Kashif Rizwan [1]

1 Department of Computer Science, Federal Urdu University of Arts, Science and Technology Islamabad, Islamabad 44000, Pakistan; rehanakram.ms@gmail.com (R.A.); nasirayub@fuuast.edu.pk (N.A.); muhammad.shiraz@fuuast.edu.pk (M.S.); kashifrizwan@fuuast.edu.pk (K.R.)
2 School of Electrical Engineering and Computer Science, National University of Science and Technology, Islamabad 44000, Pakistan
3 Department of Electrical Engineering, University of Engineering and Technology, Mardan 23200, Pakistan; imran@uetmardan.edu.pk (I.K.); sheraz@uetmardan.edu.pk (S.K.)
4 Computer Sciences Program, Turabah University College, Taif University, P.O. Box 11099, Taif 26571, Saudi Arabia; f.alhammdani@tu.edu.sa
5 Department of Computer Science, Allama Iqbal Open University, Islamabad 44000, Pakistan
* Correspondence: gr@uetmardan.edu.pk
† These authors contributed equally to this work.

Abstract: The advent of the new millennium, with the promises of the digital age and space technology, favors humankind in every perspective. The technology provides us with electric power and has infinite use in multiple electronic accessories. The electric power produced by different sources is distributed to consumers by the transmission line and grid stations. During the electric transmission from primary sources, there are various methods by which to commit energy theft. Energy theft is a universal electric problem in many countries, with a possible loss of billions of dollars for electric companies. This energy contention is deep rooted, having so many root causes and rugged solutions of a technical nature. Advanced Metering Infrastructure (AMI) is introduced with no adequate results to control and minimize electric theft. Until now, so many techniques have been applied to overcome this grave problem of electric power theft. Many researchers nowadays use machine learning algorithms, trying to combat this problem, giving better results than previous approaches. Random Forest (RF) classifier gave overwhelmingly good results with high accuracy. In our proposed solution, we use a novel Convolution Neural Network (CNN) with RUSBoost Manta Ray Foraging Optimization (rus-MRFO) and RUSBoost Bird Swarm Algorithm (rus-BSA) models, which proves to be very innovative. The accuracy of our proposed approaches, rus-MRFO and rus-BSA, are 91.5% and a 93.5%, respectively. The proposed techniques have shown promising results and have strong potential to be applied in future.

Keywords: smart grid; electricity theft; advanced metering infrastructure; RUSBoost; manta ray foraging optimization; bird swarm algorithm

Citation: Akram, R.; Ayub, N.; Khan, I.; Albogamy, F.R.; Rukh, G.; Khan, S.; Shiraz, M.; Rizwan, K. Towards Big Data Electricity Theft Detection Based on Improved RUSBoost Classifiers in Smart Grid. *Energies* **2021**, *14*, 8029. https://doi.org/10.3390/en14238029

Academic Editor: Hugo Morais

Received: 6 October 2021
Accepted: 16 November 2021
Published: 1 December 2021

1. Introduction

The proficient utilization of electric power vitality forces the requirement of the effective systems for the ideal utilization of accessible electric power. The electric power distribution framework incorporates a perplexing power network as an electricity grid. These electricity grids comprise countless electricity lines and devices. Observation and execution of such a complex system represents a significant day by day challenge for electrical companies in providing electricity supply and power distribution. These electrical companies have understood that data accumulated from sensors and specific estimating gadgets are increasingly more significant in making effective mark able strategies for business plans. The productive utilization of information collected from sensors and estimating

gadgets can prompt decreasing expenses and conveying better customer services. On the off chance that electricity conveyance companies guarantee that the information obtained from the sensors is abused along with the information from other data frameworks, they will improve the nature of information and subsequently the nature of choices they make. The electric power produced by electrical companies is further divided into two basic categories. One is called a traditional grid as shown in Figure 1 and the other one is a Smart Grid (SG), shown in Figure 2. The traditional grids are used in past eras. In some under developing countries still, the traditional grid is used. The traditional grid is fundamentally the interconnection of different force framework components, for example, simultaneous machines, transmission lines, power transformers, transmission substations, dissemination lines, dispersion substations, and various sorts of burdens [1]. They are situated a long way from the force utilization territory and electric force is transmitted through long transmission lines.

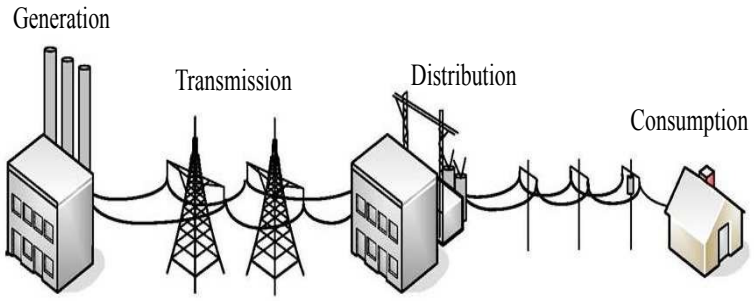

Figure 1. Traditional grid.

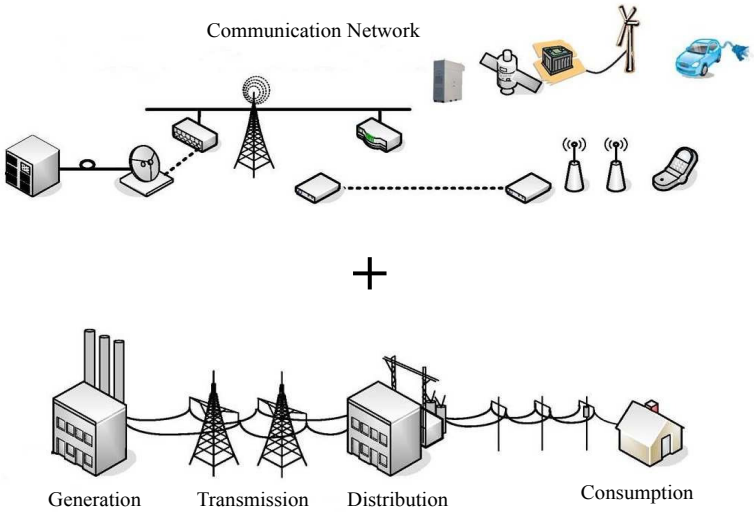

Figure 2. Smart grid.

The U.S. Department of Energy (DOE) characterizes the brilliant lattice as [2]: "A mechanized, generally appropriated vitality conveyance arrange the Smart Grid, on the other hand, will be portrayed by the two-path stream of electricity, and the data will be suitable for monitoring everything from power plants to the environment of client inclinations to singular machines. It consolidates into the network the advantages of appropriated figuring and interchanges to convey constant data and empower the close quick parity of flexibly and request at the gadget level".

The Smart Grid is, in general, a Transmission and Distribution (T & D) framework that coordinates detecting, checking, and correspondence innovation in order to keep T & D under controlled frameworks [3] to refine misfortunes and unwavering quality. Smart grids ensure that electricity is always available and forestall power outages by methods for computer innovation. This mode of keen frameworks provides a change in

perspective in electricity T & D frameworks to take into account unavoidable observing and control to promote efficiency, safety, security, and unwavering quality. The Smart Grid disperses slighter electricity since it can turn down concerns, such as obstruction and other unpleasant influences, preventing electricity outages. The Smart Grid values its ability to add new generators to the transmission network, allowing for more significant consolidation of sources of renewable energy.

It is increasingly important to demonstrate that a smart grid is able to connect not just traditional generators, gas turbines, and petroleum product generators, among instances, but also renewable energy generators including solar panels and wind turbines. It allows communication in both directions among generators, customers, and grid controllers in order to address the problem of unit responsibility, watch the electricity stream, screen data, and demonstrate any break progressively. The primary objective of the smart grid is to limit electrical losses and extend creation. Additionally, the electrical losses are not kidding sort of issue of various electric organizations.

There are two classifications of such issues, one is Technical Losses (TLs) and the other is Non-Technical Losses (NTLs). The TLs can be estimated with a low blunder rate and may diminish utilizing better facilities [4]. Notwithstanding, the decrease in NTLs is a troublesome issue and it is important to recognize the electrical losses esteeming a high proportion. The NTLs are brought about by controls in purchaser facilities. TLs are brought about by physical impacts (for example transformer issues), which have power dissemination imperatives [5].

A definitive victim due to NTL is the user as these losses are changed over into financial losses by DISCOM, which further brought about the burden of additional tariffs. In this way; electric utility grids ought to be urged for creating projects to decrease NTL. The general method of identifying extortion is to complete nearby reviews however the expense of on location assessment for various users may not be repaid with the estimation of the vitality recouped. Another methodology is the perception of the 'load profile' of users. By user's profile information, it is conceivable to recognize critical deviations in the conduct that can be related to NTL.

One of the significant difficulties confronting electric power suppliers overall is electricity theft that is the only act of utilizing electricity from the service company without the company's approval or assent. Electricity theft, which could occur as a result of billing discrepancies, meter altering, and unlawful association, and unpaid bills are generally done at the user end [6].

There are three sorts of electricity theft:

A. Outright theft, which is consummated by:
   - Tapping an overhead line to make a new, unauthorized interconnection;
   - Induction is a term that refers to the electromagnetic induction, which is used to collect energy from a power line without establishing physical contact to the line.

B. Fraud, which is consummated by:
   - By passing a meter, you can block it from calculating the amount of energy used;
   - Changing the settings on a meter to give the consumer a more favorable reading. Mechanical and digital/smart metering methods are separated into this category.

C. Billing Issues of Systematic non-payment of bills:
   - Intentional and unintended billing irregularities (poor record keeping practices).

Despite the fact that only a single approach is openly referred to as theft, these all issues entail the use of unpaid electricity.

A. Outright Theft

When someone steals power when they are not already a recognized client of a utility, it is called outright theft. An electrical thief attempts to establish a new, unauthorized T & D system connectivity without the permission of the T & D system's owner.

(1) Tapping

Tapping, which is making an unlawful association to lines buried or above ground on the distribution transformer's line [7]. In current working, tapping can be used to connect a building or piece of instrument to the electric grid where there was previously no connection. Taping has been used in the United States as an E-theft strategy since at least the 1890s [8]. Tapping exposes both the culprit and innocent bystanders to the risk of electrocution and death as a result of this type of power theft [9–11].

(2) Induction Coupling

To steal the electricity by placing a big coil under a high-voltage power line is one way that has gained media attention [12]. Inherently, such a technique is a sort of tapping via induction coupling. However, there are concerns with such schemes, aside from a few anecdotal accounts of electricity theft by induction coupling, e.g., [13–15]. Primarily, due to the massive amount of copper necessary to build a suitably big coil, a return on investment is often improbable [16]. Although induction coupling electricity theft is rare, induction power light bulbs have been found in creative installations where fluorescent tube lights are powered using induction [17].

B. Fraud

Electricity theft is accomplished by taking measures to keep record of less consumption by electricity bill or tampering with the power utility's metering equipment to make it record less (or none) utilization than was used.

(1) Bypassing Existing Meters

Bypassing a meter is similar to tapping, however, it is performed by linking the house wiring directly to the wires entering around the meter into the meter wiring [8,18–21]. This type of the electricity theft can either totally disengage the meter from the system or pull off the meter linked to the system additionally to the bypass, allowing the meter to record some consumption, though less than the earlier. Another method of bypassing is to use another (spare) meter for part of the billing time duration to prevent recording complete use. It would be unproductive for a thief to tell a utility to turn off electricity to a premises during whatever technique he or she used to bypass a meter, as a result, all connections should be moulded to live wires. As a result, this kind of theft carries a high level of personal risk.

(2) Meter Tampering

Electric meter tampering has been a problem for almost a century, and it is known that by the late 1890s, designing meters to avoid tampering was a top priority, and it is still a problem today. Mechanical meters and digital/smart meters are the two types of meter tampering having great interest (which are not immune to tampering).

To enhance the flawless distribution of energy sources toward the community, more extensive utilization of smart meters has expanded prospects of recognizing power theft by utilizing power utilization information imparted by smart meters to the utility server inside a consistent time period. These meters might be exposed to abnormalities identified with meter protections, estimation interferences, and system interruptions. Progressed Metering Infrastructure (AMI) has seemed well and good and store a lot of information conveyed by brilliant meters. By applying reasonable AI strategies on such a lot of keen meter information it has obtained conceivable to distinguish inconsistencies identified with NTL.

Since the development of calculations of AI, AMI has been broadly applied for the evaluation of the security of the force framework to defeat the inconveniences of regular strategies [21]. This area gives an outline of the detailed work for NTL recognition by utilizing both ordinary and current information examination strategies. There has been an expanding enthusiasm for the advancement of procedures based on the extraction of examples of shopper utilization conduct from recorded information. These strategies can

be managed, unaided, or semi-regulated [6]. Unaided techniques decide oddities without earlier information about clients' conduct, and administered strategies decide both typical just as anomalous conduct utilizing a regulated characterization that requires pre-grouped information [22]. This research proposes a deep learning-based algorithm for detecting theft in SG. The following are the primary contributions of this paper.

- A detailed review of different methods is provided in related work section;
- The solution of most relevant challenges are described in detail in proposed methodology;
- The proposed algorithm is validated by comparing with other models;
- The proposed model outperforms by achieving highest accuracy and minimizing electricity theft.

2. Related Work

The following is a work of several researchers' on theft detection models: In [23] study employs a method based on the power line communication concept, which is used to identify electricity theft. A high-frequency signal is introduced to the distribution network that fluctuates in amplitude and frequency as the load on the lines rises or falls. If any unlawful connections are formed between the poles, the gain values will be changed, the illegal connection in the electricity will be detected, and the authorities will take appropriate measures to neutralize it. However, this method has not been tested for detecting theft in consumers with unreadable connections [24]. It creates user load profile information based on a customer's historical power use pattern, which is used to detect atypical electricity flows and so identifies a class of users that might be further synthesized to detect suspected fraud consumers. Many ideas are used in the article, including Extreme Learning Machine (ELM) and Support Vector Machine (SVM). In these detecting processes, a variety of processes are used. To begin, consumer use data are pre-processed. Data selection, data separation, and data normalization are the three processes in the procedure. Then there is the feature selection process, which takes the most significant aspects of the data and extracts them automatically. The data are then classified using ELM based on anomalous usage patterns. The data are then further categorized using SVM in order to detect suspected electricity fraud. However, because we are using SVM, this is not a problem. Because SVM is not accurate in categorizing data to the degree that it is, there is a risk of fraud detection failure.

On the basis of time and speed, a comparison of K-Means and N-K Means clustering [24] was performed in this work. K-means based on cluster centroids or means, this is a very successful technique for grouping uniform and non-uniform data. Based on normalization, the (N-K) method is proposed. This method employs normalization, which is useful for clustering and establishing initial centroids based on available data and weight. After the database updates are performed, K-means produces efficient results. We use a converted approach to calculate initial centroids based on the weighted average core of the dataset. We standardize and pre-process the dataset before performing the N-K means technique. It is mostly dependent on the suggested approach in three steps. Data pre-processing techniques are used to turn raw data into a comprehensible format at the initial stage. The second stage involves applying normalization to a defined range of data items to return them to their original state. To get the clusters in the third step, we use the N-K means technique. The paper describes an efficient methodology in which we pre-processed our dataset using a normalization approach before generating effective clusters. To find the standardization, weights are assigned to each attribute value. Based on speed and time of execution, this method has proven to be superior to the classic K-means method. In comparison to K means clustering, experimental findings show that the suggested N-K means method has improved overall performance and time complexity.

In [25], the concept of using computational tools to categorize consumers' power use characteristics is presented. To obtain the results, the study employs a two-step procedure. C means based on fuzzy clustering is used to find customers with similar usage profiles using a distance-based method, and then fuzzy classification is performed to the fuzzy

cluster values and fraud matrix values using a distance based technique. The deflection is then used to grade the material. The more valuable the grade, the more probable it is to be cheated. The fuzzy C means clustering approach used for clustering improves the chances of finding similarities between normal and deviant customer usage patterns. Five attributes are used to create user profiles: average consumption of a specific client over the previous six months, maximum consumption over the previous six months, using these three parameters. Following the data that were used for classification, the categorization is completed using data from the preceding 12 months. As a consequence, we can evaluate the degree of irregularity in consumption and identify problematic customers using the appropriate threshold. The drawback of this method is related to accuracy issues. Although fuzzy clustering and classification produce adequate accuracy, due to the use of just a 6-month data cycle, there is still a chance that the fuzzy clusters training may not create a good load profile [26]. The Atkinson index is used to determine how serious the disease is because the Atkinson index is primarily concerned with quantity distribution over a range of incomes. To successfully use the Atkinson index to quantify aspects, such as pollution, this technique leverages principles, such as the relative Lorenz curve. However, as a result of this modification, the Atkinson index will only be used to compute negative outcomes; no comparisons between good and bad results will be possible using this technique. We would be able to detect power imbalances more effectively if we used this technique [27]. K means and hierarchical clustering were used to compare invasion datasets.

Clustering in a hierarchical structure: When an event happens, the election method chooses a Cluster Head (CH) for each cluster, who then sends out Cluster Configuration Messages (CCMs), which are identified as, where ID is the identification and W is the node of the energy factor. Among all the nodes in the cluster, the CH has the most energy. K-means: To discover the cluster, use a distance measure from a group to compare the consumed data to that of a cluster. Consumer data from one comparable cluster and consumer data from another distinct cluster. The simplest unsupervised clustering approach is K-means clustering. This technique separates the parameter k into n datasets into k clusters, resulting in both strong intra- and inter-cluster equality. K is a positive integer number that has been established. It takes the least amount of time and delivers significantly greater results when compared to hierarchical clustering. The Atkinson index is still the most effective approach for identifying disparities in the distribution of specific values, according to the research. The Gini index [28] is a useful tool, however, it is not without flaws. When understanding inequality, for example, the Gini index computation is cumbersome and causes calculating difficulties for specific values. The Atkinson index, on the other hand, is highly easy to decompose to accommodate a variety of changes, but the Gini index's decomposability is limited. As a consequence, it can be stated that when it comes to uneven distribution techniques, the Atkinson index outperforms the Gini index [29]. It necessitates sophisticated power meters that can wirelessly send customer consumption data to power authorities to monitor power use every half-hour. Customers' load profiles would be generated based on the data gathered thus far. This research employs intelligent systems and fuzzy logic ideas. The data used to compile the findings are for the period of one month. The information was first obtained from the smart meters. After that, the data were pre-processed to prepare for load profiling. Following that, an anomaly in consumption was detected using load profiling, and the client was categorized into five kinds using fuzzy logic, resulting in the identification of fraud. Such methods, however, need a large amount of infrastructure, and the findings may not be true, as the reference month chosen for testing may be a vacation month with low usage [30]. Poor parzen window estimates plague segmentation and clustering algorithms that employ the Gaussian kernel function to construct affinity matrices, such as spectral clustering approaches. The final results are based on this parameter, and they alter when its value fluctuates. We create a vector that corresponds to each row in a dissimilarity matrix, which is then used to generate an affinity matrix with the help of a Gaussian distribution function in this study, which employs optimization techniques in a novel estimation methodology. The affinity matrix generated

by the suggested approach is quite valuable and includes extra information, such as the number of clusters; nevertheless, comprehensive clustering without the use of other techniques is not achievable. This study offers a portfolio optimization system that manages stock portfolios using a Neuro-Fuzzy methodology in sequence. In order to maximize profits from a stock portfolio, the proposed portfolio reduction methodology Neuro-Fuzzy system reasoning is used. When compared to existing envelope models, the Neuro Fuzzy model now delivers substantially better confidence. The strategy presented in this paper was developed by the BSE Sensex stock index. The results of the experiments reveal that the success of models may be measured by their return on assets and risks. Thus, when compared to previous portfolio models, the results generated utilizing the suggested methodology in performance opinion tests utilizing live stock exchange data yielded a considerably greater return on investment values [30]. The notion of genetic algorithm is combined with the Support Vector Machine in this study (SVM). The billing data collected from the authorities were first filtered using criteria such as ignore clients who had not used the service in the previous 25 months, and so on. After that, the load profiling is finished. After that, the feature extraction and data normalization were finished. SVM was used to classify the data, which was then divided into four groups. Consumers in class 1 are the most likely to perpetrate fraud, while those in class 4 are the least likely. The hyper parameters of SVM are then reduced to a only chromosome using Genetic Algorithm (GA) optimization, and the fraud is identified with minimum effort. However, this technique has a problem with accuracy because the SVM is a poor classification mechanism, and while the GA decreases the effort, the precision remains deprived [31]. The rectified Gaussian distribution is a straightforward but effective variant of the standard Gaussian distribution. The variables of the rectified Gaussian must be positive in order to practice concave energy functions. The cooperative and competitive distributions are two multi-dimensional instances of the rectified Gaussian's power. The cooperative distribution may be used to express pattern translations, demonstrating the rectified Gaussian's promise for modeling pattern manifolds. Finding tractable learning methods is crucial for making the corrected Gaussian usable in real applications. It is unclear if the learning for the corrected Gaussian will be more manageable than it was for the Boltzmann machine. Perhaps the rectified Gaussian's real value variables are easier to deal with than the Boltzmann machine's binary variables [32]. On invasion-based datasets, a comparison is made between K means and C means clustering. The dataset includes all K means and C means clustering inequality measurements. The confusion matrix is used to analyze the results of these clustering techniques. This methodology is based on three intrusion datasets: KDDCup99, NSLKDD, and GureKDD, and it employs several pre-processing approaches. These data are preprocessed and standardized before being used as input for models. Based on their clustering accuracy and computing time, it can verify. Clustering's fundamental purpose is to identify things that are similar and different. The similarity between the items in the cluster is used to evaluate the algorithm's performance. We will choose the non-similar measurements that produce superior outcomes when comparing K means and C means. In the KDD corrected dataset, Euclidean distance has higher accuracy than other measurements. Similarly, adopting K means as the second choice yields the best results. In contrast to C means, the results suggest that K means delivers greater clustering accuracy. K means is a fantastic alternative for designing intelligent invasion detecting software [33]. This document explains estimation software. The greatest challenge for a software developer is accuracy. It may look at project decisions such as source allocation and direction, which could be used to schedule and lengthen the project.

We provide a unique system that uses fuzzy logic to estimate important elements of software test evaluation, such as cost and time, as well as a neural network system to perform desire estimations for program development, in this paper. In this paper, we show how, as compared to previous fuzzy models with enrolment objectives, the Bayesian regularization technique provides minimal condition numbers, which represent abandon at certain places and the status number grows at some locations [34]. This research proposes

a fuzzy logic-based technique for detecting power theft and improving power quality. The fault signal is used to demonstrate power theft by comparing the overall load delivered by the administration power plant to the total load utilized by the user. The Fuzzy logic controller is given a load (energy) and a voltage as inputs, and the controller provides the relevant fluctuations in output voltage to increase the power nature and reduce power theft. The replications are carried out in MATLAB, and the simulation results are supplied. It enables us to predict how effectively intelligent control in electrical models will perform. As a result, incorporating creative authority into the electrical model has the potential to significantly improve power efficiency while also securing a large number of non-legal activities. Surveillance of power theft will be a barometer of a country's economy's ability to meet its power generation needs. As a result, the development of automated intelligence in power systems will be a huge step forward [35]. This study presents the idea of utilizing temperature-based energy meters to predict electricity theft using smart meters. It is a model based on the evolution of technical loss estimation based on continuous resistance. They used the model and made improvements in predicting resistance by taking temperature into account. The theft can be predicted by computing the variables and estimating the circuit since the variables are dependent on the material used to create transmission lines. If the theft rate is more than 4%, the model is useful for forecasting. The model employs user power use profiling, with data from the meter being gathered every 30 minutes, and a temperature profile is provided as input. This system works well for identifying illicit grid connections, but the complexity and infrastructure requirements are substantially higher. Because circuit calculations and approximations are required, they must be completed.

From customers to the utility billing system, a variety of strategies for stealing energy have been found and implemented. Support Vector Machine (SVM) [36,37] is a popular method for analyzing energy usage profiles. This method entails training of an SVM on a historical dataset and then evaluating it on a new dataset in order to detect abnormalities or deviations in consumer energy consumption profiles. Smart meters become the backbone of smart grids by using AMI as upgraded digitize technology as shown in Figure 3. Smart meters play a key role among various sensors and information sources that can show vitality theft, practically speaking, and the individual strategies. These individual strategies show numerous faults. AMI interruption recognition framework that utilizes data combination to consolidate the sensors and utilization information from a smart meter to all the more precisely distinguish faults. There are three key parameters involved in this technology, such as measurement, consumption, and distribution. These monitoring capabilities, coupled with huge scope AMI information accumulation guarantee to fundamentally moderate the issue of vitality robbery, a particularly inescapable issue in creating nations. Among many proposed methods, one technique refers to Advance Metering Infrastructure Intrusion Detection System (AMIDS) [38]. It is an attack graph-based information fusion method for conceptually combining the data gathered.

These are three types of AMI specific information sources [39].

Cyber-Side Networks: Data combination arrangement which makes utilization of an AMI explicit assault diagram to recognize vitality theft endeavors with least number of faults positives.

Anti-Tampering Sensors: Information mining systems to recognize vitality theft through nonintrusive burden checking.

Nonintrusive Load Monitoring: A managed approach that can distinguish singular apparatus utilization and a solo approach that learns by bunching load occasions.

The energy theft is a major issue for utilities since the start of the electricity billing system. In reality, smart meters have been intended to recognize and report altering endeavors. They use some assault methods, which are well known strong state kills with simple conventional meters. Smart meter alerts have the ability to identify meters being tilted, separated, turned around, or even hacked. The electric theft problem stays, although utilities are presently confronting new significant difficulties. Using AMI non

tamper proof meters, they present a noteworthy arrangement of new assault strategies to achieve energy theft. These well-known used strategies are incorporate intruding on estimations, increasing special effusion to the meter firmware, messing with the availability of meter storage. They are also beneficial in capturing the meter communication to a block or modifying the utilization value being reported [40].

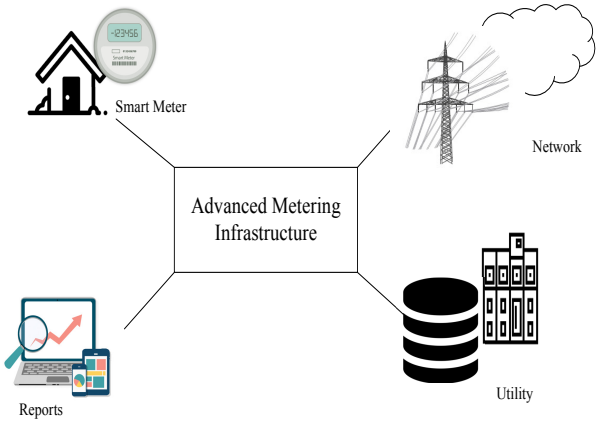

Figure 3. Advanced metering infrastructure.

Different detecting techniques have threat models, which record energy threats according to information sources as shown in Table 1.

Table 1. Threat classified by different detection techniques.

Type of Attacks	Description
Cyber Attacks	Compare meters through remote network exploit
	Modify the firmware/ Storage on meters
	Steal credentials to login to meters
	Intercept/alter communications
	Flood the NAN bandwidth
Physical Attacks	Break into meter
	Reverse/disconnect the meter
	Physically extract the password
	Abuse optical port to gain access to meters
Data Attacks	Bypass meters to remove loads from measurements
	Stop reporting entire consumption
	Remove large appliances from measurements
	Report zero consumption
	Alter load profiles to hide large loads
	Report negative consumption (acts as a generator)

The proposed data detection techniques utilize diverse data sources to assemble adequate measures of proof about an ongoing attack before denoting an action as a malicious energy theft [25]. Various test datasets show that through a viable data combination and utilizing the relationship among the activated alarms, AMIDS can distinguish different sorts of energy theft endeavors precisely by utilizing exclusive sensors. The difficulty of discovering forms in data that do not imitate to estimated behavior is known as anomaly detection. In various application fields, these non-imitating designs are stated to as anomalies, outliers, dissonant observations, exemptions, aberrations, wonders, oddities, or impurities. To keep a consumption profile for each client and observe the profiles for any abnormalities, anomaly detection techniques are preferred. Different techniques and algorithms are applied to different datasets to see the anomaly or issue in load records. Many approaches are used to avoid cyber-attacks, which are further categorized into the game based, state based and artificial intelligence based models. State based detection use many sensor devices to detect electricity theft [41]. The problem of power theft detection is represented

as a game between the electrical thief and the electrical utility in a game theory-based technique. These strategies could provide a low-cost, reasonable, but not ideal option for decreasing energy theft [42]. Artificial intelligence based models use different learning methods such as deep leaning and machine learning [25]. Machine learning includes clustering and classification while deep learning includes many neural network architectures. Currently, most used neural network architecture is convolutional neural network, as shown in Figure 4 [25]. As a classifier, different datasets are trained and tested using CNN. On the other side, when we talk about supervised machine learning algorithms, such as RF, then a decision tree is used which comprises the tree concept for prediction [43].

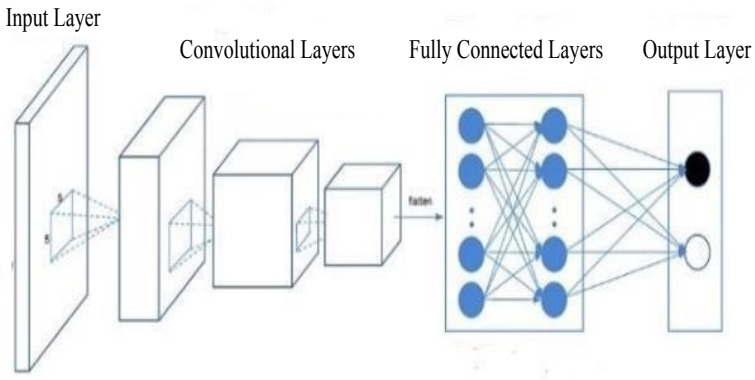

Figure 4. CNN Model Layers.

3. Design

3.1. Model Overview

The precision of estimation is a critical factor in detecting electricity theft. The proposed model compared with the traditional Support Vector Machine [24], Linear Regression (LR), and convolutional neural network [25]. SVM have higher complexity with high loss rate. LR is sensitive to outliers and keenly observe a relationship between the mean of the independent variable and the dependent variables. CNN have issues like overfitting, and class imbalance while training the model. The proposed smote is implemented for balancing data, as stated earlier. To improve the accuracy of electricity theft detection RUSBoost with manta ray foraging optimization and RUSBoost with bird swarm classifiers are used. Finally, classifiers are employed to forecast electricity theft using selected features and to give excellent performance. The schematic of the proposed model is illustrated in Figure 5.

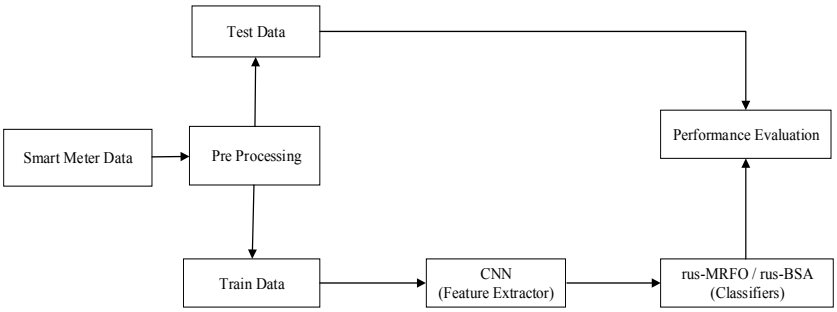

Figure 5. Proposed model.

3.2. Classifiers and Techniques

In our proposed model, CNN [25] is designed to learn the features between different hours of the day. These features are carried from massive varying smart meter data by the operations of convolution, downsampling, and RUSBoost manta ray foraging optimization or RUSBoost bird swarm algorithm. The problem of data sampling and boosting analyzed by several techniques [44]. RUSBoost [45], a novel hybrid data sampling/boosting tech-

nique meant to improve the performance of models, in this study. RUSBoost reduces the time it takes to build a model, yielding dramatically improved classification accuracy. The proposed algorithm is the metaheuristic Manta Ray Foraging Optimization (MRFO) [46] imitates the foraging activities of manta rays.

The Bird Swarm Algorithm (BSA) is proposed in [47], a new bio-inspired algorithm. In bird swarms, BSA is a simplification of social behaviors and interactions. To solve difficulties, BSA always follows a set of guidelines. It imitates the foraging, attentiveness, and flight behaviors of the bird. As a result, swarm intelligence can be retrieved easily from bird swarms.

Several strategies are used to investigate the problem of data sampling and boosting [44]. We used RUSBoost algorithm which reduces the time to build a model, with dramatically improved classification accuracy. The proposed metaheuristic algorithm Manta Ray Foraging Optimization (MRFO) [45] simulates the foraging activities of manta rays. Chain foraging, cyclone foraging, and somersault foraging are three foraging operators in this algorithm.

Chain foraging is the first foraging approach [46]. Manta rays begin foraging by establishing an organized line by line. Male manta rays are piggybacked onto females and swim on top of their backs to match the female's pectoral fins' beats. As a result, plankton overlooked by prior manta rays will be snatched up by those following them. By working together, they can get the most plankton into their gills and increase their food rewards. The cyclone foraging strategy [48] is the second foraging strategy. Hundreds of manta rays congregate when plankton concentrations are extremely high. Their tail ends spiral together with their heads to form a spiraling vertex in the cyclone's eye, and the filtered water rises to the surface. The plankton is drawn into their wide jaws as a result of this. Somersault foraging [49] is the final foraging approach. This is one of nature's most beautiful scenes. When manta rays discover a food supply, they do a sequence of backward somersaults, circling the plankton and attracting it to them. Somersault is a cyclical, random, frequent, and localized movement that helps manta rays improve their food intake. These foraging activities are uncommon, yet they are quite effective.

The Bird Swarm Algorithm (BSA) is a new bio-inspired algorithm. BSA is a simplification of the social behaviors and social interactions in bird swarms. BSA always follows some rules to optimize problems. It mimics the birds foraging behavior, vigilance behavior, and flight behavior. Thus, swarm intelligence can be efficiently extracted from bird swarms. Our proposed model, Smote, is designed to balance the imbalanced collected data. After these data samples are divided into two categories, Train data and Test data including features. These features are carried from massive varying smart meter data by the operations of convolution, downsampling, and RUSBoost manta ray foraging optimization or RUSBoost bird swarm algorithm. Moreover, the proposed classifiers are a combination of sampling, boosting, and optimization. These rus-MRFO and rus-BSA classifiers are trained based on the obtained features to detect whether the consumer thieve electricity or not.

Finally, the confusion matrix and Receiver Operating Characteristic (ROC) curves are used to evaluate the accuracy of the proposed models. Further, we compared our model with state-of-the-art models.

4. Data Gathering

The New York Independent System Operator (NYISO) is responsible for organization and stimulating speculation in New York's electric infrastructure. We do not create electricity or own transmission lines; instead, we collaborate with power manufacturers, utility companies, and other stakeholders to satisfy New Yorkers' electrical demands on a daily, hourly, and minute-by-minute basis. In reality, the NYISO's authoritative statistics and planning knowledge is relied on by government officials, Wall Street investors, and energy sector professionals all around the world.

Our dataset comprised of approximately 25,000 user's data and metadata, as shown in Table 2. Malicious samples, on the other hand, are impossible to get because energy theft may never or rarely occur for a given consumer. Because of the enormous number and variety of consumers, as well as the lengthy period of measurements, this dataset is a good source for smart meter data analysis study.

Table 2. Dataset Description.

Description	Value
Time duration of data	1 January 2014 to 31 September 2016
No. of total consumers	25,000
No. of normal users	22,532
No. of electricity thieves	2468

5. Results and Discussions

5.1. Simulation Environment

The proposed models are implemented using the Python packages Keras, NumPy, TensorFlow, Imblearn, Sklearn, and pandas for simulation purposes. Models are run on a system with an Intel Core i5, 6 GB RAM, and 1 TB storage. The dataset used to detect electricity theft was obtained from NYISO. NYISO contains real-time data of consumers. It features daily data on electricity usage and generation from 2014 to 2016.

5.2. Simulation Results

After running simulations on the NYISO dataset, the results are presented in the form of graphs. The comparison of different approaches for detecting electricity theft is shown in these graphs. Figure 6 depicts the performance of our suggested model using a ROC curve. It can be observed in ROC curve that our proposed techniques rus-BSA achieved 93%, and rus-MRFO attained 91.5% of Area Under the Curve (AUC), while the traditional Support Vector Machine, Linear Regression, and convolutional neural network achieved 71%, 63%, and 85.1%, respectively. It can be concluded that the proposed rus-BSA outperformed state-of-the-art techniques, and the results indicate that our improved model can better distinguish between normal and theft users.

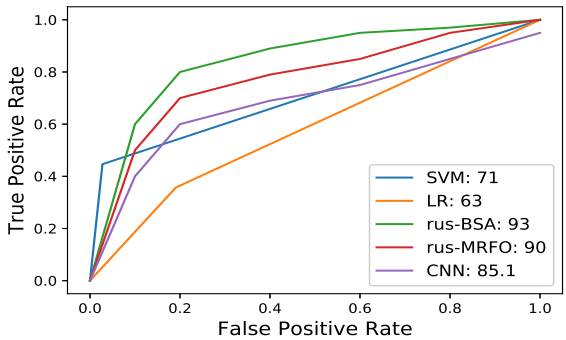

Figure 6. ROC Curve.

As seen in Figure 7, our improved methodology outperforms the competition. Our proposed approaches, rus-MRFO and rus-BSA, have a 91.5% and a 93.5% accuracy, respectively, which are better than the SVM, LR, and CNN with an average accuracy of 68%, 63%, and 85.1%, respectively. The proposed rus-BSA has around 25% higher accuracy than traditional linear regression technique, and 8% higher than commonly employed CNN technique.

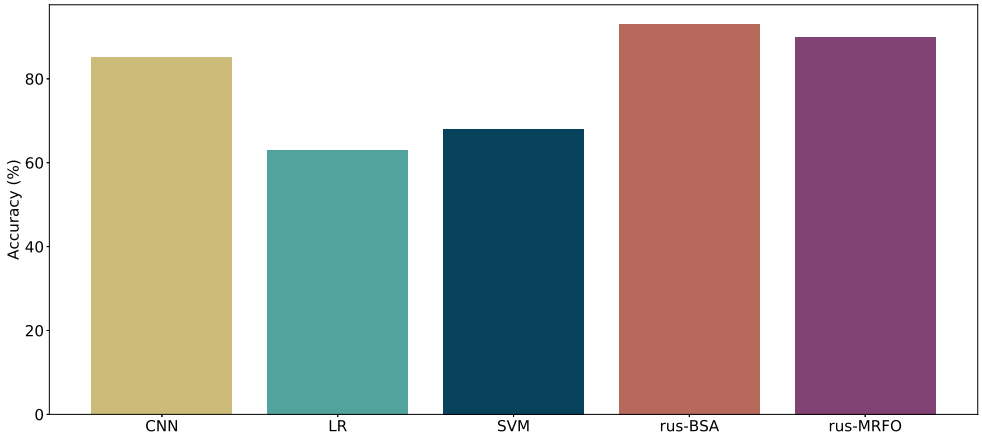

Figure 7. Accuracy Techniques.

As shown in Figure 8, the performance of proposed technique is evaluated using performance measures, such as the F1-score, accuracy, precision, and recall. The rus-BSA outperformed the baseline techniques with a significant margin and achieved an average values of 92%, 93%, 92.3%, and 94% of F1-score, accuracy, precision, and recall, respectively. The performance metrics values are also shown in Table 3.

Table 3. Performance metrics score.

Techniques	F1-Score	Accuracy	Precision	Recall	AUC
CNN	86.2%	85.1%	87.43%	88%	85.1%
LR	61%	63%	67%	65%	63%
SVM	71%	68%	65%	72%	71%
rus-BSA	92%	93.5%	92.32%	94.02%	93%
rus-MRFO	87%	91.5%	89%	92.87%	90%

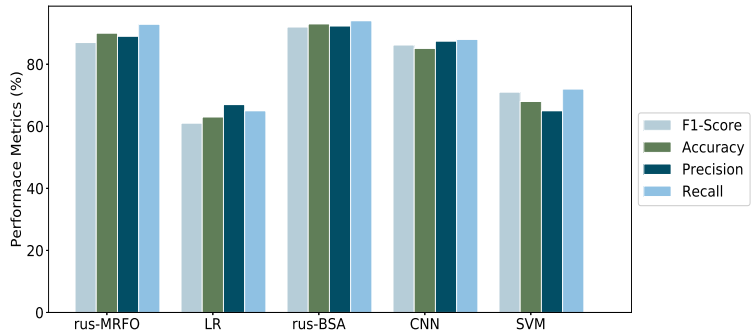

Figure 8. Performance Metrics.

The performance of the benchmark approaches is inferior to that of our suggested solution.

In addition, the accuracy and loss curve of rus-MRFO is given in Figures 9 and 10. The rus-MRFO achieved 91% and 90.6% of training and testing accuracy, while the model loss for training is 9% and 10.4% for testing. Similarly, the proposed techniques rus. BSA achieved 93% of training accuracy, and 91.3% of testing accuracy, and model loss is 7%, and 9.7%, for training and testing, respectively, as depicted in Figures 11 and 12.

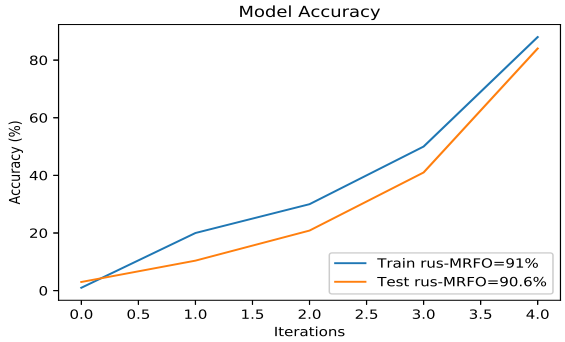

Figure 9. Accuracy rus-MRFO.

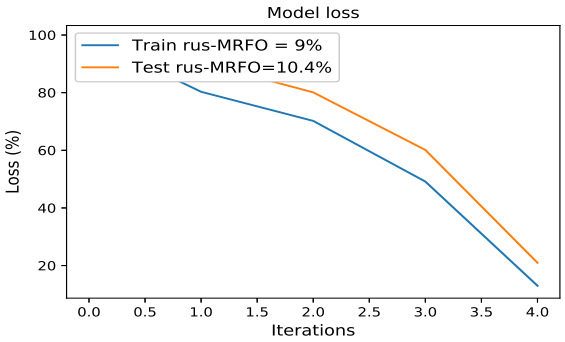

Figure 10. Loss rus-MRFO.

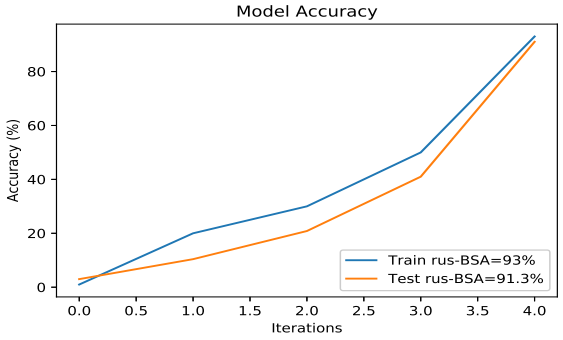

Figure 11. Accuracy rus-BSA.

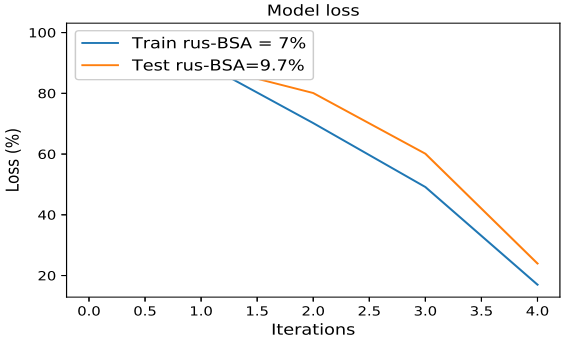

Figure 12. Loss rus-BSA.

Figure 13 depicted the overall performance error of different techniques and proposed model is calculated using different measures, such as Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Mean Squared Error(MSE). The performance error of our proposed models rus-MRFO and rus-BSA in terms of MAPE, RMSE, and MSE is 10%, 14.24%, 31.87% and 7%, 10%, 21.5%k, respectively.

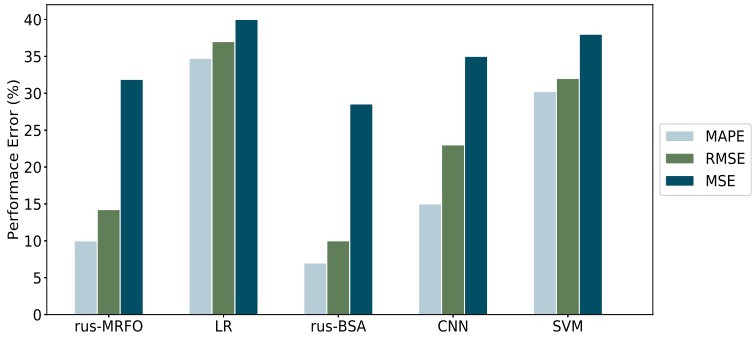

Figure 13. Overall performance error.

6. Conclusions

Despite being under-reported and publicized, electricity theft is a serious challenge for utilities. Recent advancements in modern energy systems have improved resistance to electricity theft while also introducing new vectors and tactics for power thieves. On the other side, the electronic nature of digital meters creates new vulnerabilities to power theft and meter manipulation, which must be addressed through cybersecurity measures. This study looks at the issues that go into detecting energy theft, the reasons that drive unlawful users, and the methods that have been used in the past to reduce theft. Based on their location, load capacity, and type, it determines the approximate energy consumption patterns of several customers. A dataset detailing the energy consumption trends of numerous clients is built using historical data. The SVM model is trained using the input training data, and various clients' power usage patterns are evaluated as needed. SVM achieves the accuracy rate of 71% which is less than the accuracy rate of CNN. The overall accuracy rate of CNN is 85.1% and LR is lowest with 63%. In proposed model, Smote algorithm is used for balancing data. Proposed classifiers with boosting technique outperform accuracy and performance. The overall accuracy rate of rus-MRFO is 90% and rus-BSA is 93%. Both classifiers give excellent performance in terms of precision, accuracy, recall, and f-measure. In future, to achieve better accuracy, we will use enhanced techniques and extend our methods to the real-time environment for electricity theft detection.

Author Contributions: Conceptualization, Data curation, Methodology, Resources, Validation, Writing—original draft, Writing—review & editing, R.A., N.A., I.K., M.S. and K.R.; Conceptualization, Data curation, Methodology, Resources, Validation, Writing—original draft, Writing—review & editing, Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, and Visualization, I.K., G.R., and S.K.; Writing—review & editing, Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, and Visualization, F.R.A. All authors have read and agreed to the published version of the manuscript.

Funding: The APC is funded by Taif University Researchers Supporting Project Number (TURSP-2020/331), Taif University, Taif, Saudi Arabia.

Acknowledgments: The authors would like to acknowledge the support from Taif University Researchers Supporting Project Number(TURSP-2020/331), Taif University, Taif, Saudi Arabia.

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

Abbreviations

The following abbreviations are used in this manuscript:

TLs	Technical Losses
NTLs	Non-Technical Losses
SM	Smart Grid
AMI	Advanced Metering Infrastructure
CNN	Convolutional Neural Network
MRFO	Manta Ray Foraging Optimization
BSA	Bird Swarm Algorithm

References

1. 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]
2. Ghulam, H.; Khan, I.; Jan, S.; Shah, I.A.; Khan, F.A.; Derhab, A. A novel hybrid load forecasting framework with intelligent feature engineering and optimization algorithm in smart grid. *Applied Energy* **2021**, *299*, 117178.
3. Hafeez, G. Electrical Energy Consumption Forecasting for Efficient Energy Management in Smart Grid. Ph.D. Thesis, COMSATS University of Islamabad, Islamabad, Pakistan, 2021.
4. Batra, S.G. A Comparative Study of Various Combinatorial Approaches for Minimization of Transmission and Distribution Losses. Ph.D. Thesis, Thapar University, Punjab, India, 2015.
5. Guerrero, J.I.; León, C.; Monedero, I.; Biscarri, F.; Biscarri, J. Improving knowledge-based systems with statistical techniques, text mining, and neural networks for non-technical loss detection. *Knowl.-Based Syst.* **2014**, *71*, 376–388. [CrossRef]
6. Smith, T.B. Electricity theft: A comparative analysis. *Energy Policy* **2004**, *32*, 2067–2076. [CrossRef]
7. Bihl, T.J.; Hajjar, S. Electricity theft concerns within advanced energy technologies. In Proceedings of the 2017 IEEE National Aerospace and Electronics Conference (NAECON), Dayton, OH, USA, 27–30 June 2017; pp. 271–278.
8. Hallberg, J.H. Theft of current: How to detect, prosecute and prevent II. *Electr. World Eng.* **1905**, *45*, 884–886.
9. Suriyamongkol, D. Non-Technical Losses in Electrical Power Systems. Doctoral Dissertation, Ohio University, Athens, OH, USA, 2002.
10. Taylor, A.J.; McGwin, G., Jr.; Brissie, R.M.; Rue, L.W., III; Davis, G.G. Death during theft from electric utilities. *Am. J. Forensic Med. Pathol.* **2003**, *24*, 173–176. [CrossRef]
11. Kim, V. *Father and Daughter Burned in Alleged Electrical Theft*; Los Angeles Times: Los Angeles, CA, USA, 2011.
12. Lopez, C.; Sargolzaei, A.; Santana, H.; Huerta, C. Smart Grid cyber security: An overview of threats and countermeasures. *J. Energy Power Eng.* **2015**, *9*, 632–647.
13. Chrasekhar, A.; Vivekananthan, V.; Khelwal, G.; Kim, W.J.; Kim, S.J. Green energy from working surfaces: A contact electrification–enabled data theft protection and monitoring smart table. *Mater. Today Energy* **2020**, *18*, 100544. [CrossRef]
14. Kumar, M.E.; Reddy, G.T.; Sudheer, K.; Reddy, M.P.K.; Kaluri, R.; Rajput, D.S.; Lakshmanna, K. Vehicle theft identification and intimation using gsm & iot. *IOP Conf. Ser. Mater. Sci. Eng.* **2017**, *263*, 042062.
15. Bihl, T.J.; Hajjar, S. *Electromagnetic Harvesters: Free Lunch or Theft*; Industry Tap: Sheffield, UK, 2013; pp. 271–278.
16. Deardorff, D.L. A Solution to the RWP for Exam 1-Stealing Power. *Retrieved August* **2006**, *28*, 2015.
17. BBC. Floods in Drought-Hit Kenya Spark Aid. 2011. Available online: http://news.bbc.co.uk/2/hi/8441708.stm (accessed on 2 February 2021).
18. Weslowski, J. Utilities launch assault to halt theft of power. *Electr. Light Power* **1976**, *10*, 54.
19. Bihl, T.J.; Zobaa, A.F. Data-mining methods for electricity theft detection. In *Big Data Analytics in Future Power Systems*; CRC Press: Boca Raton, FL, USA, 2018; pp. 107–124.
20. Somefun, T.E.; Awosope, C.O.A.; Chiagoro, A. Smart prepaid energy metering system to detect energy theft with facility for real time monitoring. *Int. J. Electr. Comput. Eng.* **2019**, *9*, 4184.
21. Venkatesh, T.; Jain, T. Synchronized measurements-based wide-area static security assessment and classification of power systems using case based reasoning classifiers. *Comput. Electr. Eng.* **2018**, *68*, 513–525.
22. Hodge, V.; Austin, J. A survey of outlier detection methodologies. *Artif. Intell. Rev.* **2004**, *22*, 85–126. [CrossRef]
23. Christopher, A.V.; Swaminathan, G.; Subramanian, M.; Thangaraj, P. Distribution line monitoring system for the detection of power theft using power line communication. In Proceedings of the 2014 IEEE Conference on Energy Conversion (CENCON), Johor Bahru, Malaysia, 13–14 October 2014; pp. 55-60.
24. Dangar, D.; Joshi, S.K. Electricity Theft Detection Techniques for Distribution System in GUVNL. *Int. J. Res. Anal. Rev.* **2014**, *7*, 513–524.
25. Angelos, E.W.S.; Saavedra, O.R.; Cortés, O.A.C.; de Souza, A.N. Detection and identification of abnormalities in customer consumptions in power distribution systems. *IEEE Trans. Power Deliv.* **2011**, *26*, 2436–2442. [CrossRef]
26. Jumale, P.; Khaire, A.; Jadhawar, H.; Awathare, S.; Mali, M. Survey: Electricity Theft Detection Technique. *Int. J. Comput. Eng. Inf. Technol.* **2016**, *8*, 30.

27. Saxena, H.L.; Richariya, D.V. Intrusion Detection System using K-means, PSO with SVM Classifier: A Survey. *Int. J. Emerg. Technol. Adv. Eng.* **2014**, *4*, 653–657.
28. Creedy, J. Interpreting inequality measures and changes in inequality. *N. Z. Econ. Pap.* **2016**, *50*, 177–192. [CrossRef]
29. Nagi, J.; Yap, K.S.; Nagi, F.; Tiong, S.K.; Koh, S.P.; Ahmed, S.K. NTL detection of electricity theft and abnormalities for large power consumers in TNB Malaysia. In Proceedings of the 2010 IEEE Student Conference on Research and Development (SCOReD), Kuala Lumpur, Malaysia, 13–14 December 2010; pp. 202–206.
30. El-Bhissy, K.; El-Faleet, F.; Ashour, W.M. Clustering Using Optimized Gaussian Kernel Function. *Int. J. Artif. Intell. Appl. Smart Devices* **2014**, *2*. [CrossRef]
31. Socci, N.D.; Lee, D.D.; Seung, H.S. The Rectified Gaussian Distribution. In Proceedings of the 11th Annual Conference on Neural Information Processing Systems, Denver, CO, USA, 1–3 December 1998; pp. 350–356.
32. Sahu, S.K.; Jena, S.K. A study of K-Means and C-Means clustering algorithms for intrusion detection product development. *Int. J. Innov. Manag. Technol.* **2014**, *5*, 206–213. [CrossRef]
33. Singh, S.P.; Johri, P. A Review of Estimating Development Time and Efforts of Software Projects by Using Neural Network and Fuzzy. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2012**, *2*, 306–310.
34. Rengarajan, S.; Loganathan, S. Power theft prevention and power quality improvement using fuzzy logic. *Int. J. Electr. Electron. Eng.* **2012**, *1*, 2231–5284. [CrossRef]
35. Sahoo, S.; Nikovski, D.; Muso, T.; Tsuru, K. Electricity theft detection using smart meter data. In Proceedings of the 2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, USA, 18–20 February 2015; pp. 1–5.
36. Memisevic, R.; Zach, C.; Pollefeys, M.; Hinton, G.E. Gated softmax classification. *Adv. Neural Inf. Process. Syst.* **2010**, *23*, 1603–1611.
37. Nagi, J.; Yap, K.S.; Tiong, S.K.; Ahmed, S.K.; Mohammad, A.M. Detection of abnormalities and electricity theft using genetic support vector machines. In Proceedings of the TENCON 2008—2008 IEEE Region 10 Conference, Hyderabad, India, 19–21 November 2008; pp. 1–6.
38. Jiang, R.; Lu, R.; Wang, Y.; Luo, J.; Shen, C.; Shen, X. Energy-theft detection issues for advanced metering infrastructure in smart grid. *Tsinghua Sci. Technol.* **2014**, *19*, 105–120. [CrossRef]
39. McLaughlin, S.; Holbert, B.; Fawaz, A.; Berthier, R.; Zonouz, S. A multi-sensor energy theft detection framework for advanced metering infrastructures. *IEEE J. Sel. Areas Commun.* **2013**, *31*, 1319–1330. [CrossRef]
40. Veillette, M. Process for Detecting Energy Theft. U.S. Patent No. 9,013,173, 20 April 2015.
41. Lo, C.H.; Ansari, N. CONSUMER: A novel hybrid intrusion detection system for distribution networks in smart grid. *IEEE Trans. Emerg. Top. Comput.* **2013**, *1*, 33–44. [CrossRef]
42. Cárdenas, A.A.; Amin, S.; Schwartz, G.; Dong, R.; Sastry, S. A game theory model for electricity theft detection and privacy-aware control in AMI systems. In Proceedings of the 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 1–5 October 2012; pp. 1830–1837.
43. Das, R. A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Syst. Appl.* **2010**, *37*, 1568–1572. [CrossRef]
44. Weiss, G.M. Mining with rarity: A unifying framework. *ACM Sigkdd Explor. Newsl.* **2004**, *6*, 7–19. [CrossRef]
45. Seiffert, C.; Khoshgoftaar, T.M.; Van Hulse, J.; Napolitano, A. RUSBoost: Improving classification performance when training data is skewed. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.
46. Zhao, W.; Zhang, Z.; Wang, L. Manta ray foraging optimization: An effective bio-inspired optimizer for engineering applications. *Eng. Appl. Artif. Intell.* **2020**, *87*, 103300. [CrossRef]
47. Meng, X.B.; Gao, X.Z.; Lu, L.; Liu, Y.; Zhang, H. A new bio-inspired optimisation algorithm: Bird Swarm Algorithm. *J. Exp. Theor. Artif. Intell.* **2016**, *28*, 673–687. [CrossRef]
48. Uhl, X.M.; White, K. *Sylvia Earle: Oceanographer and Conservationist*; The Rosen Publishing Group, Inc.: New York, NY, USA, 2019.
49. Helfman, G.; Burgess, G.H. *Sharks: The Animal Answer Guide*; JHU Press: Baltimore, MD, USA, 2014.