National College of Ireland

Electricity Theft Detection Using Machine Learning Algorithms: China

MSc Research Project
MSc in Data Analytics

Srushti Prakash Ghadge
Student ID: X20234082

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Srushti Prakash Ghadge
Student ID:	X20234082
Program	MSc In Data Analytics **Year:** Sept 2021-2022
Module:	Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	15/08/2022
Project Title:	Electricity Theft Detection Using Machine Learning Algorithms: China
Word Count:	9264 **Page Count** 24

I heat this momentarily that the information in this (my submission) is research information I conducted for this project. All information besides my contribution will be fully referenced and listed in the relevant bibliography section at the project's rear.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. Using other authors' written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Sign: **Date:**

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online submission to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project for your reference and if a project is lost or mislaid. It is not sufficient to keep a copy on the computer.	□

Assignments submitted to the Programme Coordinator Office must be placed in the assignment box outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Electricity Theft Detection Using Machine Learning Algorithms: China

Srushti Prakash Ghadge
X20234082

Abstract

This research is based on the investigation of the consumer's actual power consumption to identify the fraud consumer who steals the energy from the power lines, which can cause an imbalance in a power line which most affects the other consumers, utility companies, and the government, for this research purpose. These smart energy meters collected the consumer dataset at a regular time interval. The dataset collected by the state grid corporation of China (SGCC) is utilized in this research work to identify the fraud consumer from the bunch of honest consumers. As this research work goes into depth about the machine learning and deep learning algorithms and pattern recognition problem, it found that for the machine learning and deep learning model, when the dataset is imbalanced, it becomes ought for the machine learning and deep learning models to learn accurately about both class features and characteristics. So, to solve this issue, this research paper proposed the Hybrid algorithm of deep learning and machine learning boosting algorithm, which is known for managing imbalanced datasets, and the deep learning algorithm is utilized to extract more hidden features from the data so machine learning model can classify the information and correct fraud consumer data information can be collected. To balance and manage the imbalanced dataset, class weights were added during the model training instead of over and under-sampling the dataset, destroying the dataset's actual characteristics. This proposed methodology in this research successfully identifies the fraud consumer by around 96%, and the model can achieve an AUC score of approximately 96%. This research also included the implementation of models for comparison, including random forest, K means clustering, and Decision Tree.

1 Introduction

There are three processes of the power system operation generating, transmission, and distribution in the power system operation. Due to different supply voltage levels, considerable power is lost during these three processes. These losses are categorized into two parts technical loss and non-technical loss(Ponce et al., 2017). When the loss in the power line is analyzed due to technical tool failure and faulty operation of system equipment leads to the power loss, this type of loss is due to the impact of technicality known as a technical loss in power lines. This technical loss can be analyzed during the power audits and neglected by regular maintenance of machines and power lines. Meanwhile, non-technical losses are considered the illegal use of energy consumption by tempering with utility meters or bypassing the meter connection, and not regularly paying the electric bills are the types of non-technical losses (Haes Alhelou et al., 2019). This type of behavior is regarded as a crime, and stealing and tampering with the electricity utility company properties is considered a punishable crime in every country's law (Golden and Min, 2012). These types of illegal energy consumption use significantly affect the financial loss and quality of the power to the utility companies and countries' developments, there has been a significant rise in electricity stealing and a rise in fraud among consumers. In developing countries such as China, India, Turkey, Malaysia, Brazil, Pakistan, Bangladesh,

and many other developing counties, the illegal use of electricity is relatively high(Gaur and Gupta, 2016; Sudhir Kumar Katiyar, 2005).

In many developing countries, people steal power by hooking wires with poles or tampering with the meters; thus, utility company employees take a meter reading of power consumption at a lower value than the actual. The illegal use of energy consumption affects the operation of the conventional grid, and the power grid causes severe revenue loss and power loss. This electricity loss problem also affects the prosperous nation.

Fraud consumer energy consumption patterns are more abnormal than honest consumers. Manual energy theft detection is challenging because consumers have different power consumption (Lewis, 2015). It changes according to time, so taking the consumer misbehaviors from their power consumption by any misbehaviors algorithm is complex and involves many costs (Wong et al., 2021). In many countries, long-term use smart meters record periodic information about energy consumption. To solve this issue, many other researchers are utilizing AI-based approaches. This vital information about consumer power is considered a valuable key that can be analyzed and utilized using modern AI-based approaches (Deep learning and machine learning algorithms). The machine-learning algorithm can determine and identify the hidden features in the dataset based on the information (Gunturi and Sarkar, 2021). Further, the model can classify the data into fraud and honest consumer categories based on consumption information. However, a machine learning algorithm can extract hidden features. In this project work, we discuss those issues to identify fraud using a machine learning algorithm and try to solve the problems. Since it is a precise and lightweight solution, this project's ICT solution can be commercialized.

1.1 Research Question and Objectives:

In this research work, we are analyzing the electric theft dataset. Since electricity theft can result in abnormal patterns of electricity consumption, the data-driven electricity-theft detection approaches have received extensive attention recently due to the availability of smart-meter readings and electricity consumption data from smart grids. Next, we demonstrate how machine learning could be able to detect abnormalities in energy thefts.

RQ: "*To what extent can identification and detection of the electricity theft using the machine learning algorithms (Boosting algorithm, Random Forest, K mean clustering, Decision Tree) be used to minimize electricity fraud to support Chinese electricity boards"?*

Sub-RQ: "*Can daily and monthly analysis enhance the identification and detection of electricity frauds?"*

In this project, our primary goal is to detect fraud consumers who consume electricity illegally.
Obj1: Investigation of electricity theft detection methods.
Obj2: Data collection and pre-processing from smart meter data.
Obj3: Implementation and evaluation of electricity theft detection models
Obj3(a): Implementation, evaluation, and results of boosting-based hybrid learning algorithm
Obj3(b): Implementation, evaluation, and results of Random Forest.
Obj3(c): Implementation, evaluation, and results of k mean clustering
Obj3(d): Implementation, evaluation, and the results of Decision Tree.
Obj4: Comparison and evaluation of developed models.

Contribution: The following is a summary of this paper's main research contributions: In order to better evaluate electricity theft in smart grids, the study paper developed a model structure to better extract the information from the power consumption pattern dataset, segment the user, and manage and balance without adding extra synthetic data, conducted thorough tests using a sizable, accurate dataset of electricity consumption. According to experimental findings, our CNN model with XG Boost performs better than other techniques currently in use.

The report structure of the paper is as follows, the second section is about the analysis of various recent research works, and the type of methodologies they applied in their research work. In the third section, we discuss the basic idea of the XG boost algorithm, Random Forest, K mean clustering, and Decision Tree, the next section discusses the results and performance of the novel approach for solving this issue and the last section is about the research work conclusion and what is the future scope of the research work utilized in this paper.

2 Related Work on Electricity Theft (2014-2016)

This section of the research paper is based on information about how the other recent researcher worked on the problem of electric theft detection and analyzed the critical evolution of those methods. The plans utilized for this research work are finalized based on these methods.

2.1 The Conventional Approach to the Theft Detection

The electric power is transmitted to the consumer via the electricity distribution network. In this complex network, uncertainties occur in the power distribution network, such as random error meters, uneven loads consumption, surge supply voltage, etc. (Comden et al., 2019). For monitoring and control proposed in the electric power distribution system, the state estimation method (Zhang and Han, 2020), alongside the digital technology capable of providing critical information about the distribution network.

In the research work (Zhang et al., 2020), the interval state estimation model is proposed to formulate various uncertainties and unbalanced variables of the distribution network. It is an interval arithmetic model. The interval state estimation model provides the upper and lower bounds of the state variables. They are using this interval state estimation model to monitor the distribution network bus and distribution system by applying the state estimation method in the power network to analyze and create a math model of all real-time information and prior information about evaluating multiple uncertainties for energy theft detection(Huang et al., 2013).

The state estimation approach for the power system was analysed based on the weight most minor square technique because the weight least square technique is quite attractive and efficient. There is also one central uncertainty voltage control which plays a vital role in the operation of the distribution network.

The research work developed and simulated the state estimation algorithm in MATLAB software (Pegoraro and Sulis, 2011). The simulation considers the math model's voltage amplitude and phase angle state variables. Also, state estimation is based on a math relation between amplitude and angle of the voltage and measurement from the distribution system. So, the state estimation model formulates reactive power compensation to voltage control with uncertainties (Yang et al., 2020).

Based on the similar concept of the control system state-space model for electricity theft detection in distribution system using the wireless communication system. In the electricity

distribution network, a high percentage of the economic loss and power loss is due to electricity theft. Thus, proposing an electronics approach for electricity theft detection allows sensors to detect at a remote location. The wireless communication system is based on a global system for mobile communication. So, a global mobile communication technology system collects all the consumers' meter readings without human iteration. Thus, the wireless communication system gives detection alert messages automatically to the electricity utility companies which alerts electricity utility companies to eliminate various issues such as bypassing meter readings or tampering with the meter reading (Mufassirin et al., n.d.).

However, one of the flows of this method is that the state-space model requires much more detailed information about each piece of equipment the consumers utilize. A lot more information to process for the standard system is not possible. Hence, the operating cost is very high, and the information is not easy to capture from the consumers.

2.2 Critical Review of Machine Learning-Based Approaches

A machine learning algorithm handles a large amount of data to train machines efficiently by predictive analysis. A supervised Machine learning algorithm-based model can take label data during training to generate predictions. There are many real-world problems where machine learning algorithms are utilized, such as false consumer detection for credit card fraud, spam detection in banking, healthcare prediction, crime prediction, text mining, web mining, electricity theft detection, credit card fraud detection, and many others (Xia et al., 2022).

Based on the concept of machine learning algorithms, fraudulent electricity consumption by the consumer is detected using a support vector machine learning technique. Support vector machine utilized for both the classification and regression problem. So, the electricity theft detection problem needs to classify the consumer into two categories: honest and fraud. Thus, using a support vector machine dramatically deals with the electricity theft erection problem for organizing a consumer as per their load profiles pattern. Support vector machine-based classification model has a set of kernels. This set of kernels has a mathematical formulation, and using this formulation, mapping various consumers with the specific feature in high dimensional space. This support vector machine-based classification model distinguishes the honest or fraudulent consumers using their load profiles pattern (Nagi et al., 2010).

The support vector machine-based classification model and decision tree are used in supervised machine learning algorithms to solve classification-based problems. Using a decision tree and support vector machine-based classification model detects the actual consumption of the consumer and improves the detection accuracy. The decision tree-based prediction model calculated the consumer expected power consumption using various attributes like the number of appliances, the number of people, temperature, weather etc. after that support vector machine-based classification model takes both input and output of the decision tree as an input and classifying a consumer into honest and fraud consumer from their energy consumption (Jindal et al., 2016).

The support vector machine and random forest models are the most frequently used supervised learning algorithms to solve classification-based problems. Decision trees and support vector machines obtain acceptable results, but existing line loss data and consumer behavior for energy theft are not fully utilized. Thus, to solve this type of problem random forest-based classification model was utilized. Random forest build based on the bagging approach. A random forest algorithm is also used for both tasks, such as regression and classification. So,

using a random forest-based classification model, classify the honest and fraudulent consumers from the various consumer behavior (Hu et al., 2020).

Supervised machine learning algorithm-based methods are utilized to detect non-malicious changes in energy consumption patterns due to changes in the weather, decreases in the usage of some appliances, decreasing number of people, and temperature. This type of issue is solved using the k-means algorithm. Using the K-means algorithm remove the cluster from the dataset improves the false positive rate and detection accuracy by 11% and 94%, respectively (Jokar et al., 2016).

2.3 Investigation of Deep Learning Algorithm Approach for Solving the Problem

Many researchers worked on the machine learning or deep learning algorithm-based model to avoid the issues and instability of the unsupervised learning-based clustering algorithms for the energy theft detection problem. Further model is trained with training data set, which is a ratio around 70:30 or 80:20 of the total dataset, and once the model is trained and to check the model prediction performance based which is different from the training dataset. Deep learning is also represented as a deep neural network and its subset of a broader family of artificial intelligence and machine learning method based on a multi-layer perceptron network. The deep learning algorithm is the process of feature engineering (Khan et al., 2020b).

Multi-layer perceptron method for detecting fraud consumer for energy based on their power pattern detection mainly including multiple neurons layered neural network, feed-forward neural network, and recurrent neural network. A multi-layer perceptron network is employed with input, hidden, and output layers. A multilayer perceptron network was utilized with a backpropagation algorithm to decrease the mean squared error between the desire and actual output for detecting fraud. First, selecting the mean for the consumer's energy consumption and maximum demand is applied to a multi-layer perceptron network (C. Costa et al., 2013).

A deep and convolutional neural network was utilized to capture periodicity in the energy consumption pattern and get accurate detection accuracy. The feed-forward neural networks and deep learning algorithm-based convolutional neural networks are trained with the scaled data set of labeled classed energy consumption patterns and features based on daily power consumption for the feed-forward neural network and CNN. The classed dataset must have equal data points and the same number of attributes. A deep convolutional neural network takes daily power consumption as an input in the 2-D matrix, which is the spatial feature. Thus, convolutional neural networks capture spatial features and efficiently lean the periodicity in the energy consumption pattern (Zheng et al., 2018).

The approach is frequently used for the electricity theft detection combination of the convolutional neural network and the long short-term memory. For feature extraction and classification purpose convolutional neural network is widely used. CNN-based LSTM models are implemented to predict electricity prices. The CNN-based model was used for the feature extraction, and LSTM was used for the sequential dataset; thus, the combination of CNN-LSTM was used for the binary classification problem. They use LSTM to detect fraud consumers for energy theft from the historical power consumption data (Hasan et al., 2019).

The various literature analysis on machine learning and deep learning algorithms utilized to identify the fraudulent consumer from the bunch of honest consumers is a challenging task. The DL and ML algorithms use the classification-based approach to detect the fraud customer.

The basic fundamental of the classification approach is required the class data in equal amounts. For example, suppose the classification is classified into two classes: fraud and innocent consumers. In that case, the machine learning and deep learning model need the dataset of both types in an equal amount of data points for each category. For this problem of fraud detection in a real-life scenario, the dataset is highly biased toward the honest consumer data, so when the model is trained to detect the fraud consumer, then the model is also very biased towards the honest consumers, so the prediction accuracy to see the fraud consumer is significantly less. To balance and modify the information, many research works utilised the SMOTE-based synthetic data balancer in simulation work. All the other research works related to the SMOTE are discussed in the next section.

2.3.1 SMOTE-Based Approach for Solving the Imbalanced Data Issue

To solve the issue of the imbalanced dataset and missing or zero value in the energy consumption dataset, many researchers utilized this method of synthetic data process for the minority class for the oversampling of the dataset (SMOTE). Applied SMOTE function to the training dataset synthesizing a new sample from the minority class. And this function oversample that complex pieces tend to misclassify the whole dataset. Using this method, they balance the fraud consumer data with the actual consumer dataset so that the supervised machine learning algorithm, such as support vector machine, XG-boost, random forest, decision tree, etc., is trained with an equal dataset. The accuracy of model detection and fraud identification is up to the training and validation dataset mark. But the synthetic data generation process is entirely unknown to the developer during the training process. Based on the available sample data of the minimum sample, it will create the synthetic dataset; There are maximum chances that it will change some of the fraud consumer functional information characteristics in the synthetic dataset (Pereira and Saraiva, 2020).

2.4 Identified Gaps and Conclusion

This literature survey section includes information about the various methodologies utilized to solve the issue of electricity fraud detection based on their power consumption dataset. The first section had information about the conventional approach, then machine learning and deep learning-based system, and also discussed the flows in the process to overcome this flow of machine learning and deep learning algorithms SMOTE based algorithm is utilized to balance the imbalanced data by oversampling or under-sampling dataset and balance the fraud and honest consumer information in the dataset. However, by doing this, the original character of the dataset and information tampers. When the real-time data is fed into the model for the prediction, the prediction accuracy of unknown data is very low (Douzas et al., 2019). So, to overcome this issue of machine learning, deep learning algorithms with the SMOTE algorithm, this research work proposed a novel approach for fraud detection with the Deep and wide CNN algorithm hybrid with the XG-boosting algorithm with the class weight-based method to avoid the biases of the model. The next chapter is included the fundament information about the proposed approach and the collected dataset information.

3 Electricity Theft Detection Methodology Approach and Design Specification

3.1 Electricity theft Methodology Approach

This section is based on the various research methodologies implemented to solve the issue of electricity fraud consumer detection based on their power consumption patterns. To solve this

issue of fraud detection, this research paper proposed a deep learning and machine learning hybrid model alongside the class weight balancing to manage the problem of imbalance classification dataset. The research paper proposed a Depth with broadness of the CNN model structure to better extract the information from the power consumption pattern dataset and segment the consumer and is managed and balanced without adding extra synthetic data by counter balancing the weights in the XG-boost algorithm. Machine learning algorithms like Random Forest, K mean clustering, and Decision Tree is also implemented. This chapter also included information about the data collection and the basics of exploratory data analysis.

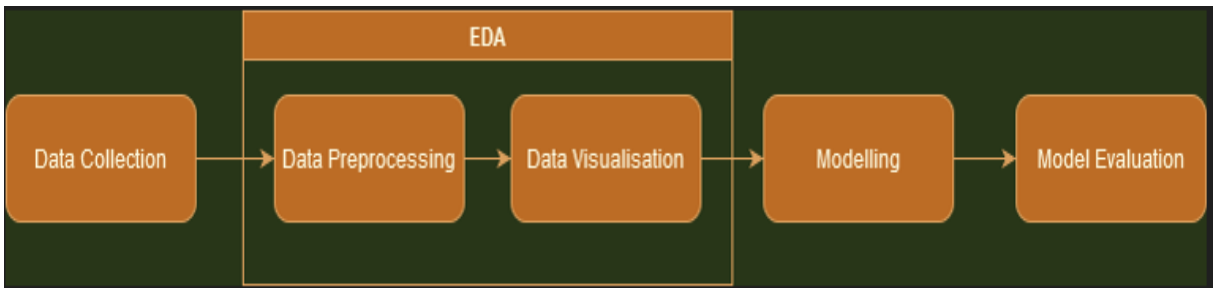

Figure 1: Methodology for electricity theft detection

3.2 Data Collection and All Essential Information about EDA on the dataset

3.2.1 Data Collection

For this research, A large dataset consists of information regarding actual power consumption by the consumer from China. It is a real-time smart meter dataset collected by the state grid corporation of China (SGCC) (Chinese government, 2017). The collected dataset by the state grid corporation of China from 1st Jan 2014 to 31st Oct 2016 (Approximately two years and eight months). The dataset file is set as .csv files. The dataset of China's state grid corporation represents the daily power consumption data of 42,372 consumers for 1035 days. The dataset is split into two types of consumers honest and fraud consumers (Pan, et al., 2020). We utilize the state grid corporation of China data to detect fraudulent consumers for electricity consumption.

3.2.2 Exploratory data analysis and Data Visualization

Exploratory data analysis in data science is a critically important process. From the EDA process, the machine learning and profound learning model developer get a better idea about the problem which is tried to solve using this deep learning and machine learning model parameters selection and analysis of the dataset information and further what type of processing is still required to performed machine learning, and deep learning prediction performance is improved. EDA process is further analyzed into two parts A) Dataset processing and B) Data visualization.

Data Processing:

Processes the Null values: We have found the missing values or NA's values from the dataset. These missing values are due to a faulty meter or random error while collecting the meter reading. So, this missing value or NA's value remove or replaced with zero. However, we manage this absence or the NA's weight with the median of each consumer power usage.
Process the date timeline in proper sequence: The dataset date is not in the continuing format from the dataset. So, please keep it in the continuing design using the date-time module.
Process the null or zero value in the consumer: by eliminating zero or null value in the consumer and processing the dataset. The dataset has information about the daily power

consumption of the consumer. Using this information creates a mean dataset of power consumption.

Data visualization:

The dataset must be split into equal classes for the training model to avoid model biases. Data visualization class samples in the training dataset are biased toward the honest consumer and only 8.9% of fraud consumers of the total number of consumers. To model better performance accuracy, balanced the dataset based on the class weight method. Based on the outcome, the model is trained on the dataset.

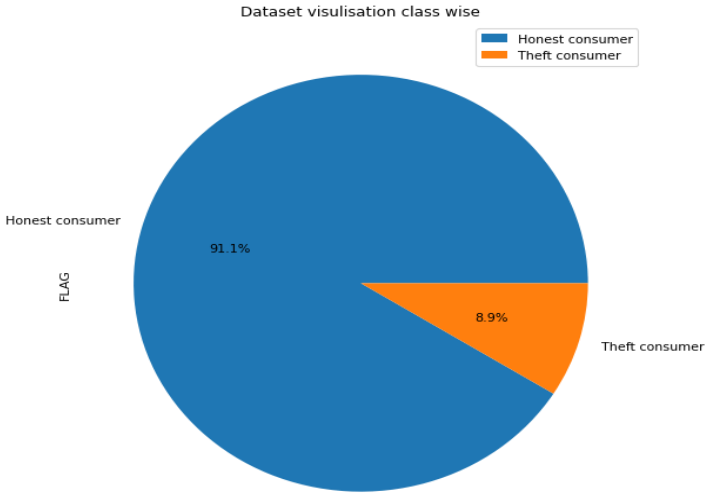

Figure 2: Data visualization class samples in the training dataset

3.3 Modeling

After all necessary data pre-processing, the models can be applied on this dataset. During the implementation phase, we create models to categorize various attributes depending on the structure of the model using the appropriate machine learning method. A scientific method was used to assess the model's accuracy. For our suggested project, we are applying the Boosting algorithm with CNN, Random Forest, Decision Trees, and K means clustering model models. These models were seen to perform better than every other model in the earlier works

3.4 Algorithm Performance Evolution Method

3.4.1 Confusion Matrix

A confusion matrix is an n * n tabular visualization with two dimensions of predicted value versus the actual value. Where n represents the target variable which is positive or negative. A confusion matrix between the expected value and real value has four attributes: true positive, false positive, true negative, and false negative. True Positive is the combination that states that the model's actual and predicted value is positive. True Negative is the combination states that the model's actual and predicted value is negative. False Positive is the combination that represents a type 1 error, and in this combination model actual value is negative, but the model predicts the positive value. False Negative is also represented as a type 2 error, and in this combination model's actual value is positive, but the model predicts the negative value.

3.4.2 Classification Report

With the Confusion matrix's help, it evaluates various classification reports. This is defined as follows: Accuracy, precision, Recall, and F1-Score.

Accuracy:

The accuracy of the models can be calculated as the ratio of the number of correct predictions to all kinds of predictions made by the classifier.

Accuracy = (True Positive + True Negative) /
(True Positive + False Positive + True Negative + False Negative)

Precision:

It can be defined as the ratio of the number of correct predictions to all kinds of positive classes predicted correctly by the classifier model.

Precision =True Positive / (True Positive + False Positive)

Recall:

It is defined as the out of the total positive classes number of positive classes provided by the model.

Recall = (True Positive) / (True Positive + False Negative)

To evaluate both precisions and recall simultaneously, F1-Score was utilised. I remember, equal to the accuracy, F1-Score must be 1; it is the best value of F1-Score.

F1-Score = 2*(recall*precision) / (recall + precision)

AUC- ROC Curve:

The AUC favorable plots between the actual favorable and false-positive rates at various threshold values and separate the noise from the signal. Thus for the different threshold setting AUC-ROC curve was utilized. ROC represents the probability, whereas AUC represents the separability. For better performance, the model value of AUC should be higher.

3.3 Design Specification

The dataset is prepared for the other data pre-processing procedure, which is crucial for the machine learning model, once it has been scaled and converted. The scaled dataset will then be divided into training and testing halves. To prevent the model from overfitting, 70–80% of the whole dataset is partitioned and randomly mixed into the training dataset. The remaining 30–20% of the dataset will be used to validate the machine learning model. It examined the accuracy of the trained model's performance on the incoming unknowable dataset. After the dataset is divided into its class populations, the weights for the fraud customer and the honest consumer are chosen. The training dataset and class weightage are applied to the machine learning-based supervised learning algorithm to identify the fraud consumer based on the dataset features once the class weightage is chosen.

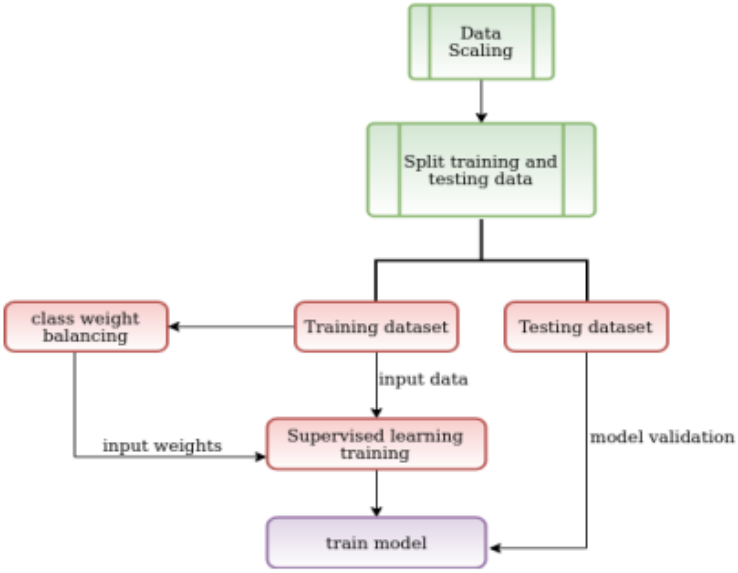

Figure 3: Design Specification

3.5 Conclusion: The methodology is transformed to meet the requirements of this project, and this research uses that modified methodology. This methodology is applied to the project design process flow and data from smart meters is collected. The project uses the design architecture as it was intended to be utilized. The next part demonstrates the implementation, evaluation, and results of models for fraud detection.

4 Implementation and Evaluation of Electricity Theft Detection Models

4.1 Introduction

This section includes the implementation, evaluation, and results of different models used to identify fraud detection. The different machine learning models such as Boosting algorithm, Random Forest, K mean clustering, and Decision Tree are implemented for that purpose. Pre-processing is also completely detailed in this section along with the model implementation. The models are evaluated using the accuracy, confusion matrix, and classification reports in the project. Python was used to implement the models because of how simple it is to use and the vast array of machine learning choices that are available with a variety of libraries.

4.2 Data Pre-processing

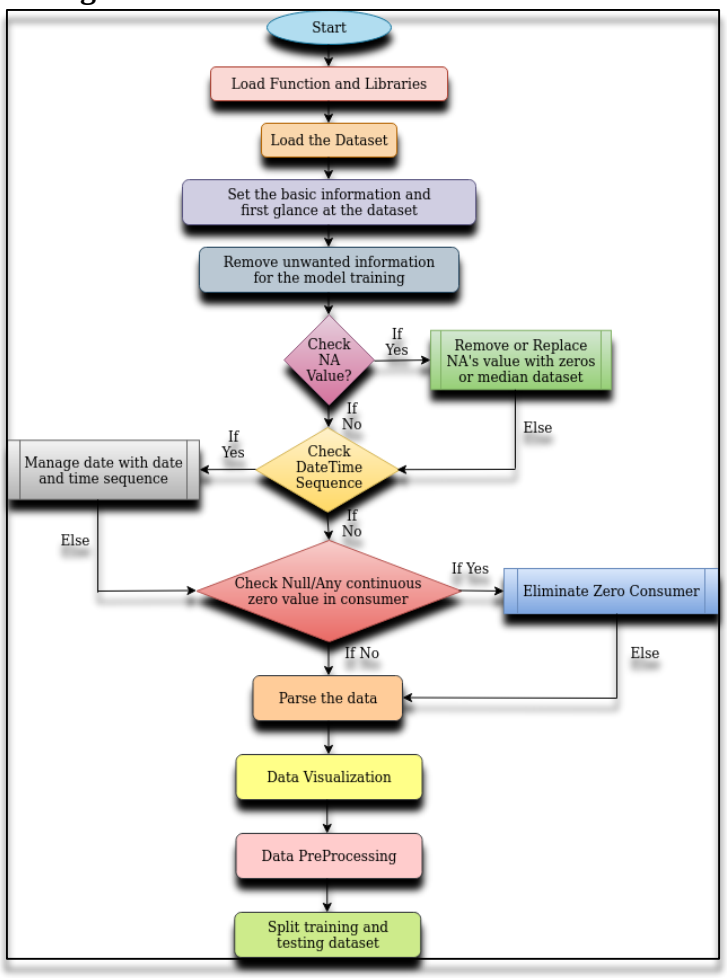

Figure 4: Programming flow-data processing

Figure 4 represents the Programming flow for the data processing steps. Using python programming language to load the function and libraries. Then, it loads the dataset file is set as .csv using the Pandas library, Checking the basic information about the dataset using the data info function and the first glance at the dataset. Then, Remove unwanted information from the dataset for the model training. Check for null or missing values. If yes, remove or replace with zeros or median of the dataset or else proceed further. Check the date-time sequence. If yes, manage the date-time line using the date-time function. Else, move on. Check null or any continuous zero value in the consumer. If yes, then it eliminates zero consumers. Else parse the data. Data visualization has been done using the Matplotlib library. Data pre-processing has been done using Scikit learn library. Split training and testing dataset for the random shuffle.
In Data pre-processing the data standardization and scaling are essential for the data pre-processing for numerical features. Many machine learning algorithms require data scaling because variables measured at different scales impact numerical stability. So, in the data standardization process, putting the other variables on the same scale. Standardization can be achieved by using the standard scaler function of Sklearn. The normal scaler function ensures the standardizing and transforming of the data so that the mean is 0 and the variance is one, bringing all the variables at the same magnitude. This approach computes it separately for each variable in the dataset. In the random shuffle of data pre-processing shuffles the data to prevent the overfitting of the classification process. Using random shuffling prevents the model's

overfitting and increases the performance accuracy. In this step, samples are randomly mixed, and the label position changes according to the class of each instance regularly.

The train test split method is used for the data pre-processing for splitting the dataset to perform an unbiased model and identify problems of overfitting or underfitting. Splitting data using a train-test-split function from machine learning library Sklearn of model selection package. This approach minimizes the model's biases in the evaluation and validation process. Once the data set is split into the training and testing set, it is applied to the deep learning model, and then the hybrid model with the xgboost model is utilised in the following diagram.

4.3 Implementation, Evaluation, and Results of CNN Model with XG-Boost Algorithm

4.3.1 Implementation

Choosing the optimization algorithm: there are five types of algorithms momentum, Nesterov momentum, Adagrad, RMSProp, and Adam are used to increase the speed of the convergence in terms of history from last weight updates. The Adam optimizer is used as a default choice in this research work. Using Adam optimizer increased the convergence speed and automatically adapted the effective learning rate.

xgboost model parameters and class weights selection is as follows. Scale-pos-weight is used for the imbalance classes and helps fast convergence. In the high imbalance classes value greater than zero should be used. The booster parameter is used to run the model at each iteration. There are two types of model: one is based on the tree model, and the second is based on a linear model. Verbosity mode is when silent is set to 0 activated and 1. It means no messages are printed. Validate-parameter is used to perform the validation process of the input parameter for checking parameter is used or not. it is used for parallel processing to run the boost number of parallel threads.

Machine learning-based Boosting algorithm and Class weight balancing approach

Boosting is a method used in machine learning-based models to reduce training error in predictive analysis by building a solid classifier from the number of weak classifiers. Facilitating process is also represented as a sequential ensemble. Enabling technique corrects the training error made by the previous model by adding some weights to the model.

Extreme gradient boosting algorithm: Extreme gradient boosting builds upon the supervised machine learning decision tree-based ensemble algorithm, and it is performed for both problem classification and regression. XG-Boost (Extreme Gradient Boosting) is an open-source machine learning library that improves model performance and speed. Extreme gradient boosting is a promoting method in ensemble machine learning. The critical feature for extreme gradient boosting is weight quantities. Using weight quantities, get the best node split. This boosting method splits data into a smaller dataset for the process of parallelization. After that, it Optimizes the gradient boosting machine by parallelizing, pruning the tree, utilizing missing values, and avoiding overfitting. Thus, extreme gradient boosting built a decision tree for automatically selecting features and improving the algorithm's efficiency (Chen and He, n.d.).

Class weight balancing method: The machine learning and deep learning algorithm-based models are not very useful with the imbalance dataset. If training a model with the imbalance dataset, it is biased toward the majority class, thus using overcome this type of issue using different approaches such as oversampling and under-sampling, and class weight methods. The class weight approach gives equal importance to all the data courses.In the machine learning models, the model weights are designed to be set by a 1:1 ratio, which means equal importance for both model classes. This weight of the model can also be redefined and applied to the model weights. These weights are re-assigned to the model by the following math equation.

Updated weights = Total number / (Number of classes * Total number of samples in class (i))

Based on this fundamental, the biasness of the model is reduced for the imbalanced classification dataset (Fernando and Tsokos, 2021).

4.3.2 Evaluation and Results

The proposed model was implemented as the procedure discussed above chapter on design and implementation. I considered four cases in which the models were trained and implemented to detect fraud consumers based on their power pattern information for this research work.

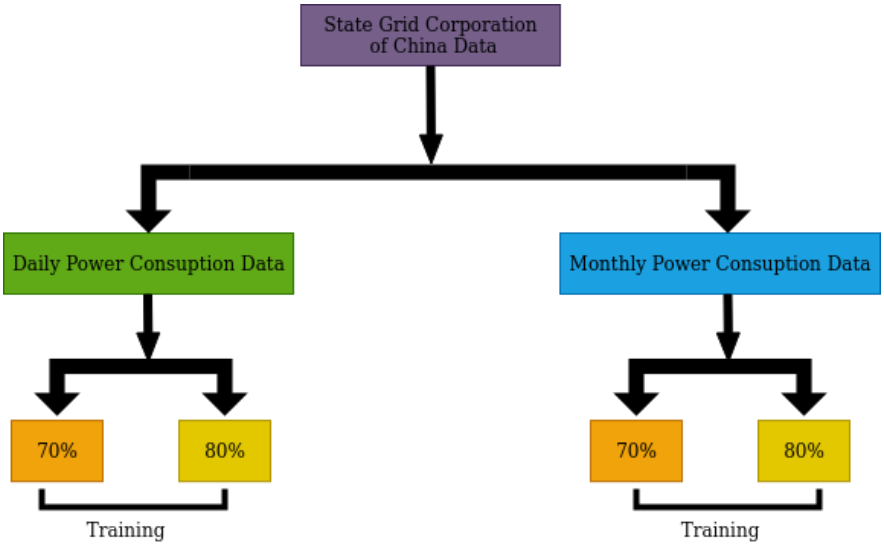

Figure 5: Case study considered in this research work

Case 1-Daily power consumption data utilized for training with 70% of the total dataset:

The smart meter in the power grid system can collect power consumption information at regular intervals based on the utility companies' terms and conditions. In this research work, the dataset utilized has information on the daily power used by the consumers. This research aims to identify the daily power consumption pattern and differentiate between honest and fraudulent consumers based on their power consumption. The proper data time interval is managed and manipulated in the research work dataset for the feature extraction. So based on this information, the dataset is split with a 70:30 training ratio. A 70% training with two types of consumers (honest and fraud). In this scenario, around 96% classification accuracy is achieved in the research work.

	precision	recall	f1-score	support
0	1.00	0.96	0.98	10969
1	0.70	0.97	0.81	1109
accuracy			0.96	12078
macro avg	0.85	0.96	0.90	12078
weighted avg	0.97	0.96	0.96	12078

Figure 6: Classification report for Daily power consumption data utilized for training with 70% of the total dataset

Case 2-Daily power consumption data utilized for training with 80% of the total dataset:

In the second case, different fluctuations daily for power consumption and differentiate between honest and fraud consumers based on the time interval of power consumption. So

based on this information, the daily power consumption is calculated using an 80:20 training ratio. An 80% training ratio for daily power consumption is around 96%.

	precision	recall	f1-score	support
0	1.00	0.95	0.98	7297
1	0.69	0.97	0.81	755
accuracy			0.96	8052
macro avg	0.84	0.96	0.89	8052
weighted avg	0.97	0.96	0.96	8052

Figure 7: Classification report for Daily power consumption data utilized for training with 80% of the total dataset

Case 3-Monthly power consumption data utilized for training with 70% of the total dataset:

In the third case, the electricity consumption over the month (30-day average) and the result between the fraudulent and honest consumers based on the 30-day power consumption using a training ratio of 70:30. The entire monthly power consumption data gets the exact data for some days. So, using this information, the model accuracy of 74% for monthly power consumption.

	precision	recall	f1-score	support
0	0.96	0.75	0.84	25710
1	0.21	0.70	0.32	2470
accuracy			0.74	28180
macro avg	0.59	0.72	0.58	28180
weighted avg	0.90	0.74	0.80	28180

Figure 8: Classification report for Monthly power consumption data utilized for training with 70% of the total dataset

Case 4- Monthly power consumption data utilized for training with 80% of the total dataset:

In the fourth case, the monthly power consumption of the consumer's data sample and the difference between honest and fraud consumers are based on the time interval of power consumption. However, for some consumers, the fraud consumer power consumption pattern is similar to the honest consumer, so to reduce this, monthly power is selected to train the model. In this case, the 80:20 training ratio for monthly power consumption obtained an accuracy of around 74%.

	precision	recall	f1-score	support
0	0.96	0.74	0.84	7297
1	0.22	0.69	0.33	755
accuracy			0.74	8052
macro avg	0.59	0.72	0.58	8052
weighted avg	0.89	0.74	0.79	8052

Figure 9: Classification report for Monthly power consumption data utilized for training with 80% of the total dataset

4.4. Implementation, Evaluation, and Results of Random Forest

4.4.1 Implementation:

The random forest technique is a powerful classifier made up of a number of weak classifiers. Each sub-training data set creates a decision tree, and the sub-training data set is constructed

using bootstrap and put back sampling. Following that, the decision trees, which are independent of one another, are trained using distinctive parameters that are randomly chosen. Finally, votes are cast on the output of various decision tree classifiers to determine the final outcome. In the implementation of the Random Forest Model, the training set and test set are separated from the effective data set that was obtained through data preprocessing. The random forest algorithm's decision trees are trained using the training data, and its parameters are adjusted to increase classification accuracy using the test set. Following the extraction of the distinctive parameters, decision trees are constructed using the training set. Additionally, these decision trees are used to construct the random forest, and each decision tree classifier casts a vote to determine the classification outcome. Lastly, use the trained model to determine whether there is behavior indicative of electricity theft under the known assessment index. The dataset here is imbalanced, since we need the dataset with balanced classes we use SMOTE to get the data ready for the Random Forest Classifier. On unbalanced datasets, classifiers do not perform upto standards due to presence of bias. The majority class or classes are ultimately appropriately classified at the expense of the minority class. Synthetic minority oversampling is one method for dealing with unbalanced datasets (SMOTE). By building synthetic examples from combinations of the nearby minority cases, this program creates new sample data.

4.4.2 Evaluation and Results

For the purposes of this research, the random forest technique was applied to 2 situations in which models were developed and used to identify consumers who were fraudsters using data from their power patterns.

Case 1- Daily Power Consumption Data: The dataset used in this study contains data on how much power consumers use on a daily basis. This study attempts to determine the daily pattern of power use and distinguish between honest and fraud consumers based on their power consumption. In the study work dataset, the appropriate data time interval is maintained and adjusted for feature extraction. As a result, the dataset is divided with a training ratio of 70:30 using this information. 70% training with two different consumer types (honest and fraud). In this situation, the research study has a 91% accuracy rate.

```
Model AUC score: 0.5359010338624524
Model accuray: 91.4018250471995

Model classification report
              precision    recall  f1-score   support

           0       0.92      1.00      0.95     11586
           1       0.62      0.08      0.14      1126

    accuracy                           0.91     12712
   macro avg       0.77      0.54      0.55     12712
weighted avg       0.89      0.91      0.88     12712


True negative: 11533
False negative: 1040
True positive: 86
False positive: 53
Model miss classficiaiton rate: 2.5795336543000094
```

Figure 10: Classification report for Daily Power Consumption Data (Random Forest)

Case 2-Monthly Power Consumption Data: In this case instance, a training ratio of 70:30 was used to compare the results between fraud and honest consumers based on the electricity use over the course of a month’s 30-day average. The exact data is obtained for a few days for

the complete monthly power consumption data. So, based on this data, the model's monthly power consumption accuracy is 91.354**%.**

```
Model AUC score: 0.5344394180641241
Model accuray: 91.3546255506608
              precision    recall  f1-score   support

           0       0.92      1.00      0.95     11586
           1       0.60      0.07      0.13      1126

    accuracy                           0.91     12712
   macro avg       0.76      0.53      0.54     12712
weighted avg       0.89      0.91      0.88     12712


True negative: 11530
False negative: 1043
True positive: 83
False positive: 56
Model miss classficiaiton rate: 8.645374449339208
```

Figure 11: Classification report for Monthly Power Consumption Data (Random Forest)

4.5 Implementation, Evaluation, and Results of K mean

4.5.1 Implementation

The iterative K-means method attempts to divide the dataset into K unique, non-overlapping subgroups (clusters), each of which contains only one group to which each data point belongs. While keeping the clusters as distinct (far) apart as possible, it aims to make the intra-cluster data points as comparable as possible. By assembling them in groups, the bunching is done to select the type of consumer that is better qualified in their force use design. Since K-means is employed in exploratory data mining, it is necessary to review the clustering outcomes to ascertain whether clusters make sense. If some of the clusters are too small or too broad, the value of k may need to be increased. K denotes a user-specified parameter in the straightforward K-Mean clustering technique. Here, the "k" number of various clusters must be specified beforehand. For the K-Mean clustering procedure, the initial centroids should be carefully chosen. Every time a run is performed, a different result will be obtained due to the initial centroids being chosen at random. Due to this flaw, clusters differ from one another, and data items within clusters may differ between clusters. The arithmetic mean value can be significantly impacted by noisy data, making the simple K-Mean algorithm noise sensitive

4.5.2 Evaluation, and Results

Case 1-Daily Power Consumption Data:

Based on daily energy consumption by consumers included in the dataset utilized for this investigation, consumers' everyday patterns of power use aim to identify honest and fraudulent consumers. The proper data time interval is upheld and modified for feature extraction. K number of clusters is defined as 2 in which the data will be grouped. In this instance, accuracy is defined as 91.46%. The random_state parameter is set to 342.

```
Model AUC score: 0.49998709910467787
KMeans model for daily power consumption based  clustering accuracy 91.46606249409987
KMeans Model Classification Report
              precision    recall  f1-score   support

           0       0.91      1.00      0.96     38757
           1       0.00      0.00      0.00      3615

    accuracy                           0.91     42372
   macro avg       0.46      0.50      0.48     42372
weighted avg       0.84      0.91      0.87     42372


True negative: 38756
False negative: 3615
True positive: 0
False positive: 1
Model miss classficiaiton rate: 8.533937505900123
```

Figure 12: Classification report for Daily Power Consumption Data (K-mean)

Case 2- Monthly Power Consumption Data:

Depending on the electricity use over the course of a month's 30-day average, outcomes betw een fraud and honest consumers. For the whole monthly power usage data, the precise data is collected for a few days. The K value defined in the algorithm is defined as 2. The model's m onthly power consumption accuracy is therefore 91.47% according to this data.

```
Model AUC score: 0.5001383125864454
KMeans model for monthly power consuption based  clustering accuracy 91.47078259227793
KMeans Model Classification Report
              precision    recall  f1-score   support

           0       0.91      1.00      0.96     38757
           1       1.00      0.00      0.00      3615

    accuracy                           0.91     42372
   macro avg       0.96      0.50      0.48     42372
weighted avg       0.92      0.91      0.87     42372


True negative: 38757
False negative: 3614
True positive: 1
False positive: 0
Model misclassficiaiton rate: 8.529217407722081
```

Figure 13: Classification report for Monthly Power Consumption Data(k-Means)

4.6 Implementation, Evaluation, and Results of Decision Tree:

4.6.1 Implementation

One of the supervised learning methods that is frequently employed is the decision tree because of its precision, simplicity, resistance to outliers and missing values, and capacity to map non-linear relationships. We are motivated by these qualities to apply decision tree-based methods to find non-technical electricity losses. Each tree-like network has a root that represents the complete dataset, internal nodes that represent test conditions on attributes, and leaf nodes that represent class labels. Branches indicate the results of the tests conducted on the internal nodes. To optimize the information gained, which distinguishes between the impurity of the parent node and child nodes, the decision tree's training phase selects the best splitter among a range of feasible splitters. The decision tree for power theft behaviors identification was proposed based on improved SMOTE, taking into account the inadequacies of existing electricity theft detection methods and the unbalance of user data. The SMOTE approach can minimize the influence of detection accuracy brought on by unbalanced data.

.

4.6 .2 Evaluation, and Result

Case 1- Daily Power Consumption Data:

In this model, a training set made up of 80% of the user data and a test set of 20% were create d for the daily power consumption using SMOTE. The accuracy that was obtained was 57.78 %

```
Model AUC score: 0.5978066104770902
Model accuray: 57.890182504719945
              precision    recall  f1-score   support

           0       0.94      0.57      0.71     11586
           1       0.12      0.62      0.21      1126

    accuracy                           0.58     12712
   macro avg       0.53      0.60      0.46     12712
weighted avg       0.87      0.58      0.67     12712


True negative: 6660
False negative: 427
True positive: 699
False positive: 4926
Model misclassficiaiton rate: 42.10981749528005
```

Figure 14: Classification report for Daily Power Consumption Data (Decision Tree)

Case 2-Monthly Power Consumption Data:

In the decision tree using SMOTE, a training ratio of 70:30 was employed to compare the out comes between fraud and honest consumers based on the average electricity use over 30 days in a month. The model's monthly power consumption accuracy is therefore 68.29% according to this data.

```
Model AUC score: 0.6260369975523224
Model accuray: 68.29767149150409
              precision    recall  f1-score   support

           0       0.94      0.70      0.80     11586
           1       0.15      0.56      0.24      1126

    accuracy                           0.68     12712
   macro avg       0.55      0.63      0.52     12712
weighted avg       0.87      0.68      0.75     12712


True negative: 8055
False negative: 499
True positive: 627
False positive: 3531
Model misclassficiaiton rate: 31.70232850849591
```

Figure 15: Classification report for Monthly Power Consumption Data (Decision Tree)

5 Comparison of Developed models and Discussion

The dataset for this research is collected from the state grid corporation China. The dataset providers' fraud consumers ladle the dataset with '0' and honest consumers with '1'. In the case study, the monthly dataset shows very low accuracy of the models compared to another case study. In the case study, the daily conditions are taken with two different scenarios with two other cases, 70% training, and 80% training accuracy.

Focus on the daily power consumption dataset for the XGBoost case study in which the model is trained with 80% slightly higher training accuracy than the 70% training accuracy models. The training data has more data, so there are very high chances for misclassification, which indicates that the model is overfitting and unable to predict an accurate fraud consumer. So, for that reason, in this research work, is considered an essential solution to the problem in the Boosting Algorithm Case. In Random Forest, daily power consumption, which is 91.40% more

accurate than monthly power consumption with a 70% training dataset, produces better results. When using K-Mean Clustering, accuracy is approximately 91.47% for both daily and monthly data consumption. In every model that has been tested, CNN with XGBoost performs better than the other tested models.

Table 1 Performance of the various case model

	Power Consumption	Target	Precision	Recall	F1-Score	Accuracy	Miss Classification rate
Boosting Algorithm	Case-1	Fraud	1.00	0.96	0.98	96.00%	4.07
		Honest	0.70	0.97	0.83		
	Case-2	Fraud	1.00	0.97	0.99	96.00%	4.42
		Honest	0.78	0.99	0.87		
	Case-3	Fraud	0.96	0.75	0.84	74.00%	25.8232
		Honest	0.21	0.70	0.32		
	Case-4	Fraud	0.96	0.74	0.84	74.00%	26.241
		Honest	0.20	0.69	0.31		
Random Forest	Case-1	Fraud	0.92	1.0	0.95	91.40%	2.57
		Honest	0.62	0.08	0.14		
	Case-2	Fraud	0.92	1.00	0.95	91.35%	8.64
		Honest	0.62	0.07			
K-mean Clustering	Case-1	Fraud	0.91	1.00	0.96	91.46%	8.53
		Honest	0.00	0.00	0.00		
	Case-2	Fraud	0.91	1.00	0.96	91.47%	8.52
		Honest	1.00	0.00	0.00		
Decision Tree	Case-1	Fraud	0.94	0.57	0.71	57.89%	42.10
		Honest	0.12	0.62	0.21		
	Case-2	Fraud	0.94	0.70	0.89	68.29%	31.70
		Honest	0.15	0.56	0.24		

6 Conclusion and Future Work

This research work proposed a hybrid approach to identify the fraud consumers from the bunch of the honest consumers based on their power consumption information which is collected from the smart meter. This smart meter energy dataset for this research work was collected from the state grid corporation China. This dataset contains information on the power consumption of around 42,000 consumers. All this information is passed and processed for the data analysis and exploration of the problem in detail. This research proposed a deep and broad CNN model hybridized for the imbalanced classification with the xgboost algorithm with the updated class weight for better information extraction. The multiple cases are examined to select the best tuned and optimized dataset for the final model training and testing. From these cases discussed, daily power consumption is chosen with the 70:30 training and testing model ratio. So far, this model can identify fraud consumers with a misclassification rate of 4.073% and catch 1077 fraud consumers from a total of 1109 consumers in the testing dataset. Compared to the monthly power consumption with a 70% training dataset, the daily power consumption

in Random Forest produces accuracy that is 91.40% higher. In K-Mean Clustering, accuracy ranges from 91.47% for daily data consumption to 91.47% for monthly data consumption. Decision trees' accuracy varies depending on the amount of data is used for modelling from 59.78% for daily data consumption to 62.60% for monthly data consumption. In the end, it was discovered that CNN with XGBoost performed considerably better than all other models tested.

Future Work: This problem is one of the significant issues in the modern world where cryptocurrency is rising, and more people are into crypto mining, most of which is a big issue of using illegal energy to run their system. This research only considered the impact on the power consumption, but still, more features also impact identifying the fraud consumers. Such as their location and type of consumers, which are not maintained in the dataset (industrial, agriculture, domestic, and others), have different power consumption patterns. If all that information is available, fraud consumers have a high chance of better prediction accuracy. However, this project can identify the fraud consumer based on their power consumption pattern. Based on this information, the utility companies can conduct the raid to analyze whether the prediction model works well. But based on this model, identifying the fraud consumer is a huge accusation to the consumers. It can also backfire on utility companies and the government, so fraud detection is required based on the number of iterations to develop this.

Acknowledgment

I would like to thank, Dr. Catherine Mulwa, my research supervisor for her tremendous encouragement and clear direction and guidance that enabled me to successfully complete my research thesis and I would like to express my gratitude to my parents for their continuous encouragement and blessings, additionally my for friends who have always encouraged and supported me during the course of my research.

References

Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. Understanding a convolutional neural network, in: 2017 International Conference on Engineering and Technology (ICET). IEEE, pp. 1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Ashish~Agarwal and Paul~Barham, 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.

Bisong, E., 2019. Google Colaboratory, Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA, pp. 59–64. https://doi.org/10.1007/978-1-4842-4470-8_7

C. Costa, B., Alberto, Bruno.L.A., M. Portela, A., W, M., O.Eler, E., 2013. Fraud Detection in Electric Power Distribution Networks using an Ann-Based Knowledge-Discovery Process. International Journal of Artificial Intelligence & Applications 4, 17–23. https://doi.org/10.5121/ijaia.2013.4602

Chen, T., Guestrin, C., 2016. XGBoost, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., n.d. xgboost: eXtreme Gradient Boosting.

Chines government, 2017. State grid corporation China [WWW Document]. Chines government. URL http://www.sgcc.com.cn/ywlm/index.shtml/ (accessed 1.16.22).

Comden, J., Colombino, M., Bernstein, A., Liu, Z., 2019. Sample Complexity of Power System State Estimation using Matrix Completion, in: 2019 IEEE International

Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). IEEE, pp. 1–7. https://doi.org/10.1109/SmartGridComm.2019.8909815
Douzas, G., Bacao, F., Fonseca, J., Khudinyan, M., 2019. Imbalanced Learning in Land Cover Classification: Improving Minority Classes' Prediction Accuracy Using the Geometric SMOTE Algorithm. Remote Sensing 11, 3040. https://doi.org/10.3390/rs11243040
Fernando, K.R.M., Tsokos, C.P., 2021. Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks. IEEE Transactions on Neural Networks and Learning Systems 1–12. https://doi.org/10.1109/TNNLS.2020.3047335
Fujita, H., Cimr, D., 2019. Decision support system for arrhythmia prediction using convolutional neural network structure without preprocessing. Applied Intelligence 49, 3383–3391. https://doi.org/10.1007/s10489-019-01461-0
Gaur, V., Gupta, E., 2016. The determinants of electricity theft: An empirical analysis of Indian states. Energy Policy 93, 127–136. https://doi.org/10.1016/j.enpol.2016.02.048
Golden, M., Min, B., 2012. Theft and Loss of Electricity in an Indian State.
Gunturi, S.K., Sarkar, D., 2021. Ensemble machine learning models for the detection of energy theft. Electric Power Systems Research 192, 106904. https://doi.org/10.1016/j.epsr.2020.106904
Haes Alhelou, H., Hamedani-Golshan, M., Njenda, T., Siano, P., 2019. A Survey on Power System Blackout and Cascading Events: Research Motivations and Challenges. Energies (Basel) 12, 682. https://doi.org/10.3390/en12040682
Hasan, Md.N., Toma, R.N., Nahid, A.-A., Islam, M.M.M., Kim, J.-M., 2019. Electricity Theft Detection in Smart Grid Systems: A CNN-LSTM Based Approach. Energies (Basel) 12, 3310. https://doi.org/10.3390/en12173310
Hu, Y., Zhang, Y., Huang, T., Hu, Z., Fan, Z., Li, C., 2020. A Detection Method for Electricity Theft Based on Random Forest Algorithm, in: 2020 10th International Conference on Power and Energy Systems (ICPES). IEEE, pp. 553–557. https://doi.org/10.1109/ICPES51309.2020.9349646
Huang, S.-C., Lo, Y.-L., Lu, C.-N., 2013. Non-Technical Loss Detection Using State Estimation and Analysis of Variance. IEEE Transactions on Power Systems 28, 2959–2966. https://doi.org/10.1109/TPWRS.2012.2224891
Jindal, A., Dua, A., Kaur, K., Singh, M., Kumar, N., Mishra, S., 2016. Decision Tree and SVM-Based Data Analytics for Theft Detection in Smart Grid. IEEE Transactions on Industrial Informatics 12, 1005–1016. https://doi.org/10.1109/TII.2016.2543145
Jokar, P., Arianpoo, N., Leung, V.C.M., 2016. Electricity Theft Detection in AMI Using Customers' Consumption Patterns. IEEE Transactions on Smart Grid 7, 216–226. https://doi.org/10.1109/TSG.2015.2425222
Khan, Z.A., Adil, M., Javaid, N., Saqib, M.N., Shafiq, M., Choi, J.-G., 2020a. Electricity Theft Detection Using Supervised Learning Techniques on Smart Meter Data. Sustainability 12, 8023. https://doi.org/10.3390/su12198023
Khan, Z.A., Adil, M., Javaid, N., Saqib, M.N., Shafiq, M., Choi, J.-G., 2020b. Electricity Theft Detection Using Supervised Learning Techniques on Smart Meter Data. Sustainability 12, 8023. https://doi.org/10.3390/su12198023
Kocaman, B., Tümen, V., 2020. Detection of electricity theft using data processing and LSTM method in distribution systems. Sādhanā 45, 286. https://doi.org/10.1007/s12046-020-01512-0
Lewis, F.B., 2015. Costly 'Throw-Ups': Electricity Theft and Power Disruptions. The Electricity Journal 28, 118–135. https://doi.org/10.1016/j.tej.2015.07.009

Mufassirin, M.M.M., Hanees, A.L., Shafana, M.S., n.d. ENERGY THEFT DETECTION AND CONTROLLING SYSTEM MODEL USING WIRELESS COMMUNICATION MEDIA.

Nagi, J., Yap, K.S., Tiong, S.K., Ahmed, S.K., Mohamad, M., 2010. Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines. IEEE Transactions on Power Delivery 25, 1162–1171. https://doi.org/10.1109/TPWRD.2009.2030890

Pan, E., Liu, S., Liu, J., Qi, Q., Guo, Z., 2020. The state grid corporation of China's practice and outlook for promoting new energy development. Energy Conversion and Economics 1, 71–80. https://doi.org/10.1049/enc2.12007

Pedregosa, F. and V.G. and G.A. and M.V. and T.B. and G.O. and B.M. and P.P. and W.R. and D.V. and V.J. and P.A. and C.D., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Pereira, J., Saraiva, F., 2020. A Comparative Analysis of Unbalanced Data Handling Techniques for Machine Learning Algorithms to Electricity Theft Detection, in: 2020 IEEE Congress on Evolutionary Computation (CEC). IEEE, pp. 1–8. https://doi.org/10.1109/CEC48606.2020.9185822

Ponce, P., Molina, A., Mata, O., Ibarra, L., MacCleery, B., 2017. Power System Fundamentals. CRC Press. https://doi.org/10.1201/9781315148991

Sudhir Kumar Katiyar, 2005. Political Economy of Electricity Theft in Rural Areas: A Case Study from Rajasthan. JSTOR 40, 644–648.

van Rossum, G. and D.F.L., 2009. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

Wong, J.C.Y., Blankenship, B., Urpelainen, J., Ganesan, K., Bharadwaj, K., Balani, K., 2021. Perceptions and acceptability of electricity theft: Towards better public service provision. World Development 140, 105301. https://doi.org/10.1016/j.worlddev.2020.105301

Xia, X., Xiao, Y., Liang, W., Cui, J., 2022. Detection Methods in Smart Meters for Electricity Thefts: A Survey. Proceedings of the IEEE 110, 273–319. https://doi.org/10.1109/JPROC.2021.3139754

Yang, H., Qiu, R.C., Chu, L., Mi, T., Shi, X., Liu, C.M., 2020. Improving Power System State Estimation Based on Matrix-Level Cleaning. IEEE Transactions on Power Systems 35, 3529–3540. https://doi.org/10.1109/TPWRS.2020.2984926

Yi-chong, X., 2005. Models, templates and currents: the World Bank and electricity reform. Review of International Political Economy 12, 647–673. https://doi.org/10.1080/09692290500240370

Zhang, H., Han, K., 2020. A Hybrid Observability Analysis Method for Power System State Estimation. IEEE Access 8, 73388–73397. https://doi.org/10.1109/ACCESS.2020.2987358

Zhang, Y., Wang, J., Li, Z., 2020. Interval State Estimation With Uncertainty of Distributed Generation and Line Parameters in Unbalanced Distribution Systems. IEEE Transactions on Power Systems 35, 762–772. https://doi.org/10.1109/TPWRS.2019.2926445

Zheng, Z., Yang, Y., Niu, X., Dai, H.-N., Zhou, Y., 2018. Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids. IEEE Transactions on Industrial Informatics 14, 1606–1615. https://doi.org/10.1109/TII.2017.2785963