Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Electricity Theft Detection in Smart Grids Based on Machine Learning Techniques

Hussein K. Imran[1*], Muhammed Salah Sadiq Al-kafaji [2]

[1] Al-Furat Al-Awsat Technical University, Al-Mussaib Technical College, Babylon, Iraq, hassonkadhim@gmail.com
[2]Al-Furat Al-Awsat Technical University, Najaf, Iraq, com.muh7@atu.edu.iq

[*]Corresponding author E-mail: hassonkadhim@gmail.com

https://doi.org/10.46649/fjiece.v4.1.12a.25.3.2025

Abstract. *Electric energy theft causes technical and non-technical losses in addition to serious destruction to energy suppliers and the power grid. Power quality and overall profitability are negatively impacted by energy theft. By combining data and energy flow, smart grids could potentially solve the issue of power theft. Power theft can be detected with the use of smart grid data analysis. However, there was room for improvement in the previous approaches' ability to detect energy theft. In this study, we introduced a machine-learning model for energy theft detection and smart grid classification based on active learning.*

For the utilities, electricity theft is a serious worry. The introduction of smart meters has increased the frequency of gathering data on residential energy usage, opening it up for sophisticated data analysis that was previously unattainable. In fact, the sustainability, efficiency, and dependability of the conventional energy infrastructure may be significantly increased by employing Smart Grid (SG) networks, which are essentially enhanced networks of linked things. The amount of data generated by the SG infrastructure is enormous and includes user power usage. By employing machine learning and deep learning algorithms on this data, it is possible to precisely identify individuals who steal electricity. electricity consumption dataset released by State Grid Corporation of China is used to train and test the models. The Dataset released by State Grid Corporation of China is used since It was the sole dataset accessible online. The total number of users is 42372. The Number of normal users is 38757. The Number of aberrant user or electricity thieves 3615. A set of six algorithms used in the research. the most accurate algorithm is Support Vector Machine SVM. The use of this technology in this field may improve the precision of energy theft detection as well as the issues and outcomes associated with power.

More specifically, the Deep Conventional Neural Network CNN model effectively completes two tasks: it distinguishes between non-periodic energy and, at the same time, preserves the general features of the power consumption data. The results of the testing show that the deep CNN model has the highest level of accuracy and performs better than previous models for energy theft detection.in this thesis a set of six which is algorithms (Random Forrst,Logistic Regression,Support Vector Machine,K Nearst Nighbour and NN Deep Learning) used, the best algorithm chosen.the work provided that the most accuarate algorithm is the SVM with accuracy of 91%.

Keywords: Energy Theft; Non-Technical Losses; Machine Learning; Deep Learning; Energy Theft Detection.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

1.INTRODUCTION

The electrical grid is one of the most important and intricate man-made systems in the modern world. The transition to a smart grid is presently occurring in tandem with a modification of the conventional energy grid due to the latest developments in observation, communication, control, and sensing. The ever-expanding distribution of distributed, renewable energy sources to achieve sustainability, self-healing, adaptability, and efficacy is known as a smart grid (SG). By utilizing advanced infrastructure alongside the existi ng power grid, the concept of SG is becoming more widely acknowledged[1]. The cyber-infrastructure enables the collection and analysis of data from several disparate scattered endpoints, such as smart meters, circuit breakers, and phasor determination units[2]. These grids often have a few upgrades that will increase their efficacy, durability, and capacity to supply homes and businesses with a steady supply of electricity. Furthermore, distributed generation (DG), distributed storage (DS), and wind, solar, and other renewable energy resources are all present in SG[3]. A smart electric instrument that measures energy use and provides more precise information than a traditional meter by driving and obtaining data over a two-way link is referred to as a "smart metering system[4]. As a result, smart metering grids employ smart sensors to enable businesses to manage and control smart gas (SG) by providing communication and information technologies[5]. In the modern human life, electrical energy is essenti al in modern life. Throughout the process of producing, distributing, and transitioning electrical energy, losses of energy frequently occur. Electrical energy losses are classifi ed as technical (TLs) and non-technical losses (NTLs) [6]. Additionally, the safety of the power system may be impacted by the actions of electrical thieves. For example, the high demand or electrical systems brought on by energy theft might result in fires, endangering everyone's safety. -For the safety and stability of the power grid, precise detection of electricity theft is therefore essential.

Power companies are now able to collect large quantities of frequency data on energy use from smart meters thanks to the installation of advanced metering infrastructure (AMI) in smart networks. This information is useful in identifying instances of electricity theft[7]. But there are always two sides to a story, and the AMI network creates a new avenue for energy theft assaults. These AMI assaults may be started in a everal ways, including using digital tools and cyberattacks. The main methods for detecting power theft are inspecting unlawful line diversions by hand, contrasting malevolent and benign meter data, and inspecting hardware or equipment issues. However, when verifying every meter in a system, these procedures are quite expensive and time-consuming.

Furthermore, these manual methods are unable to thwart cyberattacks.Numerous strategies have been proposed in recent years to address the aforementioned issues. -These techniques may be broadly divided into three categories artificial intelligence, game theory, and state-based models[8]. Numerous studies have been conducted on the subject of identifying electricity theft. Conventional methods of detecting energy theft include examining faulty meter setups or disconfirmations physically, connecting irregular meter readings to regular ones, and tracking the line where power is bypassed. These methods are expensive, time-consuming, and ineffectual. The existence of SGs increases the likelihood of solving cases of electricity theft. SGs are made up of traditional electricity networks, communication grids connecting smart devices (smart meters and sensors, for example) inside networks, and computation services to detect and control networks[9]. Employers and service providers are linked by the flow of energy and information in smart networks. This allows for the collection of a wide range of data via smart sensors or meters, including network conditions, energy usage, financing, and electrical energy cost[10]. To solve all of the aforementioned issues, the focus of this research article is on offering an efficient method for identifying electricity theft. In particular, Convolutional Neural Networks (CNNs) were first proposed in conjunction with recently proposed nature-inspired metaheuristic optimization algorithms to identify power thieves and analyze electricity usage data. Multiple convolutional layers, a pooling layer, and a fully connected layer make up the

CNN portion. In general, the CNN component is able to record the data on electricity consumption's periodicity. This model helps identify electricity theft by combining the power of the CNN component with the recommended algorithms.

2. RELATED WORK

This section provides a comprehensive literature assessment on many interconnected topics of the Smart Grid (SG). Firstly, we will discuss some essential studies on SGs. Additionally, a comprehensive analysis of methods for detecting electricity theft is provided. Next, a literature review of deep neural networks is conducted. Following that, a survey is provided on the implementation of Convolutional Neural Networks (CNNs) for the purpose of detecting instances of electricity theft. Lastly, the literature pertaining to the algorithms is presented Recently, there has been extensive research on privacy and security due to their enormous impact on the national economy, public security, and safety, which are heavily reliant on energy networks. While vulnerabilities in privacy and security are consistently being observed in grid protocols, technologies, and devices used in ensuring system-level safety and energy systems, the understanding of privacy concerns in smart grid metering and the risks of threats or theft by facilities is not always comprehensive. Subsequently, we will elucidate the latest survey publications in this field and highlight their outstanding contributions and unique characteristics. These publications are of surveys that have examined the topics of privacy and security within the field of SG. In, 2011 Line et al.. conducted a comparison of the standards for security for SG-communication networks and cable networks [10]. An enumerated several key cyber security concerns, such as confidence prototypes, security controlling, connectivity, customer confidentiality, software susceptibilities, and humanoid aspects. Additionally, some clarifications toward these difficulties be there suggested. (Deng) and (Shukla) in 2012, conducted a survey on the weaknesses and solutions, specifically focusing on the transmission subsystem in the “SG” [11]. Their focus was on identifying the vulnerabilities of Phasor Measuring Units “PMUs” and Wide Area Measuring System “WAMS” technology. An assaults were categorized addicted to 4-groups: movement investigation occurrence, denial of facility occurrence, malicious information addition attack, and high-level operations attack. The writers discussed the fundamentals of Phasor Measurement Units (PMUs), the use of PMUs for case approximation, and the application of PMUs in inverse attacks in 2013, Wang and Lu analyzed the security concerns in the grid of SG, which includes Home Area Networks (HANs), Advanced Metering Infrastructures (AMIs), and distribution and transmission subsystems [12]. They demonstrated the essentiality of security and assessed concerns regarding network security using case studies. The study primarily focused on cryptographic countermeasures that include the confirmation and management of the significant in several domains of SG. Their report included an intricate analysis of logical reasoning, as well as established procedures (such as distributed network protocol), in the subject of energy. However, since 2013, advanced and innovative safety procedures have been introduced, and it is necessary to familiarize oneself with them.

In 2013, Baig and Amoudi categorized the cyber-attacks of “SG” and solutions through 5-categories: Supervisory Control and Data Acquisition “SCADA”, Addition of Information and Reiterate Attacks, Occurrences of Smart-Meter, Network-based Attacks, and Physical Layer Attacks, which span home area nets, grids of the neighborhood, and wide area networks [13].

In 2014, Komninos et al. accessible “SG” and smart homebased protection investigation [14]. Generally supposed the statement within the surroundings of “SG” and smart-homebased are regarded as their risks of security. The article planned approximately illustrative worries and predicted the hypothetical effects as of “SG” to smart-home and equally. They brought an evaluation of the accessible collected works as the solutions of protection and limited the SG’s existing actions from 2009 to 2013. Komninos et al. deliberate some researches from the viewpoint of security countermeasure comprising secrecy, the serious reading of these schemes was not clarified.

In 2014, Mohassel et al. described an investigation on “AMI” advanced metering infrastructure [15]. They considered the central notions of “AMI”. They presented the physical and cyber security trials comprising confidentiality fleetingly. This article comprised part then provisions of safe keeping and secrecy in the grid of “AMI”. Yet, those writers do not cover full risk prototypical, and clarification on contemporary systems of safety and no articulated the secrecy preserving organizations.

3. MACHINE LEARNING

Machine Learning (ML) is the discipline of teaching computers to learn from and make predictions based on data without explicit programming. It sprang from the study of artificial intelligence (AI). Put another way, instead of strictly adhering to predetermined program instructions, machine learning algorithms work by constructing a model from sample inputs to generate data-driven predictions or judgments[16]. Data mining is a branch of machine learning that finds information that can be understood by humans from large databases, including clickstream data from websites or medical records. Furthermore, certain applications must be learned by experience, much like a human would, and cannot be designed by hand. Examples of this include handwriting recognition, computer vision, much of natural language processing (NLP), and even autonomous helicopters. To build software for every single consumer would be unthinkable, especially considering that people's tastes change over time. The ultimate goal of machine learning is to create artificial intelligence that can replicate the functions of the human brain. Tasks related to machine learning may often be divided into four groups based on the type of data that is available[17].

- Supervised learning: A series of "labeled instances" is fed into a supervised learning algorithm. In other words, the machine receives input variables, also known as features, and output variables. Its task is to learn how to create the right output given a fresh, unlabeled input. The training dataset is the collection of examples that the learning algorithm uses. When it comes to classification, the objective is to accurately classify new occurrences, with discrete labels serving as the desired qualitative output variables.
- Unsupervised learning: In unsupervised learning, there are no labels on training instances. The learning algorithm is requested to identify a hidden structure in the dataset rather than categorizing instances or creating a regression function. The most prevalent example is the clustering algorithm, which groups instances into distinct clusters based on similarities between them. Numerous applications, including computer clusters, social network analysis, market segmentation, and astronomical data analysis, employ clustering.

Semi-supervised learning: Only a limited portion of occurrences may be labeled due to the sometimes expensive labeling procedure; yet, readily accessible unlabeled data can be used for regression or classification. This strikes a balance between learning that is supervised and unsupervised.

- Reinforcement learning: Additionally, the machine is capable of generating actions that modify the state of a dynamic environment and provide rewards or penalties that it seeks to maximize or decrease. This type of machine learning is used to do certain tasks, such operating a car or competing against a rival in a game.

4. TYPES OF ENERGY LOSSES

One of the biggest issues facing power companies worldwide is energy loss in the distribution and transmission of electrical energy. Energy losses are often divided into two categories: non-technical and technical[18]. Technical loss is intrinsic to the transmission of electricity, which is generated by internal operations in the power scheme's components, such as the converters and transition lines[19].

The difference between technical losses and total losses, which are mostly caused by energy theft that occurs through physical attacks such line tapping, meter breakage, or meter reading manipulation, is known as the non-technical loss[15]. These forms of electrical fraud might result in income loss for power companies, as the losses incurred from electricity theft amount to around $4.500 million annually in the United States[20]. According to estimates, International Utility Corporation loses about $20 billion a year due to energy theft[21]. Furthermore, methods of energy theft affect the power system's security. For instance, the heavy burden of electric systems created by power theft may result in fires, which jeopardize people's safety. Therefore, the stability and safety of the electrical grid depend on the accurate detection of electricity theft. Power services achieved enormous amounts of electrical data consumption at a substantial smart meter frequency with the deployment of sophisticated metering infrastructure in SGs, which is beneficial for detecting electricity theft[22-23]. However, the sophisticated metering infrastructure grid opens the door to everal new ways to steal power. There are several ways to launch these attacks on the advanced metering system, including digital tools and cyberattacks. Important methods for identifying power theft include physically monitoring unauthorized line diversions, linking malicious meter readings to those of generous meters, and inspecting complex equipment or hardware. However, these methods incur high costs and need a significant amount of time to fully check all of the meters in a scheme. Also, these manual techniques are unable to stop cyberattacks. Over the years, many approaches have been proposed to address the aforementioned challenges. These methods are primarily divided into three categories: game theory-based, state-based, and artificial intelligence-based models[24].

5. METHODS OF DETECTION OF NON-TECHNICAL LOSS

For many years, electricity theft has been a major concern. Distribution System Operators (DSOs) have been conducting tests to identify instances of power theft; yet, the situation persists, and simple meter check methods are insufficient to identify the highest levels of fraudulent activity[25]. This section examines the most recent and feature-rich research paths on NTL revealing, and summarizing their key characteristics. Network-oriented, hybrid, and data-oriented are the three main categories into which NTL uncovering systems are categorized. Methods of data and network orientation are more broadly classified into subgroups based on the primary perception that underlies the covert detection of NTL. In addition to categorizing the various approaches, the researchers concentrated on the detection system of NTL's size, types of data, methods, features estimation metrics, and reaction times.

5.1 GROUPING OF NON-TECHNICAL LOSS DETECTION METHODS

A survey of scholarly studies on non-technical language detection (NTL) indicates that there is no one standard approach that is monitored to detect fraudulence. Aside from distribution network analysis, other areas of research that researchers assume a lot of things from include anomaly detection, machine learning, and cyber-security. The many NTL detection systems are organized into three major categories: hybrids, data-oriented, and network-oriented. The use of power grid data (such as network measurements or topology) sets data-oriented methods apart from network-oriented ones. Data-oriented methods use individual customer-related data (e.g., consumer type, energy use). Methods that use information from both sets are called hybrids. These prime groupings are depicted in Figure 1.1

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

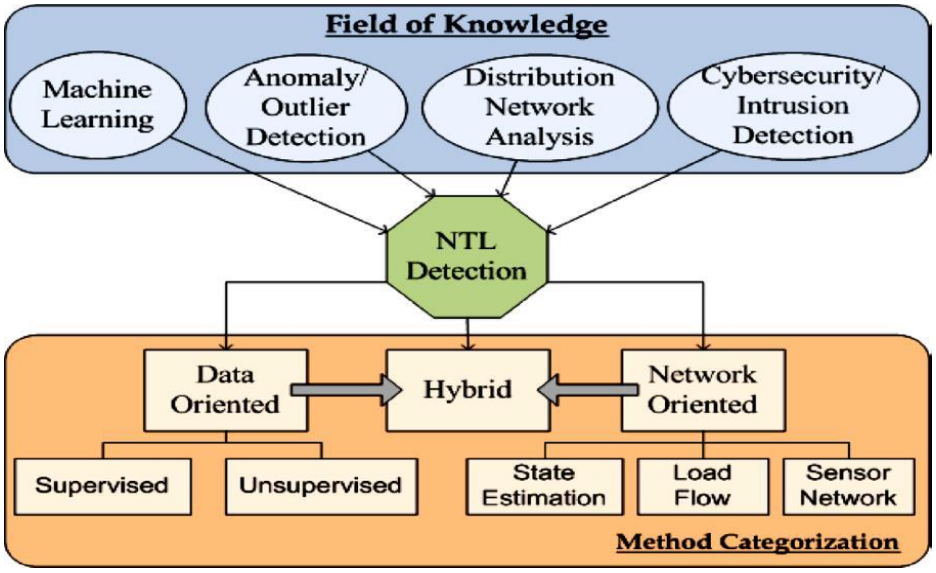

Fig. 1.NTL detection methods categorization [25].

Unsupervised and supervised data-oriented methodologies are further subdivided. Methods that don't use labels at all are considered unsupervised, whereas methods that include labels—both the acknowledged positive/fraud and negative/not-fraud classes—are considered supervised. Unsupervised approaches, however, do not make use of labels. Single-label methods are categorized as unsupervised and usually belong to the unsupervised anomaly detection category. These techniques are used when there are small sample sizes in one of the two classes (fraud class, for example). In addition to NTL detection, there are other fraud detection programs available, such as those for credit card fraud. In this stage, both labels are known; however, the absence of a positive label (fraud) precludes the application of supervised learning techniques. Network-oriented approaches often ignore labels since they are based on the study of networks and the physical laws that characterize such schemes. These methods are separated based on the primary perception/algorithm applied, such as the evaluation of the condition, load flow, or exceptional sensors for fraudulence detection. Concepts from the complete series mentioned above are used in hybrid approaches. For example, to identify NTL at the Medium Voltage (MV)/Low Voltage (LV) converter level, an estimate of the state manner may be used at the Medium Voltage (MV) level. Supervised learning may be applied to identify NTL at the customer level and then identify regions of the network that have NTLs [26]

5.2 DATA TYPE DESCRIPTIONS AND CLASSIFICATION

The many data types that have been used in the literature are organized in this section into wide categories in advance. The main goal of this categorization is to demonstrate that investigators may pick their method of NTL detection based on the data at hand and are not restricted to certain data categories when choosing their algorithm. Figure (2) displays the pyramid's data type

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

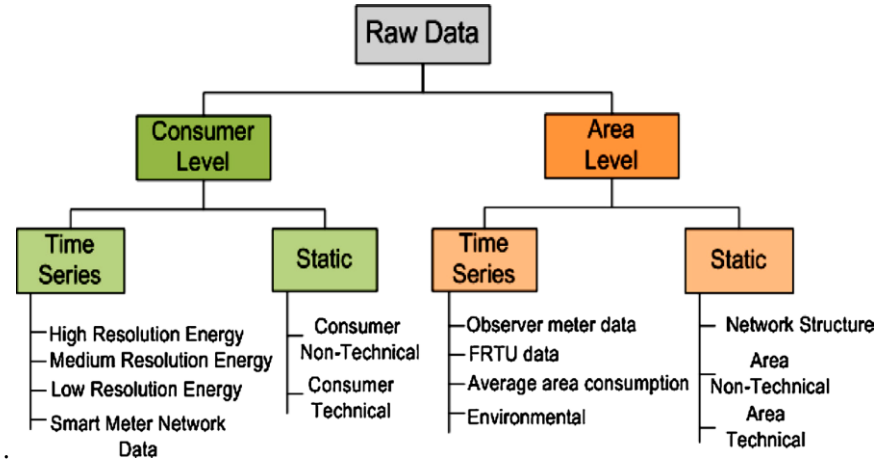

Fig. 2.Classification of data types for detecting NTL implementations [25].

Data was first scheduled based on where their physical resource was located. Information about a specific customer, such estimates of active energy, is classified as "Level of Customer" data; information about an area, like the network's topology, is classified as "Level of Area" data. Data that falls into any of the two categories previously described can also be further classified as a sequence of temporal and static data. As seen in Table (1), data may then be arranged into even more specific groupings.

Table 1. Data used for detection of NTL [25]

Level of Customer	Series of Time	Great resolution Power	estimations of energy active/reactive with a resolution of time equivalent or lesser than 10.0 min
		Moderate-resolution Power	estimations of energy active/reactive with a resolution of time amid 15.0 min. and 60 min
		Little resolution Power	estimations of energy active/reactive with a resolution of time of 30 days or extra
		Data of network of Smart meter	non-power data of network of Smart meter (voltage, alarms line resistance, amperage)
	Static	Customer practical	data supplying practical features of the substructure of customer installed power(kW), (request contracted (kW), level of voltage, converter of power (kVA), applications number, stages number, remote system usage for heating of space.
		Customer non-practical	Data expressing the behavior of the customer, e.g. review remarks, geographic region, the action of finance.
Level of Area	Series of Time	Spectator meter data	measurements of power, voltage, and amperage of a meter mounted on the side of LV of the secondary converter of the network of distribution to deliver overall feeder estimations
		Data of Remote	power, voltage, and amperage from RTUs

		technical unit (RTU)	set up in the network of MV or LV
		Average of consuming of area	Mean consuming of the observed region
		Ecological	generally temperature, but also might contain other parameters
	Static	Construction of Network	The topology of the network of LV or MV (may contain length and type of line). structure of Network associated data, such as the converter to which a customer is associated or the practical damages fraction
		Region practical	Data that describe a region from a practical point of sight (fraction of atypical customer per converter, number of converters in the region, a fraction of atypical customers in the region
		Region non-practical	Data that describe a region from a typical/financial point of opinion (fraction of residences with a group of waste, mean earnings, a fraction of borrowed residences, a fraction of literates, activities of movement opposite to fraudulence in the region, fraction of residences with water, mean number of residents, fraction of residences with roadways)

5.3 DEEP LEARNING AND NEURAL NETWORKS

A particular type of learning process is the foundation of the machine learning study field known as deep learning (DL). The endeavor to create a learning model at several layers is how deep learning is characterized. The highest levels take as input the results of lower levels, transform them, and constantly abstract more. This theory of learning levels is driven by the fact that the brain responds to outside stimuli to absorb information and learn[26].An arrangement of algorithms that resembles the structure of the human brain is called a neural network[27]. These networks use machine perception, data labeling, or data clustering to understand data. These networks identify patterns that are organized into a vector into which all data—text, audio, or image must be translated. The nodes that make up the layers of the neural network replicate how a neuron works in the human brain. These nodes are only the location of the calculations. to provide an input with a relevant value concerning the task that the network is trying to learn, a node mixes the data inputs with weights that either amplify or dampens the input (See Figure (3)).

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

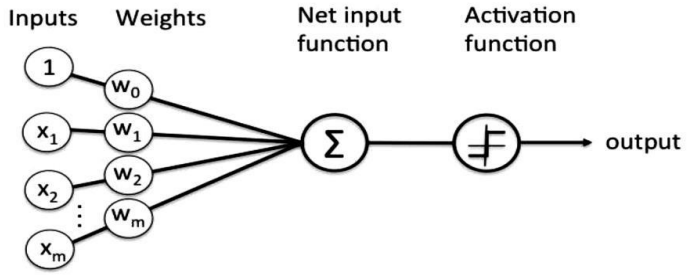

Fig. 3. Basic Neural Network Structure[23]

CNNs are essentially neural networks with one of their layers being the convolution operation rather than a linked layer . CNNs are a useful technology that has been used in situations where the input data, on which predictions are to be created, has a topology that is recognized as having a grid, such as an image (a 2-D grid) or a time series (a 1-D grid)[24]. CNNs are currently in charge of the machine vision industry. A layer of input, a layer of output, and several unseen layers make up a CNN. The pooling layers, convolutional layers, normalizing layers, and fully connected layers are often included in the unseen layers (Rectified Linear Unit-ReLU). Additional layers can be used for more complex simulations, as Figure (4) illustrates. Excellent performance in several computer vision and machine learning applications has been demonstrated by the CNN architecture. CNN uses abstract training and expectations. Because of its ongoing record-breaking efficiency, this CNN model is heavily used in new machine-learning implementations. These CNNs work based on linear algebra. The fundamental method for representing data and weights is matrix-vector multiplication[28].

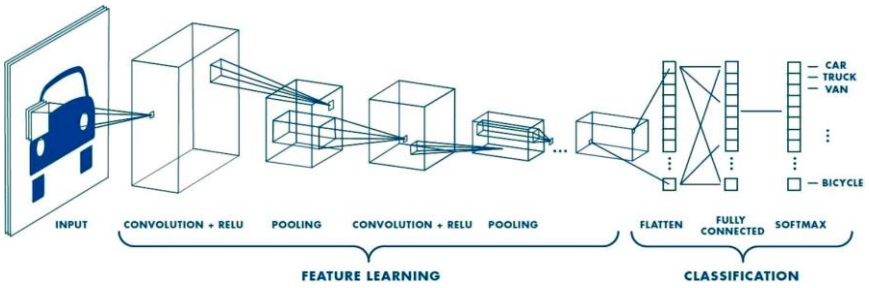

Fig. 4.Typical CNN architecture[28]

5.4 THE PROPOSED ELECTRICITY THEFT DETECTION SYSTEM

This chapter focuses on the design and implementation aspects of the proposed electricity theft detection system. A realistic electricity consumption dataset released by State Grid Corporation of China is used to train and test the models. This work is intended to identify electricity theft from the power consumption pattern of users, utilizing CNN-based deep learning and the group of algorithms. This classifier model is trained utilizing a dataset consisting of daily power consumption data of both normal and fraudulent users in a supervised manner by several steps. First, the data is prepared by a data-preprocessing algorithm to train the model. The preprocessing step also involves synthetic data generation for better performance. At the next step, the proposed model is hyper-tuned and finally, the optimized model is evaluated via the test data. The overall system is depicted in Figure (5).

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

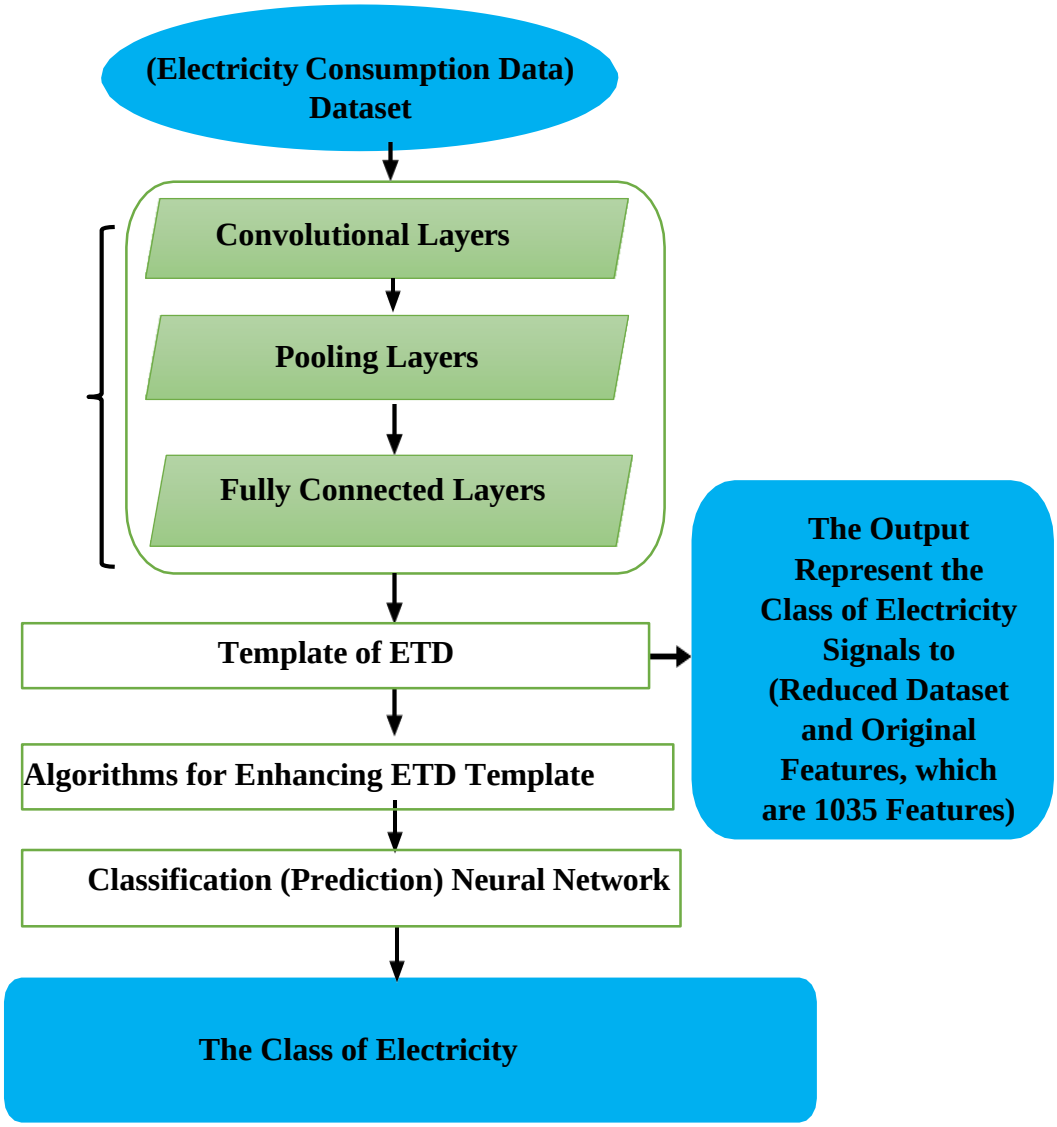

Fig. 5. The Proposed Electricity Theft Detection Model

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

5.5 THE ALGORITHIMS USED IN THIS PAPER FOR NTL DETECTION

1. Random forest Algorithim(RF)

We now need learning algorithms that scale with the volume of information while preserving enough statistical efficiency in order to benefit from the sheer magnitude of current data sets. Breiman (Breiman 2001) developed random forests, which are among the most effective data-handling techniques now in use for these situations. This supervised learning process follows the straightforward but powerful "divide and conquers" principle: sample portions of the data, develop a randomized tree predictor on each small piece, then paste (aggregate) these predictors together. It was inspired by the early work of Amit and Geman (1997), Ho (1998), and Dietterich (2000). The fact that forests have few tuning parameters and can be used to a wide range of prediction problems is what has substantially contributed for their appeal. In addition to being incredibly simple to learn and use, the approach is well acknowledged for its accuracy and its capacity to handle high-dimensional feature spaces and very small quantities. In addition, it is readily parallelizable, which means it can handle big real-world systems[26]. In machine learning, a Random Forest is analogous to a collaborative decision-making team. It creates an overall model that is more reliable and accurate by combining the predictions of several "trees," or separate models. The Random Forest Algorithm's broad appeal can be attributed to its versatility and ease of use, which allow it to efficiently address problems related to both regression and classification. The method is a useful tool for a variety of machine learning prediction tasks because of its strength in handling complicated datasets and mitigating overfitting.

The ability of the Random Forest Algorithm to handle data sets with both continuous variables—as in regression—and categorical variables—as in classification—is one of its most crucial properties. In tasks involving regression and classification, it performs better. We will learn how random forests operate in this lesson and apply them to a classification job[29].

The Random Forest Algorithm's Steps[30]

- Step 1: Each decision tree in the Random Forest model is built using a subset of features and a subset of data points. To put it simply, m features and n random records are selected from a data collection containing k records.
- Step 2: For every sample, a separate decision tree is built.
- Step 3: An output will be produced by each decision tree.
- Step 4: The final product is evaluated using either regression or classification-based majority voting or averaging.

Example of Random Forest[31]

- Think of the fruit basket in the illustration below as the data. Currently, everal samples are picked out of the fruit basket, and each sample is given its decision tree. Every decision tree will provide an output, as the illustration illustrates. The final product is decided upon by a majority vote. As you can see in the following graphic, the final result is determined to be an apple since the majority decision tree produces an apple as opposed to a banana.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

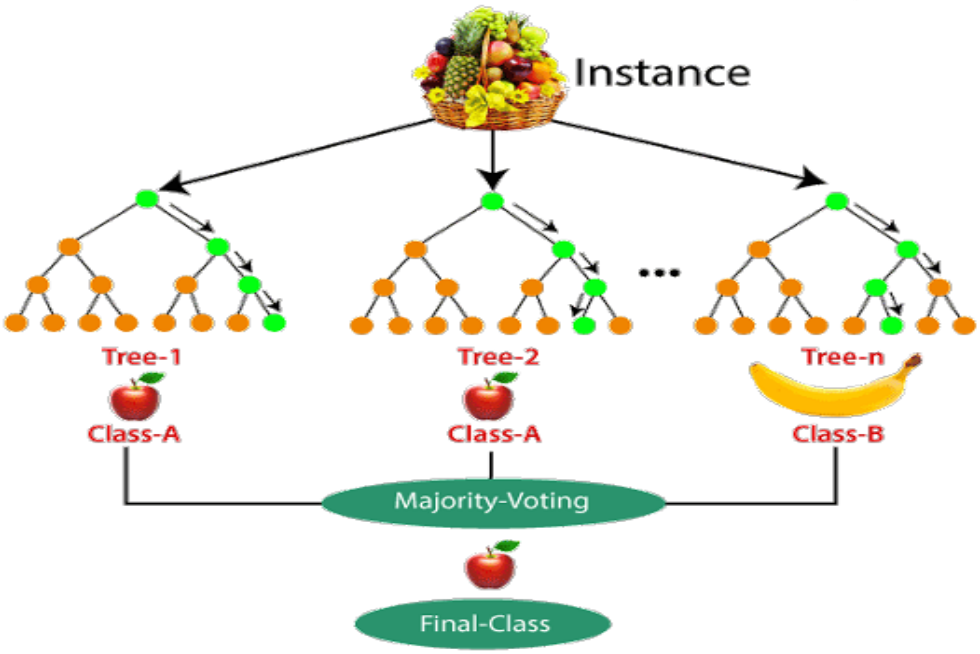

Fig. 6.Example of Random Forest[31]

Important Features of Random Forest[27]

- Diversity: Every tree is unique, and not all characteristics, qualities, or factors are taken into account while creating a single tree.

- Parallelization: Distinct data and characteristics are used to generate each tree separately. This implies that we may create random forests by utilizing the entire CPU.

- Train-Test split: Since 30% of the data in a random forest is never viewed by the decision tree, we don't need to separate the data for training and testing.

- Stability: Because the outcome is determined by average or majority vote, stability results.

2. DecisionTree

Systems that generate classifiers are among the most often used data mining approaches. Classification algorithms in data mining are can process large amounts of data. It may be applied to categorize newly available data, classify knowledge based on training sets and class labels, and create assumptions about categorical class names.Machine learning classification techniques comprise several algorithms. Fig. 7 shows the DT structure[32].

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

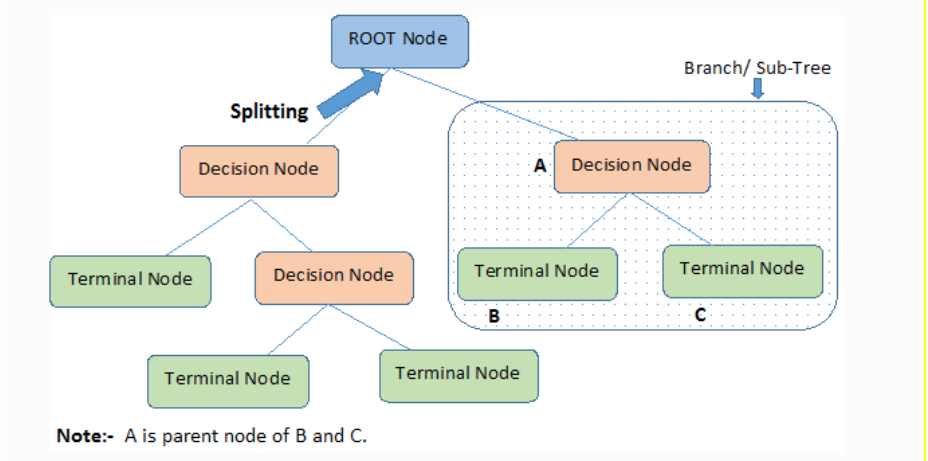

Fig. 7.Example Decesion Tree[28]

One of the effective techniques that is frequently utilized in many different domains, including machine learning, image processing, and pattern recognition, is the decision tree[33].

DT is a sequential model that effectively and cohesively combines several fundamental tests in which each test compares a numerical property to a threshold value[34]. DT is used mostly for grouping reasons. Additionally, DT is a classification model that is frequently used in data mining [39]. Every tree is made up of nodes and branches. Every subset specifies a value that the node can take, and every node represents features in a category that needs to be categorized.Decision trees have found various implementation domains due to their straightforward analysis and accuracy in a variety of data formats[35].

3. **Knighboors(KNN)**

As a nonparametric classification technique, K-Nearest-Neighbors (KNN) makes no assumptions on the elementary dataset.It is renowned for being both easy to use and efficient. The algorithm is one for supervised learning. The purpose of providing a labeled training dataset is to predict the class of the unlabeled data by classifying the data points into different groups. In classification, the unlabeled data's class is determined by a variety of factors. Mostly, KNN is employed as a classifier. It is used to categorize data according to nearby or nearest training examples in a certain area.This approach is popular because it takes little time to compute and is easy to implement. For continuous data, it determines its closest neighbors using the Euclidean distance. KNN is employed in datasets where the data is divided into distinct clusters to identify the class of the incoming input. When using data for a study when prior information about the data is lacking, KNN has greater significance[36].

4. **Logistic Regression**

In both electric and ultrasonic tomography, the primary objective is to recreate the cross-section known as the field of view. Often, we must designate cross-sections of regions that need to be imaged because they include concealed items. The imaging domain was originally designed as a specifically created pixel mesh, serving as finite elements, to identify these inclusions Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.[37].

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

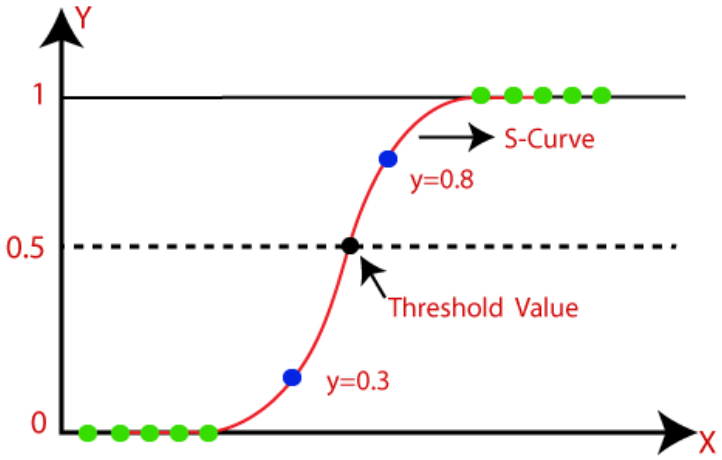

Fig. 8.Logistic Regression In Machine Learning[33]

5. Support Vector Machine

The definition of a support vector machine (SVM) is a machine learning algorithm that determines the boundaries between data points based on predefined classes, labels, or outputs. It uses supervised learning models to solve complex classification, regression, and outlier detection problems. This page describes the types, functions, and foundations of SVMs along with a few real-world applications[34]. The SVM algorithm's main goal, technically, is to locate a hyperplane that divides the data points into separate classes. The hyperplane is positioned so that the classes being considered are separated by the greatest margin[38]. Figure 9 displays the support vector representation.

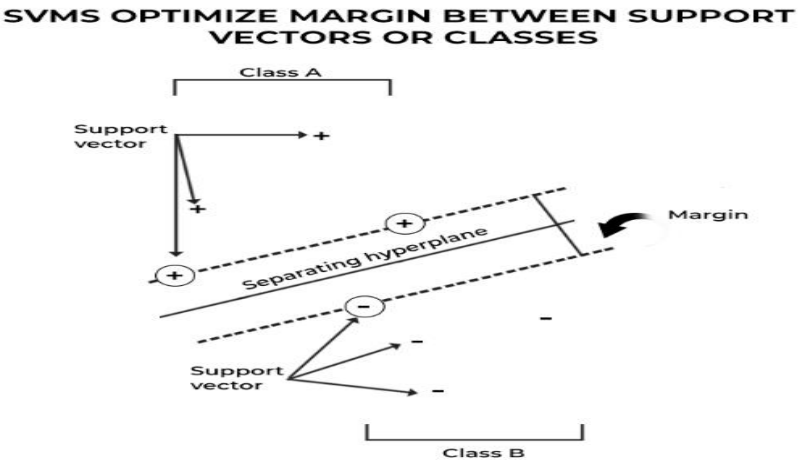

Fig. 9. SVMs Optimize Margin Between Support Vectors or Classes[34]

The greatest width of the slice that runs parallel to the hyperplane without any internal support vectors is referred to as the margin, as can be seen in the preceding picture. While it is simpler to design such hyperplanes for linearly separable issues, the SVM method attempts to optimize the margin between the support vectors in real-world circumstances, which might lead to inaccurate classifications for smaller portions of data points.

It is possible that SVMs were created to solve binary classification issues. But when computationally demanding multiclass issues become more common, some binary classifiers are built and coupled to create SVMs that can carry out these kinds of multiclass classifications using binary methods. Within the field of mathematics, a Support Vector

Machine (SVM) is a collection of machine learning techniques that utilize kernel functions and methods to modify the properties of data. The act of translating complicated datasets to higher dimensions in a way that facilitates data point separation is the foundation of kernel functions. The function adds more dimensions to map complicated data points, which simplifies the data bounds for non-linear issues.The data is not completely converted even with the addition of new dimensions because this might be a computationally demanding operation. This method, which is commonly called the "kernel trick," allows for the quick and low-cost translation of data into higher dimensions.The SVM algorithm's concept was initially identified in 1963 by Alexey Ya. Chervonenkis and Vladimir N. Vapnik. Since then, SVMs have been sufficiently well-known due to their ongoing broad implications in a variety of fields, including text classification, facial recognition, driverless automobiles, robotic systems, and protein sorting[39].

6. **NN (Deep Learning)**

Deep learning uses a neural network as its foundational technology. It is made up of layered structures made up of networked neurons or nodes. The nodes operate as a coordinated, adaptive system while processing data. They share comments on the work they have produced, grow from their errors, and keep getting better. neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain [40].

6. RESULT AND DISCUSSION

6.1 THE ELECTRICITY THEFT DETECTION SYSTEM UNDER PROPOSAL

The design and execution of the suggested power theft detection system are the main topics of this chapter. The models are trained and tested using a real-world power usage dataset made available by the State Grid Corporation of China. The goal of this work is to identify electricity theft by using CNN-based deep learning and a group of algorithms to analyze consumer power use patterns. This classifier model is trained in a supervised way over many steps using a dataset that includes daily power usage data of both legitimate and fraudulent customers. To train the model, a data preprocessing method first prepares the data. to improve speed, synthetic data is also generated during the preprocessing stage. The following action, At the next step, the proposed model is hyper-tuned and finally, the optimized model is evaluated via the test data

6.2 ELECTRICITY CONSUMPTION DATA

The State Grid Corporation of China (SGCC) provides a series of actual customer power use statistics that are used in the research. Table 3.1 displays the metadata information for this dataset. There are 1,035 columns and 42,372 rows in this dataset. The customer ID is shown in the first column, the "Flag" prediction pointer is shown in the second column, and the day columns begin in the third column and go up to the column (1,035). The dataset's metadata types are a collection of characters, numbers, and non-numeric (NaN) values, which are missing or incorrect. For each customer for more than two years, the amounts and missing or incorrect data indicate the quantity of electricity consumed (electricity signals). Furthermore, the metadata in the flag column refers to two types of consumers: normal and thief. The number of zeros in the "Flag" column denotes a normal electricity user, and the total number of them is (38,757). The thieves are represented by the number one in the "Flag" column; there are 3,615 of them in all. Ultimately, this indicates that

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

the figure (42,372) reflects the statistics on power use by users throughout a period of 1,035 days (January 1, 2014 to October 31, 2016). as displayed in table 2.

Table 2. The dataset's metadata about electricity theft

Synopsis	Value
Data gathering window timing	1st January 2014 – 31th October 2016
Total number of clients	42372
Number of normal users	38757
The number of abnormal users or energy thieves	3615

The provided dataset of energy consumption underwent many phases of modification to minimize its usage in the creation of electricity theft detection templates that employ diverse algorithms. The following steps are displayed:

- creating a new dataset by adding zeros to all of the previous dataset's null and Nan values.
- dividing a fresh dataset in half; the first half is utilized for testing (20%), while the second half is used for training (80%).
- removing the location and flag columns from the new dataset to reduce it. Since the suggested method won't require either complexity or time, the goal is to minimize those two factors.

6.3 GOOGLE COLAB (COLABORATORY) FOR MODLE TRAINING IN ML TO DETECT ETD:

You may run programs fully in the cloud using Google Colab, also known as Collaboratory by Google, a runtime environment based on Jupyter Notebooks.This is important because it allows you to train large-scale machine learning and deep learning models without the need for a powerful computer or fast internet. Deep learning aficionados will find Google Colab to be an excellent platform for testing fundamental machine learning models, gaining experience, and developing intuition about many elements of deep learning, including hyperparameter tweaking, preprocessing data, model complexity, overfitting, and more. In this model Google is used with six algorithms to compare the result and to determine what is the best algorithm to Electricity Theft Detection (ETD).

6.4 CONSTRUCTION OF AN ENERGY THEFT DETECTION (ETD) MODEL

The suggested Electricity Theft Detection model may be stated as follows: in the first phase, the dataset is processed through multiple modification procedures to decrease it, as discussed in section (4.1.2), and then the SM (sequence model) is built using Algorithm (3.4). The third stage is to create a prediction model (ETD model), which may be accomplished using two methods. The first procedure is performed using SM. The second operation employs the techniques specified in (3.4) (see Algorithm 3.4). The input is a Sequential Model with a reduced dataset, and the output is an electricity theft detection model with accuracy and loss, where this algorithm uses a set of fully connected layers, convolution layers, and a softmax layer to train and test the dataset (electricity consumption data).

In the proposed Model The Google Colab used to train and process the Data to Detect Electricity Theft.the model was ruined and a distribution of normal usage and theft was obtained as a result as shown in Figure (10)

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

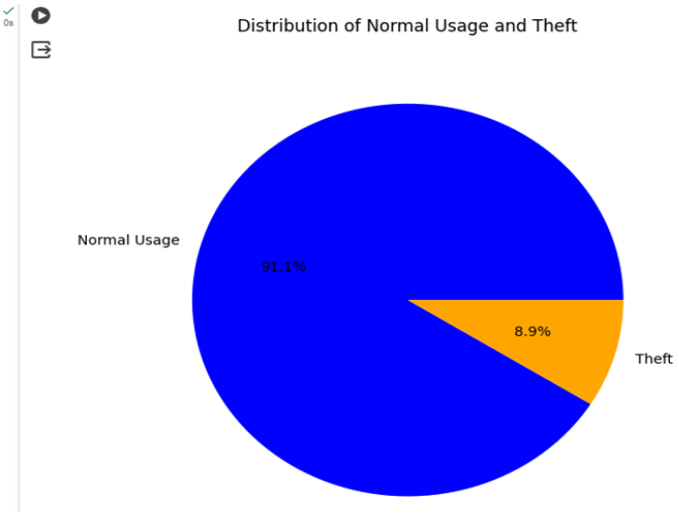

Fig. 10. Distribution of Normal Usage and Theft

Depending on the machine learning by applying the algorithms mentioned in (3.4) section the result was as follows

✓ *RANDOM FORREST ALGORITHM*

Table 3. shows the results the of Random forrest algorithm:

Accuracy	91,04570293094883
RMSE(The Root Mean Squared Error)	0.299237315003513
MAE (the average variance between the significant values in the dataset and the projected values in the same dataset)	0.08954297069051168
F1 (the harmonic mean of the precision and recall of a classification model)	[95.29772386 6.48508431]
AUC (Area under the ROC Curve)	51.67100959546927

The Confusion Matrix according to the random forest algorithm shown in figure (10)

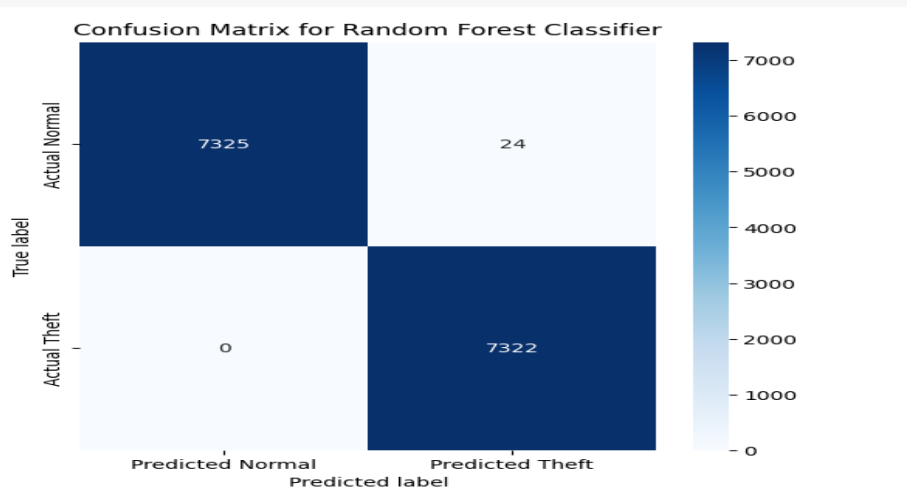

Fig. 11. The Confusion Matrix According to Random Forest Algorithm

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

✓ *DECISION TREE ALGORITHM*

Table 4. The Results of Decision algorithm are shown in table

Accuracy	85.3328365623447
RMSE(The Root Mean Squared Error)	0.38297732880230367
MAE (the average variance between the significant values in the dataset and the projected values in the same dataset)	0.1466716343765524
F1 (the harmonic mean of the precision and recall of a classification model)	[91.86358939 25.67652612]
AUC (Area under the ROC Curve)	59.53879377121369

The Confusion Matrix according to Decision tree Algorithm

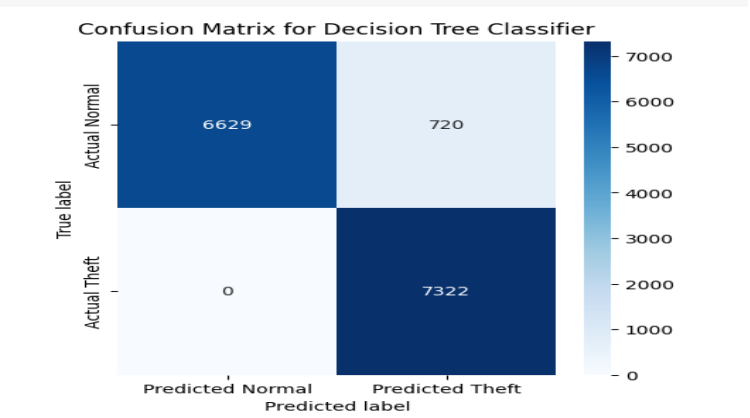

Fig. 12. The Confusion Matrix according to the Decision Tree Algorithm

✓ *KNNIGHBOORS ALGORITHM*

Table .5. Shows The Results of Knnighboors Algorithm

Accuracy	90.67312468951813
RMSE(The Root Mean Squared Error)	0,305399333831665
MAE (the average variance between the significant values in the dataset and the projected values in the same dataset)	0.09326875310481868
F1 (the harmonic mean of the precision and recall of a classification model)	[95.00698092 92.3508937]
AUC (Area under the ROC Curve)	59.50035444178938

the confusion matrix according to K-Neighbours

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

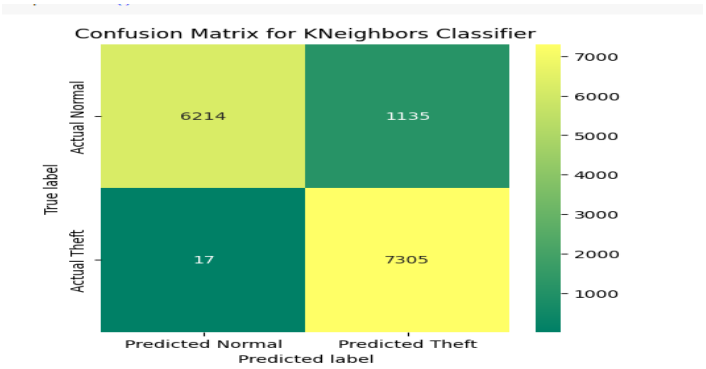

Fig. 13. The Confusion Matrix According to Knighboors

✓ *LOGISTIC REGRESSION ALGHORITHM*

Table .6. Shows The Results of Logistic Regression Algorithm

Accuracy	90.94634873323399
RMSE(The Root Mean Squared Error)	0.30089285911709535
MAE (the average variance between the significant values in the dataset and the projected values in the same dataset)	0.0905365126676602
F1 (the harmonic mean of the precision and recall of a classification model)	[95.21056435 17.4405436]
AUC (Area under the ROC Curve)	54,75648645420518

the confusion matrix according to Logistic regression

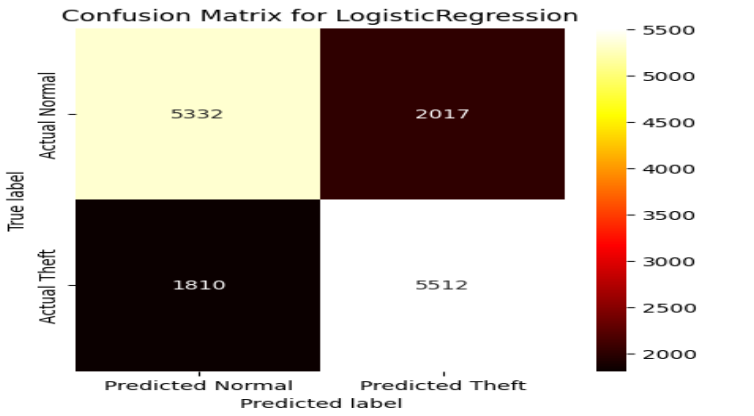

Fig. 14. The Confusion Matrix According to Logistic Regression

✓ *SUPPORT VECTOR MACHINE*

Table .7. shows the results of support vector machine algorithm

Accurcy	91,20715350223547
RMSE(The Root Mean Squared Error)	0.2965273427150443
MAE (the average variance between the significant values in the dataset and the projected values in the same dataset)	0.08792846497764531
F1 (the harmonic mean of the precision and recall of a classification model)	[95.38042433 5.34759358]
AUC (Area under the ROC Curve)	51.36868987986536

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

the confusion matrix according to suppert vector machine

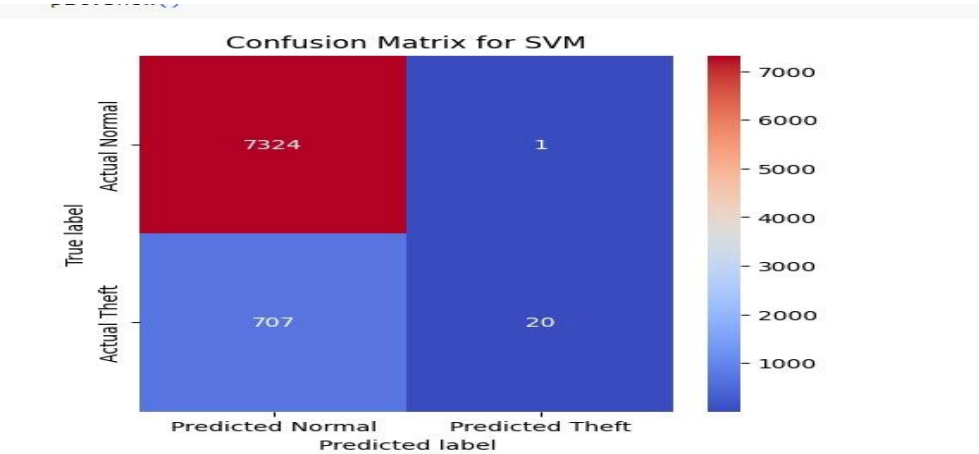

Fig.15. The Confusion Matrix According to Support Vector Machine

✓ *NN (DEEP LEARNING) Algorithm*

Table .8. shows the results of NN (deep learning algorithm)

Accuracy	90.94634873324234
RMSE(The Root Mean Squared Error)	0.32389265711712456
MAE (the average variance between the significant values in the dataset and the projected values in the same dataset	0.0895365145676602
F1 (the harmonic mean of the precision and recall of a classification model)	[94.21256435 16.9905440]
AUC (Area under the ROC Curve)	53,9564864620518

the confusion matrix according to the NN (DEEP LEARNING) Algorithm

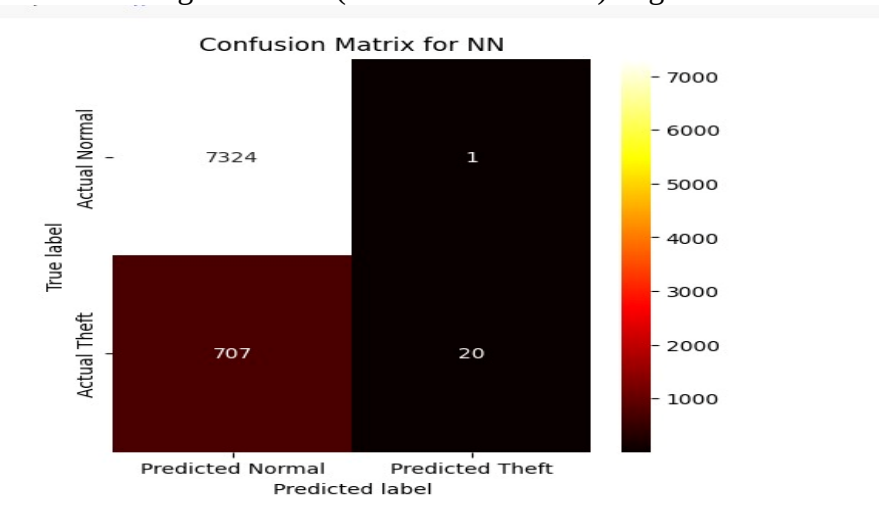

Fig.16. The Confusion Matrix According to NN (DEEP LEARNING) Algorithm

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Table .9. shows of all algorithms performance

Algorithm	F1-Score	MAE	Confusion Matrix	Ruc_auc_Score	Accuracy Score
Random forrest	[95.29772386 6.48508431]	0.08954297069051168	[7325 24 0 7322]	0.95104	91,0457029309
Decision tree	[91.863589392 5.67652612]	0.08954297069051168	[6629 720 0 7322]	0.998367	85.3328365623
K-Nighboors	[95.00698092 92.3508937]	0.09326875310481868	[6412 1135 17 7305]	0.921618	90.67312468951
Support vector machine	[95.38042433 5.34759358]	0.08792846497764531	[7324 1 707 20]	0.9993672	91,20715350223
NN (Deep Learning)	[94.21256435 16.9905440]	0.0895365145676602	[7324 1 707 20]	0.9014326	90.94634873324

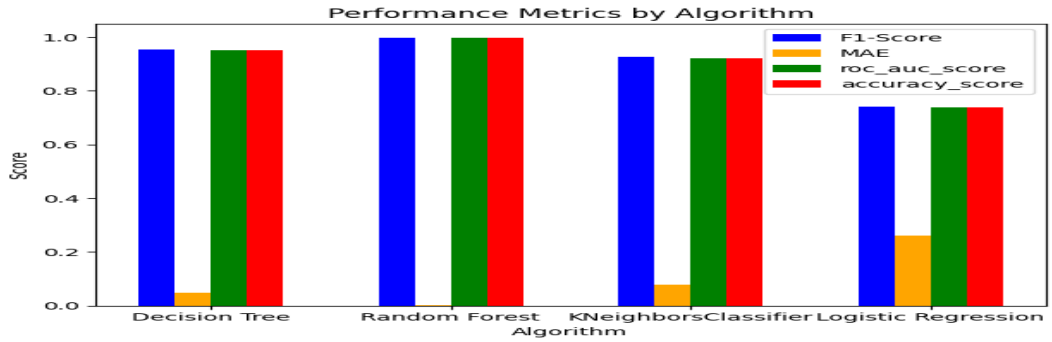

Fig.17 .The Performance Matrics Of All Algorithms

7. CONCLUSIONS

Using data on energy use from the State Grid Corporation of China, this study developed a Support Vector Machine-based model for detecting electricity theft, with a maximum detection accuracy of 91%is the best .Because around 25% of the cases of power theft could not be accurately classified, the models had limits. This might be because there is insufficient information on electrical thieves in comparison to non-thieves. Nonetheless, we made an effort to use data balancing strategies (oversampling and undersampling) to address this issue. The results of the studies demonstrate that the SVM system outperforms the majority of the other prediction systems that were evaluated, including K-Nearest Neighbors, Random Forest, and Logistic Regression.

REFERENCES

1. L. Faria, J. Melo, A. Padilha-Feltrin, Spatial-temporal estimation fornontechnical losses, IEEE Trans. Power Deliv. 8977 (2015), http://dx.doi.org/10.1109/TPWRD.2015.2469135, 1-1.
2. J. Nagi, K.S. Yap, S.K. Tiong, S.K. Ahmed, M. Mohamad, Nontechnical lossdetection for metered customers in power utility using support vectormachines, IEEE Trans. Power Deliv. 25 (2010) 1162–1171, http://dx.doi.org/10.1109/TPWRD.2009.2030890
3. J. Nagi, K.S. Yap, S.K. Tiong, S.K. Ahmed, F. Nagi, Improving SVM-basednontechnical loss detection in power utility using the fuzzy inference system,IEEE Trans. Power Deliv. 26 (2011) 1284–1285, http://dx.doi.org/10.1109/TPWRD.2010.2055670.
4. C.C.O. Ramos, A.N. De Sousa, J.P. Papa, A.X. Falcão, A new approach fornontechnical losses detection based on optimum-path forest, IEEE Trans.Power Syst. 26 (2011) 181–189, http://dx.doi.org/10.1109/TPWRS.2010.2051823.
5. C.C.O. Ramos, D. Rodrigues, A.N. de Souza, J.P. Papa, On the study ofcommercial losses in brazil: a binary black hole algorithm for theftcharacterization, IEEE Trans. Smart Grid 1 (2016), http://dx.doi.org/10.1109/TSG.2016.2560801.
6. C.C.O. Ramos, A.N. De Souza, A.X. Falcão, J.P. Papa, New insights onnontechnical losses characterization through evolutionary-based featureselection, IEEE Trans. Power Deliv. 27 (2012) 140–146, http://dx.doi.org/10.1109/TPWRD.2011.2170182.
7. C. C. O. Ramos, A. N. Souza, G. Chiachia, A. X. Falcão, and J. P. Papa, "A novel algorithm for feature selection using harmony search and its application for non-technical losses detection," Computers & Electrical Engineering, vol. 37, no. 6, pp. 886–894, 2011.
8. P. Glauner, J. A. Meira, P. Valtchev, R. State, and F. Bettinger, e challenge of non-technical loss detection using artificial intelligence: a surveyficial intelligence: a survey," International Journal of Computational Intelligence Systems vol. 10, no. 1, pp. 760–775, 2017.
9. C. León, F. Biscarri, I. Monedero, J.I. Guerrero, J. Biscarri, R. Millán, Variabilityand trend-based generalized rule induction model to NTL detection in powercompanies, IEEE Trans. Power Syst. 26 (2011) 1798–1807, http://dx.doi.org/10.1109/TPWRS.2011.2121350.
10. I. Monedero, F. Biscarri, C. León, J.I. Guerrero, J. Biscarri, R. Millán, Detection offrauds and other non-technical losses in a power utility using Pearsoncoefficient, Bayesian networks and decision trees, Int. J. Electr. Power EnergySyst. 34 (2012) 90–98, http://dx.doi.org/10.1016/j.ijepes.2011.09.009.
11. Wikipedia. Machine learning. [Online]. Available: https://en.wikipedia.org/wiki/Machine_learning.
12. Muniz, C. , Vellasco, M. , Tanscheit R. and Figueiredo, K 2009. Neuro-fuzzy System for Fraud Detection in ElectricityDistribution. In Proceedings of IFSA/EUSFLAT Conference.1096-1101.(Lisbon, Portugal, July 20-24, 2009).
13. R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, and X. S. Shen, "Energy-theft detection issues for advanced metering infrastructure in smart grid," Tsinghua Science and Technology, vol. 19, no. 2, pp. 105–120, 2014.
14. J. P. Navani, N. K. Sharma, and S. Sapra, "Technical and nontechnical losses in power system and its economic consequence in Indian economy," International Journal of Electronics and Computer Science Engineering, vol. 1, no. 2, pp. 757–761, 2012

15. S. McLaughlin, B. Holbert, A. Fawaz, R. Berthier, and S. Zonouz, “A multi-sensor energy theft detection framework for advanced metering infrastructures,” IEEE Journal on Selected Areas in Communications, vol. 31, no. 7, pp. 1319–1330, 2013.
16. P. McDaniel and S. McLaughlin, “Security and privacy challenges in the smart grid,” IEEE Security & Privacy Magazine, vol. 7, no. 3, pp. 75–77, 2009.
17. T. B. Smith, “Electricity theft: a comparative analysis,” Energy Policy, vol. 32, no. 1, pp. 2067–2076, 2004.
18. J. I. Guerrero, C. Le´on, I. Monedero, F. Biscarri, and J. Biscarri, “Improving knowledge-based systems with statistical techniques, text mining, and neural networks for nontechnical loss detection,” Knowledge-Based Systems, vol. 71, no. 4, pp. 376–388, 2014.
19. C. C. O. Ramos, A. N. Souza, G. Chiachia, A. X. Falcم‌o, and J. P. Papa, “A novel algorithm for feature selection using harmony search and its application for non-technical losses detection,” Computers & Electrical Engineering, vol. 37, no. 6, pp. 886–894, 2011.
20. P. Glauner, J. A. Meira, P. Valtchev, R. State, and F. Bettinger, “-e challenge of non-technical loss detection using artificial intelligence: a surveyficial intelligence: a survey,” International Journal of Computational Intelligence Systems, vol. 10, no. 1, pp. 760–775, 2017.
21. G. M. Messinis and N. D. Hatziargyriou, “Review of non-technical loss detection methods,” Electr. Power Syst. Res., vol. 158, pp. 250–266, 2018.
22. A. M. Giancarlo Zaccone, Md. Rezaul Karim, Deep Learning with TensorFlow: Explore neural networks with Python. 2017.
23. P. Doshi, S. Punktambekar, N. Kini, and S. S. Dhami, “Theft Detection System using Convolutional Neural Network and Object Tracking.” Vol-5 Issue-3 2019, IJARIIE-ISSN(O)-2395-4396.
24. N. Ketkar and E. Santana, Deep learning with python, vol. 1. Springer, 2017.
25. “Convolutional Neural Network - MATLAB & Simulink.” [Online]. Available: https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html. [Accessed: 06-Aug-2019].
26. Biau, Gérard, and Erwan Scornet. "A random forest guided tour." *Test* 25 (2016): 197-227.
27. Sruthi E R Understand Random Forest Algorithms With Examples (Updated 2024)
28. S. S. Nikam, “A comparative study of classification techniques in data mining algorithms,” Oriental journal of computer science & technology, vol. 8, no. 1, pp. 13–19, 2015
29. G. Stein, B. Chen, A. S. Wu, and K. A. Hua, “Decision tree classifier for network intrusion detection with GA-based featureselection,” in Proceedings of the 43rd annual Southeast regional conference-Volume 2, 2005, pp. 136–14
30. I. S. Damanik, A. P. Windarto, A. Wanto, S. R. Andani, and W. Saputra, “Decision Tree Optimization in C4. 5 Algorithm Using Genetic Algorithm,” in Journal of Physics: Conference Series, 2019, vol. 1255, no. 1, p. 012012
31. Mumbai, Apr. 2017, pp. 837–840, doi: 10.1109/I2CT.2017.8226246. P. H. Swain and H. Hauska, “The decision tree classifier: Design and potential,” IEEE Transactions on Geoscience Electronics, vol. 15, no. 3, pp. 142–147, 1977.
32. Taunk, Kashvi, et al. "A brief review of nearest neighbor algorithm for learning and classification." 2019 international conference on intelligent computing and control systems (ICCS). IEEE, 2019.
33. Psuj, G. Multi-Sensor Data Integration Using Deep Learning for Characterization of Defects in Steel Elements. Sensors 2018, 18, 292.

34. Pisner, Derek A., and David M. Schnyer. "Support vector machine." *Machine learning*. Academic Press, 2020. 101-121.
35. Widodo, Achmad, and Bo-Suk Yang. "Support vector machine in machine condition monitoring and fault diagnosis." Mechanical systems and signal processing 21.6 (2007): 2560-2574.
36. Thakolkaran, Prakash, et al. "NN-EUCLID: Deep-learning hyperelasticity without stress data." *Journal of the Mechanics and Physics of Solids* 169 (2022): 105076.
37. Fatih Ertam, Galip Aydin ”Data Classification with Deep Learning using Tensorflow”,(UBMK’17)2ndInternational Conference on Computer Science and Engineering
38. Tianmei Guo, Jiwen Dong ,Henjian LiYunxing Gao ”Simple Convolutional Neural Network on Image Classification ”,2017 IEEE 2nd International Conference on Big Data Analysis
39. “Smart grid and cyber security for energy assurance,” Nat. Assoc. State Energy Officials, Arlington, TX, USA, Tech. Rep. DE-OE0000119,2011.
40. S. Uludag, K.-S. Lui, W. Ren, and K. Nahrstedt, “Practical and securemachine-to-machine data collection protocol in smart grid,” in Proc.IEEE Conf. Commun. Netw. Security (CNS), San Francisco, CA, USA,Oct. 2014, pp. 85–90.