ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Feature Extraction of Risk Group and Electricity Theft by using Electrical Profiles and Physical Data for Classification in the Power Utilities

Supakan Janthong[1], Rakkrit Duangsoithong[2] and Kusumal Chalermyanont[3]

ABSTRACT

Non-technical loss (NTL) is one of the problems that has been a major issue in lost revenue for many years. Electricity distributors have attempted to reduce NTL by detecting electricity theft using various methods. Some events are difficult to detect that conventional meters inspection is inadequate. Moreover, many anomaly patterns found are very complex, confusing in identifying or distinguishing what types of electricity customers are at abnormal risk or energy theft that affects NTL. This paper proposes five key feature extraction methods and six classifying electricity customers using supervised learning. The main problem was studied and collected information, including kilowatt meters, electronic meters, TOU meters, and AMR meters, which cover four customer types that were recorded in the Provincial Electricity Authority (PEA) of Thailand. An electrical profile to be extracted for in-depth analysis of the behavior of each type of electricity customer, combined with the information of physical data to help enhance and increase efficiency. All features examined the relationships in each feature using Pearson correlation and handled unbalanced data using random oversampling (ROS). Then, the extracted data has been trained, validated, and tested to classify three classes: normal, risk, and theft, where we evaluate the results with performance metrics. The results show that random forest (RF) outperforms the rest of the classifiers by achieving a precision-recall area under the curve of 90% and a receiver operating characteristic curve of 78%. Significantly, the results were compared to previous studies and benchmark datasets, which revealed that the proposed method gave better results than other techniques.

Article information:

Keywords: Non-technical Losses, Feature Extraction, Supervised Learning, Electricity Customers, Risk Groups, Electricity Theft

Article history:
Received: May 20, 2023
Revised: August 24, 2023
Accepted: September 28, 2023
Published: January 20, 2024
(Online)

DOI: 10.37936/ecti-cit.2024181.252738

1. INTRODUCTION

Major power losses in the electricity network of the Provincial Electricity Authority (PEA) occur at the transmission and distribution levels. Losses can be categorized into two types: technical loss (TL) and non-technical loss (NTL). TL occurs inherently when dissipating electrical energy and operating the equipment [1], and NTL, sometimes considered commercial losses, are non-natural losses associated with the amount of unbilled energy consumed in the billing process [2]. According to the PEA's loss report [3], NTL tends to increase when compared to the past five-years statistics. Most common causes of NTL in PEA include energy theft, meter measurement inaccuracies, errors in meter reading, billing problems, and behavior for small loads [4]. Each cause results in a different amount of NTL. Considering the meter used, there are various types of meters installed in PEA such as kilowatt-hour meters, electronic meters, TOU meters, AMR meters and AMI meters. Each type has many sizes, including 5(15)A, 15(45)A, 30(100)A, 5(100)A, and 5(6)A, which were installed for customers to use differently [5]. Notably, each type of meters has limitations on data access and data

[1] The author is the 1Department of Electrical Engineering, Prince of Songkla University Hatyai, Songkhla, Thailand, Meter Department, Provincial Electricity Authority, Thailand, E-mail: 6410120056@psu.ac.th

[2,3] The authors are with the 3Department of Electrical Engineering, Prince of Songkla University Hatyai, Songkhla, Thailand, E-mail: rakkrit.d@psu.ac.th and kusumal.c@psu.ac.th

[2] Corresponding author: rakkrit.d@psu.ac.th

collection capabilities. Examples of meters installed in PEA are shown in Figure 1. In the past few years, PEA has used a technique to inspect NTL by searching the information from the system according to the street roots and having an electrician check the actual on-field site. It takes a lot of time to inspect a large number of electricity customers, especially risk group customers have the opportunity to be both normal and theft which takes even more time to inspect each individual. In some cases, there were thefts from time to time (act at night) causing no abnormalities to be detected while checking.

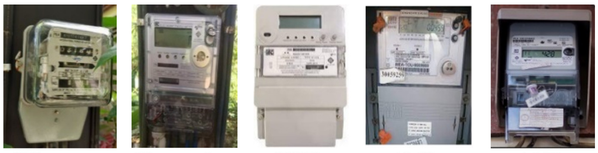

Kilowatt-hour meters
Electronic meters
TOU meters
AMR meters
AMI meters

Fig.1: *Types of meters installed in PEA.*

Given the above concerns, this paper proposes a methodology that uses two main information meters from historical data of PEA: 1) electrical profiles that are signals or time series that can be measured with different resolutions such as voltage, current, and energy; 2) physical data that describes aspects of the customer's consumption behavior, geographical location, and area conditions that all affect the meters. These characteristics are analyzed and extracted into feature-based and auxiliary features. After that, all features are classified using several supervised learning algorithms, and the results are evaluated using performance metrics. The main contributions of this work can be summarized as follows:

1. A large amount of available information in the database of PEA was analyzed for maximum benefit, not only electrical profiles but also physical data were used as supplementary data to increase the efficiency of modeling.
2. In order to ensure accuracy and reliable extraction of each extracted feature. Various knowledge has been synthesized to create features such as expert knowledge, work experience, time series analysis, and international standards.
3. Classifying between normal and abnormal may be insufficient, or the problem may not be resolved in time. In order to be able to detect and fix problems quickly, knowingly, predict before the theft happens, we then add more risk groups of customers to our classes.
4. Several supervised learning models were parameterized for optimization and evaluated to find the most efficient model.

The rest of this paper is organized as follows: In Section 2, we present the literature review. In Section 3, we describe the proposed framework, which consists of theory, methods, and experimental results. In Section 4, we provide results and discussion. In Section 5, we summarize the entire paper and provide perspectives for the future.

2. LITERATURE REVIEW

NTL detection studies show that there is no fixed method to detect electricity theft. Currently, methods for NTL detection found in the literature can be categorized as either theoretical studies (factor analysis, effect, and cause), hardware solutions (measuring equipment, sensors), or non-hardware solutions (detection, expert systems, fuzzy systems, game theory) [6]. Our approach proposes a non-hardware solution based on classification. A solution for fraud detection based on artificial neural networks (ANN) has been presented in [7]. The data they used for teaching the model consists of two types: nominal and numeric, total 14 key attributes. Some features were of low significance and did not reflect abnormal patterns. Glauner et al. used Boolean, fuzzy logic, and SVM to detect NTL. The input features for the algorithms consist only of the last 12 months for energy consumption (EC). Non-numeric data types are not taken into account [8]. The distribution of features generated distance between the neighborhoods of customers for each area, location, and different sizes of grids in [9]. However, the above model is just grouping data not distinguishing. The traditional methods used support vector machines (SVM) to understand the significant indicators of fraudulent behaviors, this makes it possible to analyze the development of selection policies to address the problems [10]. Guerrero et al. [11] propose a method to increase the precision of NTL campaigns based on null consumption analysis. Data mining and Neural network (NN) are used for customer filtering, while a second module creates rules devised from decision tree (DT) and self-organizing maps (SOM-NN). M. M. Buzau et al. show evaluating the performance of the models using various metrics such as the true positive rate (TPR), recall (RCL), false positive rate (FPR), precision (PRC), and AUC score [12], where the properties are seen from multiple perspectives for the results. In [14], they use a neural network (NN) and in [15] they use a decision tree (DT) for forecasting customer energy consumption using the variance of the predicted EC compared to the actual EC. If the specified EC is exceeded, it is considered theft. While [6] detects conditional anomalies from the data, the feature is created from extraction based on the customer, which does not mention the related effects of NTL issues.

3. PROPOSED FRAMEWORK

The main aim of the present paper is to classify the behavior of the customers among normal conditions, risk groups, and energy theft, according to the probability that there will be an anomaly in each type of meters. This method mainly uses raw data from three types of meters for feature generation, such as kilowatt-hour meters, TOU meters, and AMR meters. Figure 2 shows an overview of the proposed framework in these studies. The raw data from the PEA database was collected, cleaned, and normalized. Input data are divided into two categories: electrical profiles and physical data, both data are then extracted into feature-based and auxiliary features to find data correlation. After that, all features handle data imbalance, and split data to train, validate, and test. Finally, multiple supervised learning is classified, which evaluates the result using performance metrics.

3.1 Dataset

The information used is specific to Southern Area 3 PEA, which EC obtains from automatic readings and from checker reading units every month. We separate data into two groups: electrical profiles and physical data in Table 1. The data sources include SAP, U-CUBE, GIS, OPSA, and SMR that cover the last twelve years, from January 1, 2010, until December 31, 2022.

Table 1: *Data type used in the experiment.*

Electrical profiles	Physical data
Voltage for 1P or 3P (V), Current for 1P or 3P (A), Energy (kWh), Active power (kW), Reactive power (kVAR), Angle or power factor (pf), Total harmonic distortion of current (THD_A), and voltage (THD_V)	Seasonality, Alarm events, Geographical area, Economic activity, Number of complaints, Contracted power, Business type, Meter type, Amp rating

Approximately 1,000 customers for on-field inspections are performed with all types of customers, including residential, small, medium, and large businesses as shown in Table 2. However, during that time, there was a lot of data. We need to randomly select some intervals, which still cover every time and customer types. The technique we use is based on statistical random selection, consisting of first randomly selecting customers from all categories covering all types of customers; second filtering select groups with abnormalities in each type; and third, then randomly select an abnormal period by specifying 90 days per period. The label of physical data has recorded the results of the actual inspections on-site in 3 classes: normal condition (label: 0), risk group (label: 1), and energy theft (label: 2). More details of the label definition are shown in Table 3. For electrical profile data, we define the scope of the situation as follows: first sample, no abnormalities found and normally more than 90 days, second sample, from initial inspection to irregularity found, and third sample, from initial or irregularity until anomaly or theft is found. A framework of events is shown in Figure 3.

3.2 Data cleaning

Sometimes the installed meters is damaged while being used, causing a new meter to be replaced. In this case, many sizes of meters and various types of customers cause data to have missing values. We used equation (1) to correct errors before using the data. Moreover, since the information is available for both 1-phase and 3-phase systems, it is necessary to adjust the value to reduce the outliers by using equation (2). After that, the data were normalized using max-min scaling with equation (3). For all three equations used, we got the concept of [13] and have applied it to our work. This step is quite crucial because if the data entered into the model is incorrect, it may cause the results to be unreliable.

$$F_{(E_i)} = \begin{cases} \dfrac{E_{i-1}+E_{i+1}}{2} & , E_i \in NaN \\ E_i & , E_i \notin NaN \end{cases} \tag{1}$$

Where; E_i is a variable in an array of customer i

$$T_{(E_i)} = \begin{cases} \bar{E}_i + 2\sigma(E_i) & , if\ E_i\ >\ \bar{E}_i + 2\sigma(E_i) \\ E_i & , otherwise \end{cases} \tag{2}$$

Where; σ is standard deviation and $\bar{E}_i$ is average value of E_i respectively.

$$N_{(E_i)} = \frac{E_i - \min(E_i)}{\max(E_i) - \min(E_i)} \tag{3}$$

Where; $N_{(E_i)}$ is array normalized in w window and $\min(E_i)$ is minimum and $\max(E_i)$ maximum of variable in an array for the customer i respectively.

3.3 Feature extraction

From two data types in Table 1, we extracted two features: features based on electrical profiles and auxiliary features based on physical data. The crucial elements for data extraction are range or window definition and the techniques used for characteristic extraction. Table 4 provides each technique and window of extraction for feature-based, and Table 5 describes in detail the extracted auxiliary features. The extraction process for each method will be described in a subsection.

3.3.1 Quality byte

Features created using the quality byte (QB) measurement aim to detect all types of alarms and all physical attacks [12]. In the case of a traditional meters, the checker would go out to read the unit once a month and if there were any anomalies found on site, the case would be triggered and sent into UCUBE, which would alarm and alert the technician for inspection. In the case of an AMR meters, the system will automatically alarm and save the data in the

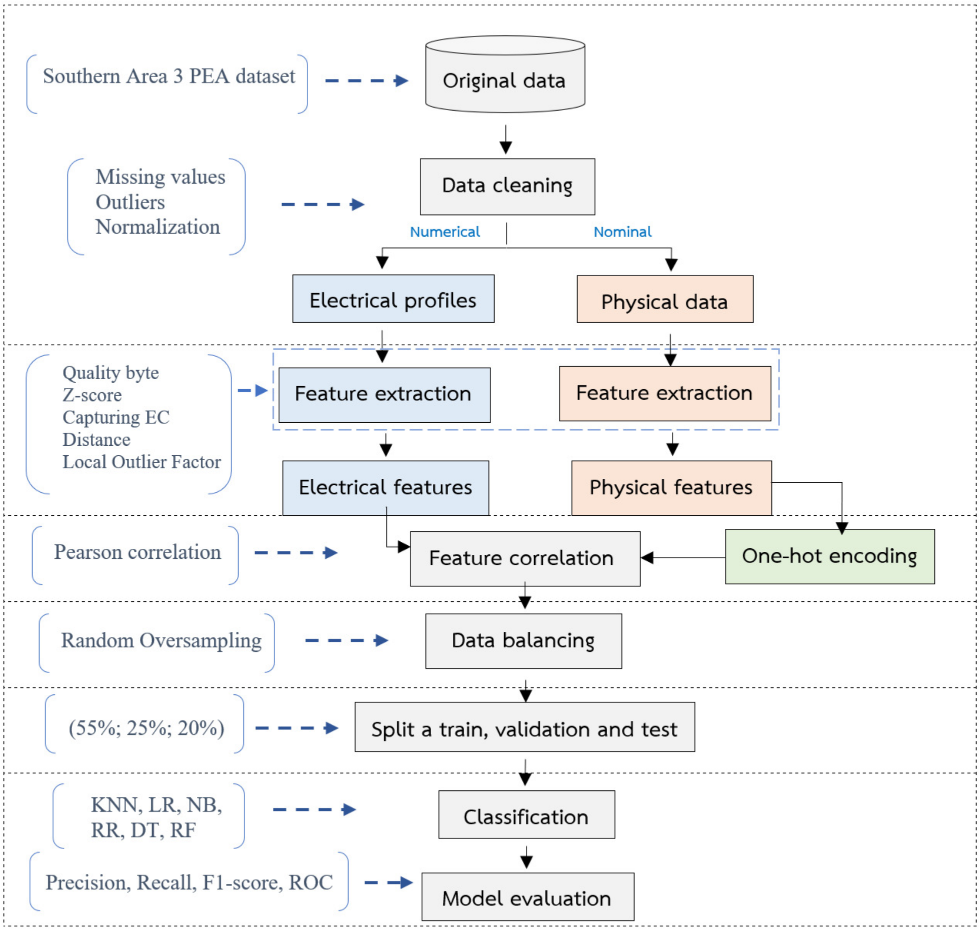


Fig.2: *Proposed framework.*

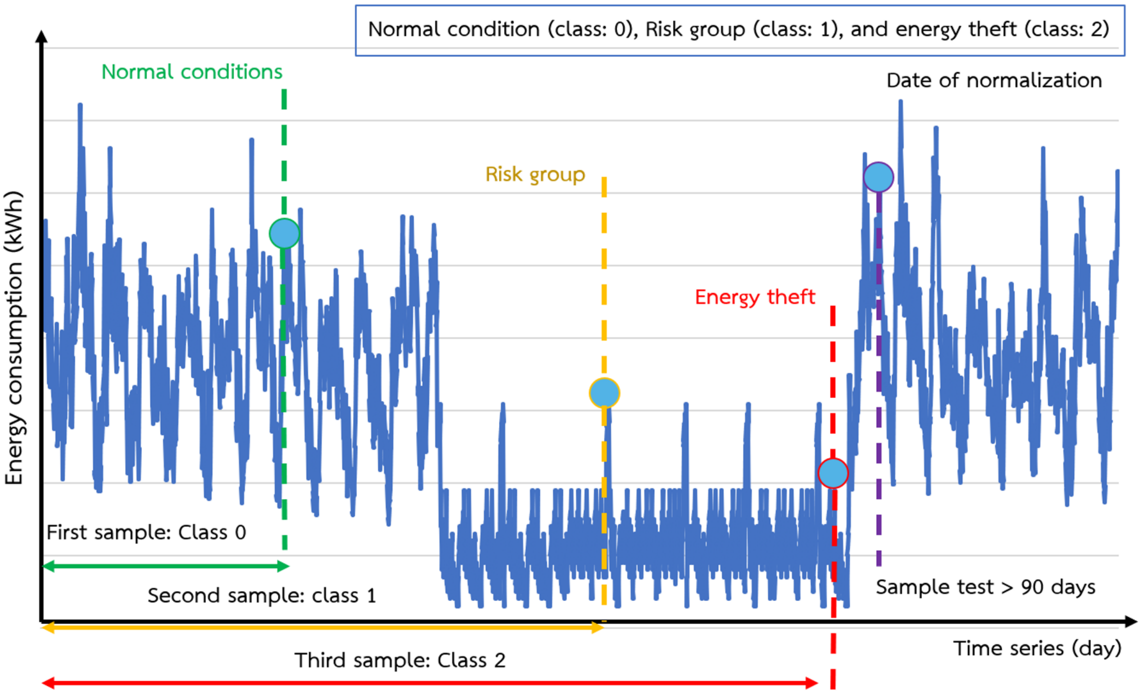


Fig.3: *EC scenarios for customer criteria.*

Table 2: *Detailed datasets.*

Period	01/01/2010 to 31/12/2022	
Classes	Number of data (customers)	Customers type
Normal conditions	600	Residential areas, Small businesses, Medium business Large business
Risk group	300	
Energy theft	100	
Total customers	1,000	

Note: All three groups of information cover and distribute four types of customers.

Table 3: *Details of class definitions.*

Period	Definitions
Normal conditions	Groups that have regularly used EC statistics and increase and decrease by no more than the specified percentage range in each condition.
Risk group	Groups were selected based on the behavioral consumption of customers that tended to be similar to anomalies. In some periods there will be sudden changes and a downward trend. The outlook for conditions is as follows: ■ Large business, 25% increase or decrease. ■ Medium and small businesses, 20–50% increase or decrease. ■ Residential areas, reduction units less than 1–99 units. ■ Null EC more than 3 consecutive months.
Energy theft	Groups that detected the energy theft or anomaly that completed the correction.

Table 4: *Data extracted as electrical profiles features.*

Feature	Technique/Type data	Window, Time, Days
Alarm	Quality byte	Month of 365 days
Consumption behaviors	Number of 0 kWh measurements	Month of 365 days
Number of measurements received in the last month	Average Z score taken during time window w	Rate A Rate B Rate C 365 days
Average Z score of daily EC measurements	Standard deviations	Rate A Rate B Rate C 365 days
EC measurements taken on type of day t	Euclidean distance Manhattan distance	Rate A Rate B Rate C 365 days
LOF score of EC daily profile	Local outlier factor	365 days Load profiles
Phase switching, Tampering	Phase imbalance, Unbalance current ratio	Month of 365 days
Current, Voltage Active(kW) /Reactive (kVAR), Phaser, Angle	Max, Min, Average, RMS, Standard deviations during time window w	Rate A Rate B Rate C 365 days

Table 5: *Data extracted as physical features.*

Feature
Location of province
Location of district
Manufacturing company
Business type
Meter type
Phase line voltage
Amp rating
Ratio
Latitude, Longitude
Number of complaints
Year install

event log. In this study, the QB measurement uses an 8-bit code to assess the quality of the measurement. Eight separate values in terms of binary code have been converted to a decimal number. Table 6 shows an example of a code assigned to correspond to a code checker that cover eight events. Table 7 shows a sample of a case where the meters is found to be damaged and the unit cannot be read, the binary value is 01010000, which can be calculated into a decimal number as 130.

3.3.2 Capturing of consumption behaviors

Consumption behaviors are different for each type of customer depending on the load used. A preliminary analysis of the EC data using statistical techniques found that the normal pattern varied periodically while that of the risk and theft group varied non-periodically. Based on this concept, we adapted it to check consumers' EC patterns. The scenarios of customer criteria are divided into three groups (Figure 3), which cover the main conditions of possible variations. For example, an anomaly event occurred before it was found, during the events that are happening, and after the incident. To detect these anomalies, the Z_{score} was used according to Equations (4)-(6). This score indicates how many standard deviations are away from the mean measurement [12]. Z_{score} divides measurements taken at three rates (rate A, rate B, and rate C) to avoid erroneous results according (show in Table 8). In Figure 4, we show some EC of small businesses; the upper part shows a load profile for three rates, and the rate C of the last n days tends to decrease. The below part EC converted into Z_{score} and found that rate C period less Z_{score} (red dotted line).

Table 6: *New Guidelines for Fattening Beef Cattle.*

Bit	Alarm	Description
7	ALF	All fraud (ALF = 1)
6	DTM	Defective meter (DTM= 1)
5	PTM	Patrol meter (PTM = 1)
4	BOR	Be out of reading (BOR = 1)
3	DSC	Defer staff check (DSC = 1)
2	WRP	Wrong PEA No. (WRP = 1)
1	MSM	Missing meter (MSM = 1)
0	EMH	Empty house (EMH = 1)

Table 7: *Example of events using QB.*

ALF	DTM	PTM	BOR	DSC	WRP	MSM	EMH
0	1	0	1	0	0	0	0

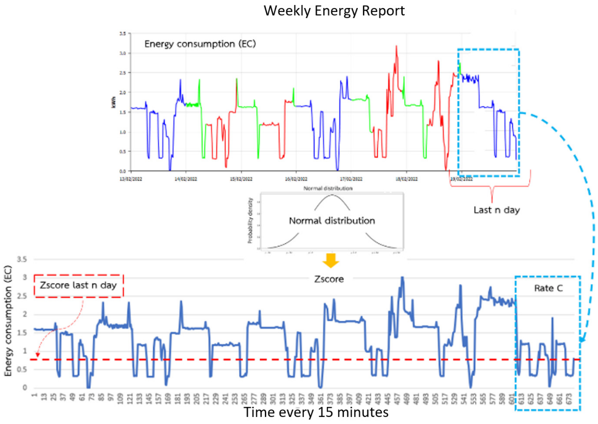


Fig.4: *Example of calculating EC into Z-score.*

$$Z_{score} = \frac{E_i - \bar{E}_i}{\sigma E_i} \tag{4}$$

$$\bar{E}_i = \frac{1}{n}\sum_{i=1}^{N} E_i \tag{5}$$

$$\sigma_{E_i} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(E_i - \bar{E}_i)^2} \tag{6}$$

Where; E_i is EC of customer with n samples, $\bar{E}_i$ is the mean and σ_{E_i} is standard deviation of EC for the customer i respectively.

Table 8: *Time rate of TOU meters.*

Rate	Rate Interval Requirements
Rate A	Peak: 09:00 a.m.- 10:00 p.m.
Rate B	Off-Peak : 10:00 p.m.-09:00 a.m.
Rate C	Holiday: 00:00 a.m.-11:59 p.m.

3.3.3 Distance Measurements

The base consumption patterns have been enumerated by using different intervals for each month. The Euclidean and Manhattan distances created the features [12]. The Manhattan distance $Dist_M$ was computed including Rate A, Rate B, and Rate C as equation (7), whereas the Euclidean distance $Dist_E$ was computed using the total results for three periods as an equation (8). Features extracted from Manhattan and Euclidean distance show in Table 9.

$$Dist_M = \left| E_i - \frac{1}{n}\sum_{i=1}^{W} E_i^r \right| \tag{7}$$

$$Dist_E = \sqrt{\sum_{i=1}^{n}\left(E_i - \frac{1}{n}\sum_{i=1}^{N} E_i^r\right)^2} \tag{8}$$

Where: W is window selected from each rate r

Table 9: *Feature of distance measurements.*

Distance Measurements	Manhattan distance of time window W Rate A
	Manhattan distance of time window W Rate B
	Manhattan distance of time window W Rate C
	Total Manhattan distance for window W
	Total Euclidean distance for window W

3.3.4 Local Outlier Factor (LOF)

An unusual customer behavior for not using electricity was analyzed at each window for a specific grouping. Local Outlier Factor (LOF) is a method used to determine values and compare them to the neighbourhoods of each type of electricity consumer. The clustering distribution of a factory for each rate of which the ungrouped value is to be assigned as a feature, the example shows in Figure 5.

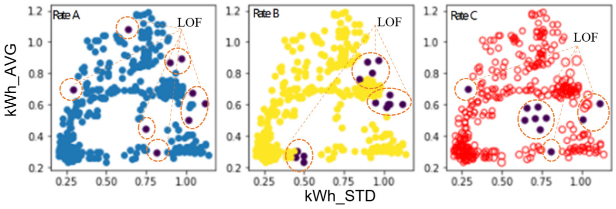


Fig.5: *Clustering of rates (A, B, C) for finding LOF.*

3.3.5 Calculation of interval statistics

The behavior of using electricity can change due to many factors, such as increasing-decreasing load, changing the season, environment, or time use period. This caused fluctuations in daily power patterns, which may be seen as abnormal when actually normal consumption. To detect this problem, basic statistics are used in the calculations consisting of max, min, median, mean, and RMS. An example of specifying a time period to calculate a feature is shown in Figure 6.

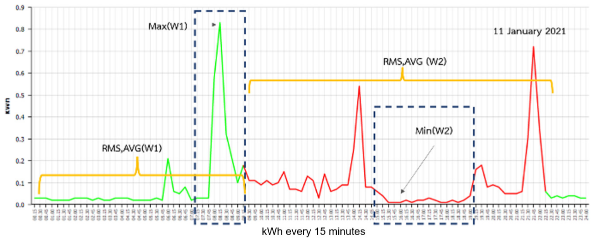
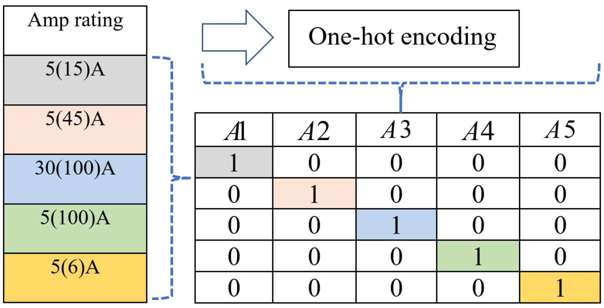


Fig.6: *Detection and calculation of statistics.*

3.3.6 One-hot encoding

One-hot encoding is a technique used in data analysis to transform nominal categorical variables into features with numerical values. Each category is converted into a binary vector, where the index corresponding to the category is set to 1, and all other indices are set to 0. The vector has a length equal to the total number of categories. In this binary vector. Figure 7 represents categorical variables as numerical values for Amp rating of meters.



Amp rating
5(15)A
5(45)A
30(100)A
5(100)A
5(6)A

A1	A2	A3	A4	A5
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

Fig.7: *Amp rating(nominal) feature into numerical values.*

3.4 Feature correlation

The dataset was extracted using the technique in Section 3.3, with a total of 32 features comprising 20 numerical data for electrical profile features (red font and red bar graph) and 12 non-numerical data for physical features (blue font and blue bar graph) in Table 10 and Figure 8. To find out which variables correlate with the classes, Pearson correlation is used as Equation 9. The correlation results show that the data were both positively and negatively correlated, with positive having 15 features and negative having 17 features. For positive correlation, there were 10 electrical features and 5 physical features, while for negative correlation, there were 11 electrical features and 6 physical features.

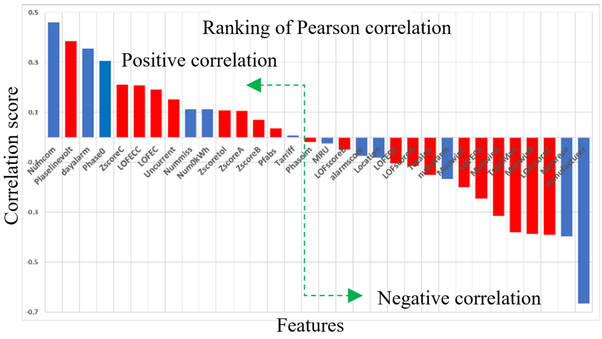


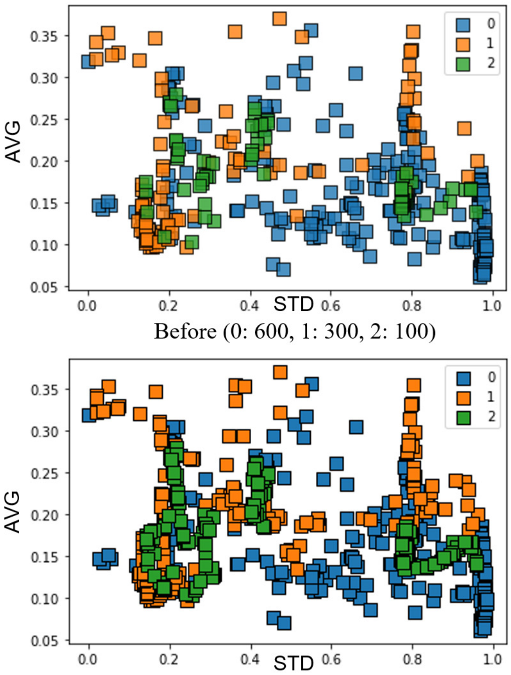
Fig.8: *The correlation of 32 features.*

Table 10: *Attribute ranking of 32 feature.*

Ranking	1	2	3	4
Feature	Numcom	Phaselivolt	Dayalarm	Phase0
Correlation	0.4597	0.3863	0.3550	0.3056
Ranking	5	6	7	8
Feature	ZscoreC	LOFECC	LOFEC	Uncurrent
Correlation	0.2097	0.2095	0.1909	0.1522
Ranking	9	10	11	12
Feature	Nummiss	Num0kWh	Zscoretol	ZscoreA
Correlation	0.1128	0.1128	0.1073	0.1057
Ranking	13	14	15	16
Feature	ZscoreB	Pfabs	Tariff	Phaseim
Correlation	0.0709	0.0353	0.0064	-0.0176
Ranking	17	18	19	20
Feature	MRU	LOFscoB	Alarmsco	Location
Correlation	-0.0238	-0.0484	-0.0729	-0.0801
Ranking	21	22	23	24
Feature	LOFECB	LOFscoA	Totaleu	Numalarm
Correlation	-0.1035	-0.1153	-0.1497	-0.1653
Ranking	25	26	27	28
Feature	ManwinC	LOFECA	ManwinA	TotalMan
Correlation	-0.1980	-0.2441	-0.3140	-0.3806
Ranking	29	30	31	32
Feature	ManwinB	LOFscoC	Numrece	Manufact
Correlation	-0.3871	-0.3901	-0.3961	-0.6644

3.5 Data balancing

Unbalanced data is one of the factors that affects model learning because it can cause the model to be biased and predict closer to the majority. Considering the amount of data used, which include class 0: 600 samples, class 1: 300 samples, and class 2: 100 samples (ratio; 60: 30: 10), shows that they are unbalanced data. Therefore, random oversampling (ROS) was used to balance the data. The idea for adjustment is resampling for generating data, which involves adding minority data to a number similar to that of the majority by using the nearest neighbor. The regrouping results are shown in Figure 9.



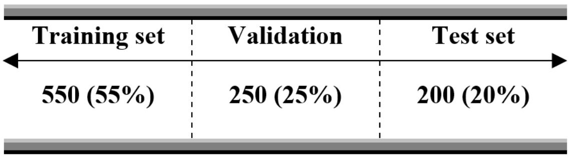
Before (0: 600, 1: 300, 2: 100)

After (0: 600, 1: 600, 2: 600)

Fig.9: *Balance the data using ROS.*

3.6 Split a train, validation and test

After data pre-processing and balancing, the dataset was split for training (55%), validation (25%), and testing (20%). Three data sets were divided as shown in Figure 10.



Training set	Validation	Test set
550 (55%)	250 (25%)	200 (20%)

Fig.10: *Split the data into three sets.*

3.7 Model configuration and classification

Supervised learning is used for pattern classification in three abnormal groups, which include decision trees (DT), k-nearest neighbors (KNN), naive bayes (NB), logistic regression (LR), random forest (RF), and ridge regression (RR). The models selected for classification have different structures and functions, where we need to find the model that best fits our dataset and achieves the best performance. Selecting the suitable classification algorithm and corresponding tools involves considering various factors, such as the nature of data, the complexity of the problem, the interpretability of the model, and computational efficiency. An overview of the mentioned algorithms and some ideas for selecting tools:

Decision Trees (DT):

Decision trees create a tree-like structure of decisions to classify data points by recursively splitting based on features. Suitable for both small and large datasets. It is easy to understand and visualize, but it can be prone to overfitting.

K-Nearest Neighbors (KNN):

KNN is a simple and intuitive classification algorithm that classifies data points based on their proximity to other data points. Suitable for small to medium-sized datasets with non-linear decision boundaries. However, it is sensitive to feature scaling and requires careful tuning of the k parameter.

Naive Bayes (NB):

NB is a probabilistic classification algorithm based on Bayes' theorem, assuming that features are independent given the class label. Often used for text classification and spam filtering. It is simple, computationally efficient, and can handle high-dimensional data.

Logistic Regression (LR):

LR is a widely used linear classification algorithm that models the probability of a data point belonging to a particular class. It works well for binary and multi-class classification tasks. However, it assumes a linear relationship between features and log-odds and can handle large datasets.

Random Forest (RF):

Random forest is an ensemble method that constructs multiple decision trees and aggregates their predictions. It works well for a wide range of classification problems. It reduces overfitting compared to individual decision trees, offers feature-crucial insights, and can handle high-dimensional data.

Ridge Regression (RR):

Ridge regression is a regularization technique applied to linear regression to prevent overfitting by adding a penalty term to the loss function. It is useful when dealing with L2 regularization in features and preventing overfitting. It helps stabilize model coefficients but might not work well if the relationship between features and target is highly non-linear.

Crucial things when choosing algorithms and tools we also need to consider other factors such as dataset size, features, interpretability requirements, computational efficiency, and potential overfitting issues.

The hyperparameter is used to optimize parameter values as shown in Table 11. To execute the model, we use a machine with an Intel Core i7-12700H CPU at 2.30 GHz 16 GB of RAM and compile on a Python Jupyter notebook using the Scikit-Learn library.

Table 11: *Hyperparameter and configuration.*

Model	Hyperparameter	Optimizing values
DT	criterion	gini, entropy, log_loss
	max_features	auto, sqrt, log2
KNN	K	3,5,7
	p	1,2,3
NB	P	default
	V	1e-09
LR	C	0.001, 0.01, 0.1
	R	L1 norm, L2 norm
RF	n_estimators	10, 50, 100
	criterion	gini, entropy, log_loss
RR	class_weight	dict, balanced
	solver	auto, svd, cholesky

3.8 Model evaluation

Model evaluation is a key part to show how well the model can be classified. To compare outcomes, the feature extracted is entered to each model that has a hyperparameter in Table 10. In this study, performance matrices are used to evaluate the model as follows: Receiver operating characteristic (ROC), Precision, Recall, and F1-score as shown in equation (9)-(11).

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{11}$$

Where: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)

4. RESULTS AND DISCUSSION

4.1 Experimental results

Figure 11 presents the classification results of the extracted data using six classifiers, where we are dividing the evaluation into three parts. First, the left image shows a color-coded heatmap with a gradient from light yellow to dark red. That is reported in conjunction with numerical scores including the precision, recall, F1, and support scores. Results show that decision trees (DT) had the highest precision (80.3%), ridge regression (RR) had the highest recall (96.6%), decision trees (DT) were the highest (79.7%) for F1-score, and support score was the number of actual occurrences of classes in the specified data set. Second, the receiver operating characteristic (ROC) is a measure of the predictive quality of a classifier that shows the relationship between the true positive rate on the Y axis and the false positive rate on the X axis. In good graphs, true positives are one and false positives are zero. From the experimental results, ridge regression (RR) shows a curve that is close to the true positive rate and has the highest mean of all 3 groups, indicating that prediction results are more correct than wrong predictions. Third, precision-recall curves show the trade-off between precision with recall and calculate the average precision for the precision-recall curve. From the experimental results, we found that the random forest (RF) showed high precision for each class and gave the highest average precision (90%). However, if considering the scope of work used, the desired outcome is a high accuracy prediction and a low FPR.

4.2 Feature type comparison

The dataset used in this study consisted of numerical and non-numerical data that covered both feature-based and auxiliary features. In order to demonstrate the significance of features that affect feature performance, we have compared the results into five cases with RF classifier, as shown in Table 12. The results show that when we separate between electrical and physical features, electrical features give higher results. While considering positive correlation and negative correlation separately, positive correlation is better than negative correlation and the combination of electrical and physical features. Nevertheless, we cannot only take positive correlation into account in feature extraction. Because the features extracted do not cover all customer conditions and meter types. Therefore, each feature will play a crucial role in enhancing the learning ability of the model to be able to classify better.

4.3 Comparison of previous work

Table 13 presents a comparison of evaluations using different types of criteria, divided into three levels. Notice that [14]-[18] use only the EC value from the smart meters and does not provide specific details for the type of customer. There are only two groups of data used for detection: normal and theft, making it unclear if they need to be used to detect other forms of anomaly. Furthermore, there is no separation of data into risk groups for analysis, making it difficult to predict abnormal events in advance. In this case, it detects when an abnormality has already occurred. However, in this study, there is a disadvantage in physical data that is personal data, so we must be careful about data privacy.

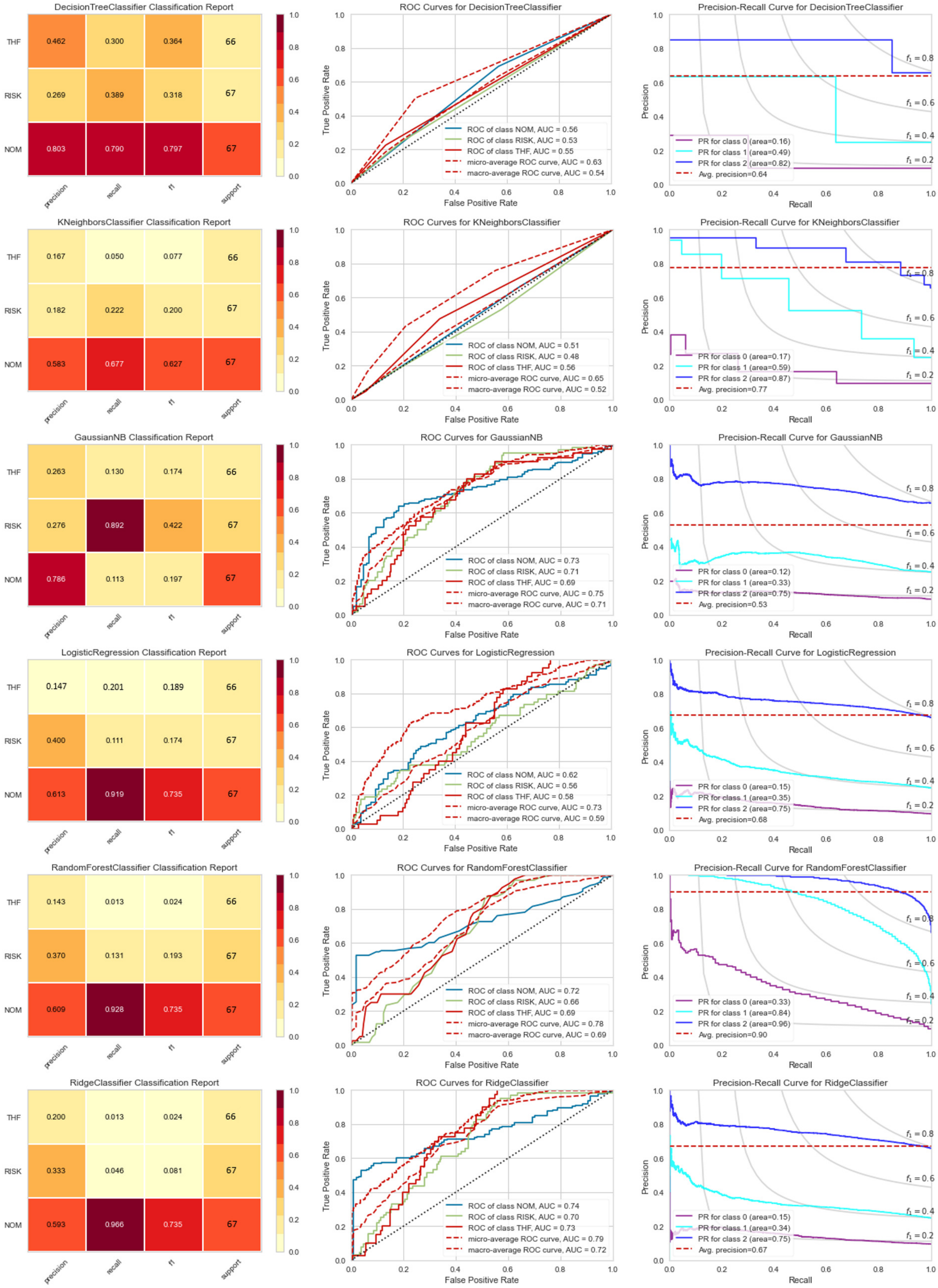


Fig.11: *Comparison of other studies.*

Table 12: *Comparative analysis by type of feature.*

Features	Precision	Recall	F1 score
Electrical features (20 features)	0.7845	0.7549	0.7699
Physical features (12 features)	0.6120	0.6647	0.6505
Only positive correlation (15 features)	0.85140	0.8317	0.8484
Only negative correlation (17 features)	0.4624	0.4736	0.4801
Both electrical and physical features (32 features)	0.8022	0.8354	0.8154

Table 13: *Comparison of other studies.*

Criterion	[14]	[15]	[16]	[17]	[18]	Our
Covering many types of meters	L	L	L	M	L	H
Covering many types of customers	M	L	L	M	L	H
Applying both 1-phase and 3-phase systems.	L	L	L	M	M	H
Learn from multi classes	L	L	L	L	L	H
Adjusting to other abnormal patterns	M	M	M	H	H	H
Anomalies found prior the incident	M	L	L	M	L	H
Method complexity	L	L	L	L	L	L
Data privacy	H	H	H	H	H	M

Note: Criteria level; H: High, M: Medium, and L: Low

4.4 Comparison analysis with benchmark datasets

In order to analyze and compare the proposed method with other techniques from other papers. We have chosen to use the datasets in this study as a benchmark for comparative evaluation. The data used will be balanced (0:600, 1:600, 2:600), extracted, and classified using different techniques. The data will be split into training, validation, and testing (55%; 25%; 20%), where the parameters are defined accordingly in each paper. In [14], [15], [16], [18], [19], [20], [21], and [22], there are several techniques that have been used for data extraction, including statistics, expert knowledge, PCA, EMD, and various types of supervised learning. Significantly, focus the features used in each paper between electrical profiles and physical data, most of which use only energy consumption. The results are shown in Table 14. It is found that techniques from other papers give an overall efficiency of about 70–80%; however, when compared to the study we offer, our study also gives better results. Because features are extracted from many perspectives, whether electrical profiles components, time series, or reference international standards, that helps improve classification performance.

4.5 Limitations

The limitations in this study are with respect to the information used, which can be summarized into two issues as follows:

1. We use data from various meters, where each type of meter has a different resolution of data. For example, kilowatt-hour meters read the unit once a month, electronic meters read every 30 minutes, and AMR meters read every 15 minutes. This factor affects the selection of the data intervals to be extracted.
2. In this paper, data from the Southern Area 3 region of PEA is analyzed. If the model is used in other areas, economic factors and spatial significance need to be thoroughly evaluated.

5. CONCLUSION

The present work applies the handcraft of feature extraction to five methods for data extraction and six supervised learning methods for data classification, where results are evaluated using three metrics. In experimental results, feature extraction is the most crucial factor affecting model learning, whether it is a nominal or numerical feature. However, feature correlation shows that some data are inversely correlated as a result of various information. Subsequently, data balancing contributed to the model's ability to learn better and reduce bias. Finally, classifying the features to compare each model and using performance matrices to evaluate the results. The results showed that each model revealed different strengths and weaknesses, but overall, random forest (RF) is the most suitable for implementation. This model can be applied to PEA applications to improve inspection efficiency and significantly, to help monitor and predict anomalies before the incident.

In future work, we intend to explore the integration of extracted features with deep learning and how to apply a combination of supervised and unsupervised learning models for classification fields.

ACKNOWLEDGEMENT

The authors would like to thank the Provincial Electricity Authority and Department of Electrical Engineering, Faculty of Engineering, Prince of Songkla University, for the financial support provided under the PEA-PSU program. The authors also gratefully acknowledge the Meter System Development Division, PEA Area 3 (Southern) Yala Province, for data support and valuable comments.

Table 14: *Comparison with benchmark datasets.*

Comparison	[14]	[15]	[16]	[18]	[19]	[20]	[21]	[22]	**This paper**
Extraction	-	statistics	expert knowledge	temporal, locality, similarity, infrastructure	PCA	EMD	-	time and freq. domain, PCA	QB, Z-score, capturing EC, LOF, distance
Using electrical profiles	yes	yes	yes	yes	yes	yes	yes	yes	yes
Using physical data	no	no	yes	yes	no	no	no	no	yes
Classifier	NN	DT	-	**RF**, LR, SVM	SVM	KNN	RF	DNN	KNN, LR, NB, RR, DT, **RF**
Number of classes	2	2	2	2	2	2	2	2	3
Precision	0.6200	0.7867	0.4403	0.7887	0.6207	0.7662	0.6134	0.7907	0.8022
Recall	0.6218	0.7897	0.4732	0.7867	0.5850	0.7742	0.6424	0.7799	0.8354
F1 score	0.6578	0.7246	0.4874	0.6351	0.5781	0.7612	0.6307	0.8157	0.8154

References

[1] M. E. De Oliveira, A. Padilha-Feltrin and F. J. Candian, "Investigation of the Relationship between Load and Loss Factors for a Brazilian Electric Utility," *2006 IEEE/PES Transmission & Distribution Conference and Exposition: Latin America*, Caracas, Venezuela, pp. 1-6, 2006.

[2] G. M. Messinis and N. D. Hatziargyriou, "Review of non-technical loss detection methods," *Electric Power Systems Research*, vol. 158, pp. 250-266, May 2018.

[3] Performance report according to the PEA operational plan for the year 2021 quarter 1-4 (1 January – 31 December 2021), Available from `https://www.pea.co.th/en/About-PEA/Operating-Results/Performance-Report`.

[4] Provincial Electricity Authority, Knowledge of loss of electricity, One Point Knowledge, OPK Code 59120015, 22 Mar 2016, pp.1-22.

[5] Provincial Electricity Authority (PEA), Provincial Electricity Authority Regulations On the practice of meters, 2019.

[6] J. L. Viegas, P. R. Esteves, R. Melicio, "Solutions for detection of non-technical losses in the electricity grid: A review," *Renewable Sustainable Energy Reviews*, vol. 80, pp. 1256–1268, 2017.

[7] B. C. Costa, B. L. a. Alberto, A. M. Portela, W. Maduro, O. Eler, andB. Horizonte, "Fraud Detection in Electric Power Distribution Networks Using an ANN-Based Knowledge-Discovery Process," *International Journal of Artificial Intelligence & Applications (IJAIA)*, vol. 4, no. 6, pp.17–23, 2013.

[8] P. Glauner *et al.*, "Large-scale detection of non-technical losses in imbalanced data sets," *2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, Minneapolis, MN, USA, pp. 1-5, 2016.

[9] P. Glauner, J. A. Meira, L. Dolberg, R. State, F. Bettinger and Y. Rangoni, "Neighborhood Features Help Detecting Non-Technical Losses in Big Data Sets," *2016 IEEE/ACM 3rd International Conference on Big Data Computing Applications and Technologies (BDCAT)*, Shanghai, China, pp. 253-261, 2016.

[10] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 1162–1171, 2010.

[11] J. I. Guerrero, I. Monedero, F. Biscarri, J. Biscarri, R. Millan, and C. Leon, "Non-Technical Losses Reduction by Improving the Inspections Accuracy in a Power Utility," *IEEE Transactions on Power Systems*, vol. 8950, no. c, pp. 1–11, 2017.

[12] M. Buzau,J.Aguilera, P.Romero,and A.G´omez-Exp´osito, "Detection of Non-Technical Losses Using Smart Meter Data and Supervised Learning," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, 2016.

[13] Z. Zheng, Y. Yang, X. Niu, H. -N. Dai and Y. Zhou, "Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to

Secure Smart Grids," in *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1606-1615, April 2018.

[14] V. Ford, A. Siraj, and W. Eberle, "Smart Grid Energy Fraud Detection Using Artificial Neural Networks," no.1, pp. 0–5, 2014.

[15] C. Cody, V. Ford and A. Siraj, "Decision Tree Learning for Fraud Detection in Consumer Energy Consumption," *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Miami, FL, USA, pp. 1175-1179, 2015.

[16] J. No, S. Y. Han, Y. Joo, and J.-H. Shin, "Conditional abnormality detection based on AMI data mining," *IET Generation, Transmission & Distribution*, vol. 10, no. 12, pp. 3010–3016, 2016.

[17] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, 2016.

[18] J. A. Meira *et al.*, "Distilling provider-independent data for general detection of non-technical losses," *2017 IEEE Power and Energy Conference at Illinois (PECI)*, Champaign, IL, USA, pp. 1-5, 2017.

[19] . N. Toma, M. N. Hasan, A. -A. Nahid and B. Li, "Electricity Theft Detection to Reduce Non-Technical Loss using Support Vector Machine in Smart Grid," *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, pp. 1-6, 2019.

[20] S. Aziz, S. Z. Hassan Naqvi, M. U. Khan and T. Aslam, "Electricity Theft Detection using Empirical Mode Decomposition and K-Nearest Neighbors," *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, Karachi, Pakistan, pp. 1-5, 2020.

[21] R. Yadav, Y. Kumar, "Detection of non-technical losses in electric distribution network by applying machine learning and feature engineering," *Journal European des Systems Automatizes*, vol. 54, no. 3, pp. 487-493, June 2021.

[22] L. J. Lepolesa, S. Achari and L. Cheng, "Electricity Theft Detection in Smart Grids Based on Deep Neural Network," in *IEEE Access*, vol. 10, pp. 39638-39655, 2022.

Supakan Janthong received the B.Eng. degree in electrical engineering from Prince of Songkla University, Songkhla, Thailand, in 2015, He is currently pursuing the M.Eng. degree. He is currently an Electrical Engineering with Meter Department, Provincial Electricity Authority. His research interests include data science, signal processing, deep learning, and power system analysis.

Rakkrit Duangsoithong received the B.Eng. degree in electrical engineering from Chiang Mai University, Thailand, in 1995, the M.Eng. degree in electrical engineering from Prince of Songkla University, Songkhla, Thailand, in 2001, and the Ph.D. degree from the University of Surrey, Guildford, U.K., in 2013. He is currently an Assistant Professor with the Electrical Engineering Department, the Faculty of Engineering, Prince of Songkla University. His current research interests include machine learning, computer vision, signal processing, and data analysis.

Kusumal Chalermyanont received the B.Eng. degree in electrical engineering from Prince of Songkla University, Songkhla, Thailand, in 1993, the M.Eng. and Ph.D. degree in electrical engineering from University of Colorado at Boulder, USA in 1999 and 2003. She is currently an Assistant Professor with the Electrical Engineering Department, the Faculty of Engineering, Prince of Songkla University. Her current research interests include power electronics, electric drives, renewable energy, and smart grid.