Effect of Check Meter Quantity on Theft Detection in Distribution Networks with Rooftop PV and Net Metering

Kaira Maxine V. Gonzales
Electrical and Electronics Engineering Institute
University of the Philippines Diliman
kaira.maxine.gonzales@eee.upd.edu.ph

Carlos Demetri S. Vicencio
Electrical and Electronics Engineering Institute
University of the Philippines Diliman
carlos.demetri.vicencio@eee.upd.edu.ph

Adonis Emmanuel D.C. Tio, Ph.D.
Electrical and Electronics Engineering Institute
University of the Philippines Diliman
adonis.tio@eee.upd.edu.ph

***Abstract*—The increasing demand for renewable energy has led to the rise of rooftop photovoltaics (PV) and net metering (NM). However, these technologies add a new dimension to energy theft detection, necessitating a more complex approach. This study aims to explore the effect of varying the number of check meters (CM) in distribution networks with rooftop PV and NM on theft detection algorithms, specifically for detecting meter tampering in households. The distribution network was modeled using the Ausgrid Dataset and the IEEE European Low Voltage Test Feeder on OpenDSS and Python. Several power flow simulations were conducted by varying levels of PV and NM penetration, and the number of houses connected per CM. The simulation data was then used to extract the following features: Gamma Deviance (GD), Log Cosh Loss (LCL), Percent Loss Error (PLE) and Poisson Deviance (PD). These were used as input to three machine learning algorithms: Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Decision Tree (DT). The performance of each algorithm was measured through its accuracy. Then, Kendall's Tau correlation test was used to quantify the correlation between the theft detection accuracy and CM quantity. The results showed that LCL and PD exhibited a moderate to high correlation between theft detection accuracy and CM quantity at a 99% confidence level, and are more robust to varying levels of PV and NM penetration for all algorithms. This provides valuable insights into the development and implementation of theft detection systems which will not only reduce the risk of electricity theft but also further promote the transition to sustainable energy.**

***Index Terms*—rooftop photovoltaics, net metering, electricity theft detection, machine learning, check meters**

I. Introduction

As the shift to renewable energy continues, the Philippine Energy Plan outlines policies that reinforce the integration of more rooftop photovoltaics (PVs) and the implementation of Net Metering (NM), an incentive mechanism that allows households and commercial establishments to sell excess electricity generated to the distribution utility [1]. With the environmental and financial benefits provided by rooftop PVs and NM, more consumers are opting for these technologies. However, these also provide new ways for consumers to tamper with meter readings. Aside from reporting a lower energy consumption, pilferers could now also report more energy fed into the grid to increase profit [2]. Acts of electricity theft such as this results in financial losses and power quality problems [3]. As such, researchers must find ways to strengthen electricity theft detection methods.

One such method that has shown huge potential is through the usage of machine learning (ML) algorithms [4]. Previous studies from [5] and [6] have analyzed the capability of ML algorithms to detect theft occurrences on networks with PV and NM using features derived from household consumption and check meter readings as inputs. The use of check meters provides data on the electricity supplied to downstream households which can then be contrasted with the sum of the downstream household readings. However, while these studies provide more information on the performance of energy theft detection methods on networks with PV and NM, they used an unrealistic system with several check meters which could be costly to implement. This project aims to address the practicality of the current works by studying the effect of check meter quantity on the accuracy of theft detection in distribution networks with PV and NM.

II. Electricity Theft Detection

A. Electricity Theft

Electricity theft is a considerable issue within the energy sector as it results in significant financial losses and power quality issues. One type of electricity theft attack is stored demand tampering wherein meter readings are manipulated to gain profit [7]. The pilferer could either decrease their energy consumption, or increase their energy generation for cases with NM.

With the rise of smart grid infrastructures, electricity theft may be addressed using smart meters to facilitate easier monitoring and billing. Smart meters could be used along with a check meter, a device used by distribution utilities to measure the supplied electricity to one or more consumers. Readings of the check meter can be compared to the total readings of the downstream household smart meters as a means to monitor and identify theft occurrences since its readings do not change even if meter tampering occurs [6].

B. Electricity Theft Problem Definition

The check meter readings and the total downstream household meters readings will never match regardless of theft occurrence due to technical losses in the distribution of electricity [6]. As such, classification-based ML algorithms that use features derived from these measurements as input should be capable of distinguishing theft occurrences from technical losses [7]. However, factors such as PV and NM penetration and check meter quantity may affect the performance of theft detection models.

1) Electricity Theft Detection in Networks with PV and NM: The installation of NM introduces new challenges on the distribution networks as the added reverse power flow introduced by NM can increase line losses [8]. Additionally, it allows pilferers to report higher generation values to gain profit. These two challenges pose possible issues on current electricity theft detection methods, so more studies on electricity theft detection for households with both PV and NM are needed to account for the new emerging technology implemented in distribution networks.

In recent years, few studies have worked towards addressing the challenges introduced by PV and NM. The research by [9] applied theft detection algorithms in households with PV and NM. They used root-mean-square-error (RMSE), Support Vector Machines (SVM), Least Square Error (LSE), and Autoregressive Integrated Moving Average (ARIMA) to detect pilferers who increased their generated electricity by checking for significant errors between the actual energy generated per household as recorded by PV smart meters and the estimated energy generated based on solar irradiance and temperature. Their RMSE and SVM models were able to achieve 94.15% and 81% accuracy, respectively, while LSE had the poorest performance among all algorithms.

The study in [10] also used ML algorithms to detect pilferers who increased their generated electricity. They used a hybrid CNN+GRU+Dense theft detection model and other basic machine learning algorithms to compare the individual household smart meters readings with solar irradiance readings and SCADA meter readings which serve a similar purpose to check meters. Their hybrid model had a detection rate of 99.3% while SVM had a detection rate of 88.3%.

The study by [6] focused on understanding the effect of varying PV and NM penetration on electricity theft detection. To do this, they modeled the IEEE European Low Voltage Test Feeder on Open DSS with eight check meters throughout the network. Then, malicious customers were generated by applying a multiplier to randomly selected households to either decrease consumption or increase generation. The percent loss error between the check meter readings and the sum of the individual household meter readings downstream each check meter was then used as input to their SVM and Artificial Neural Network (ANN) models under varying PV and NM penetration levels in order to determine whether the difference is significant enough to be considered as theft. Their results show that the performance of both algorithms decline with increasing PV and NM levels. Moreover, varying the number of households with NM resulted in greater variation in the algorithm performance as compared to varying the number of households with PV.

Building on this information, [5] used the same NM dataset, test feeder model, and check meter topology as [6], but they explored different PV and NM penetration levels and different features and algorithms. For the features, they also explored Gamma Deviance, Log Cosh Loss, Poisson Deviance and Squared Error aside from Percent Loss Error. For the algorithms, they used SVM, ANN, K-Nearest Neighbors (KNN), and Decision Tree (DT). Their results show that using Gamma Deviance and Log Cosh Loss as features resulted in high performance for SVM, ANN and DT, even with increased PV and NM penetration. Meanwhile, all five features performed poorly in KNN, even with minimal households with PV and NM in the system.

2) Role of Check Meters: Increasing the number of check meters in a network decreases the number of households connected to each check meter which subsequently decreases the line losses captured by the check meter. Doing so could affect the performance of the ML algorithm, but it is also costly. The studies by [5] [6] and [10] used 9, 8, and 8 check meters respectively in their simulations to simplify the detection model. This implementation is impractical due to the development, installation, and maintenance cost of check meters. However, reducing the check meter quantity may also affect the performance of the theft detection system.

Since these studies did not explore other check meter quantities, there remains to be no literature on the effect of check meter quantity on theft detection in distribution networks with PV and NM.

III. Methodology

The methodology is divided into three main stages: dataset creation, theft detection machine learning implementation, and performance assessment.

A. Dataset Creation

1) Household Data Acquisition and Cleaning: The Ausgrid solar home half-hour dataset [11] which includes load and PV generation profiles for 300 customers was used as the sample data for this study. Households with controllable loads were discarded and only data from Dec 5-11, 2010 were used to match the weather in the Philippines [12], resulting in seven days worth of half-hour meter readings for 161 customers.

2) Network Modelling: The IEEE European Low Voltage Test Feeder [13] was used to perform the power flow simulations in OpenDSS, an open-source simulation tool used for electrical utility distribution systems [14] This test feeder is connected to a substation through a distribution transformer and consists of 55 loads. Six check meter (CM) configurations with varying CM quantity and placement were used. Per CM configuration, the households were distributed among the different CMs based on proximity and convenience. Table I shows the different CM configurations, and Figure 1 shows

the different placement of CMs per configuration. The load of the test feeders were modeled with five varying levels of PV and NM penetrations as shown in Table II.

TABLE I
CHECK METER CONFIGURATION

CM Configuration	CM Qty	Households per CM
C1	24	2-3
C2	10	4-6
C3	5	9-12
C4	4	11-17
C5	2	23-32
C6	1	55

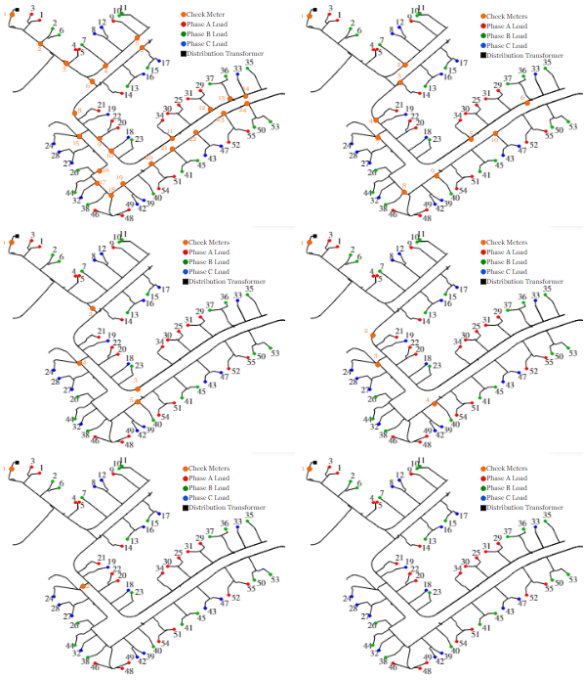

Fig. 1. Check Meter Placements for C1 to C6 (left to right; top to bottom)

The load of the test feeders were modeled with five varying levels of PV and NM penetrations as shown in Table II

TABLE II
PV AND NM CONFIGURATION

PV + NM Configuration	PV + NM Penetration (%)	PV+ NM Houses Quantity
P0	0	0
P1	25	13
P2	50	27
P3	75	41
P4	100	55

3) Power Flow Dataset Creation: For each combination of CM configuration and PV and NM configurations, 220 power flow simulations were performed, each with a different set of household profiles. Power flow simulations were performed in OpenDSS using the cleaned Ausgrid household profiles as input. Per power flow simulation, 55 household profiles were randomly selected from the 161 customers in the cleaned Ausgrid dataset and assigned to the 55 loads in the network. Each power flow simulation resulted in 7 days worth of household and CM readings with a 30-minute interval. Then, the sum of readings per day was obtained to simplify the output which resulted in 7 readings for each of the household and CMs. Since there is no theft present, the totalled meter readings were labeled as benign.

Given the 6 CM configurations, and 5 PV and NM configurations, 30 benign datasets containing 220 power flow simulations each were created. All in all, 46 200 rows of data was made with each row containing the total household and CM readings for a single day.

The malicious dataset was created by duplicating the benign dataset then adding a single pilferer per power flow simulation. A random theft multiplier k was applied to the pilferer's meter readings which would decrease their reading (if the net consumption is positive) or increase energy fed back to the grid (if net consumption is negative). The value of k was chosen randomly from a uniform distribution bounded from 0.1 to 0.8. Equation 1 shows the equation used to determine the malicious value for the pilferer's net meter reading (X).

$$X = \begin{cases} X * k & \text{if } x > 0 \\ X - (|X| * k) & \text{if } x < 0 \end{cases} \tag{1}$$

4) Feature Extraction and Labelling: Four different features were extracted from the benign and malicious datasets which were then used in the machine learning classifiers. These features are: Gamma Deviance (GD), Log Cosh Loss (LCL), Percent Loss Error (PLE), and Poisson Deviance (PD). By using these features, the difference or ratio between the check meter readings and sum of individual household readings are scaled to assist the theft detection model. Equations 2 to 5 present how these are obtained. In these equations, CM refers to the check meter reading in a particular area and $\sum_{n=1}^{i} M_n$ corresponds to the sum of individual household readings under a certain check meter. Features with and without theft were then labeled with 1 or 0 respectively to guide the classifier.

$$GD = 2(log(\frac{\sum_{n=1}^{i} M_n}{CM}) + \frac{CM}{\sum_{n=1}^{i} M_n} - 1) \tag{2}$$

$$LCL = log(cosh(\sum_{n=1}^{i} M_n - CM)) \tag{3}$$

$$PLE = \frac{|\sum_{n=1}^{i} M_n - CM|}{CM} \tag{4}$$

$$PD = 2(CM \log(\frac{CM}{\sum_{n=1}^{i} M_n}) + \sum_{n=1}^{i} M_n - CM) \tag{5}$$

B. Theft Detection Algorithm Implementation

The features extracted were splitted into two groups: 80% for training and 20% for testing. After splitting the data, three different machine learning algorithms were used namely, Support Vector Machine (SVM), Artificial Neural Networks (ANN), Decision Tree (DT). A total of 360 classifiers were

used as there were 3 algorithms, 30 datasets, and 4 features to be used.

1) Support Vector Machine (SVM): SVM is a supervised machine learning algorithm that plots data points in a multi-dimensional space, then finds an optimal separator to classify the data [15]. In this study, the radial basis kernel function was used to handle non-linear data. To implement this, the Support Vector Classifier (SVC) library from Scikit-learn [16] was used, and grid search and ten-fold cross validation were used to tune the cost and gamma hyperparameters.

2) Artificial Neural Networks (ANN): ANN is a computational network that maximizes the use of multiple layers of interconnected nodes in order to recognize certain patterns and map the inputs to their corresponding outputs [17]. The MLPClassifier function from Scikit-learn [18] was used to perform this. The following hyperparameters were tuned using grid search and five-fold cross validation: solver, activation function, batch size and learning rate.

3) Decision Tree(DT): DT is another type of supervised learning algorithm that uses several hierarchical conditional control statements to classify the given data [19]. This was done through the DecisionTreeClassifier function from Scikit-learn [20]. Grid search and ten-fold cross validation were used to tune the minimum samples split and minimum samples leaf hyperparameters.

C. Performance Assessment

The performance of the different algorithms at different configurations was evaluated through its accuracy. This metric is dependent on the value of four different variables: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Equation 6 shows the formula for this metric.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Kendall's B Tau Correlation Test was then used to assess the correlation between CM quantity and theft detection accuracy for each feature-classifier pair with the different PV and NM configurations. Kendall's B Tau is a non-parametric correlation test used for ordinal and continuous variables with several ties and small sample size [21]. The null hypothesis which states that there is no correlation between CM quantity and theft detection accuracy was tested at a 99% confidence level. The tau correlation and p-value was calculated using IBM SPSS Statistics software, a tool that can analyze bivariate statistics [22]. Equation 7 shows how tau was obtained where P is the number of concordant pairs, Q is the number of discordant pairs, and X_0 and Y_0 are the number of pairs tied only to the X and Y variable respectively [23].

$$Tau = \frac{P - Q}{\sqrt{(P + Q + X_0)(P + Q + Y_0)}} \tag{7}$$

IV. Results and Discussion

A. Feature-Algorithm Analysis

Figures 2 to 5 present the box and whiskers plot for each feature-algorithm pair. Each figure contains six columns with decreasing CM quantity, and each boxplot represents the accuracy of the feature-algorithm pair over varying PV and NM penetration levels within the same CM configuration. The box represents the upper and lower quartile with respect to the median accuracy while the ends of the whisker mark the minimum and maximum accuracy value for each feature-algorithm pair. The span of each boxplot reflects the range of the feature-algorithm pair where a thinner boxplot entails more robustness to the presence of PV and NM.

Figures 2 and 3 show that using GD and PLE with 24 CMs to train SVM, ANN, and DT is the least robust to the presence of PV and NM. This means that the performance of the classifiers is not consistent as households with PV and NM are introduced in the network. Additionally, the results for GD and PLE show that using ANN with 5 CMs resulted in the highest median accuracy for both features at 95.94

Figsures 4 and 5 show that using LCL and PD with 24 CMs to train ANN resulted in the highest median accuracy for both features at 100% and 99.03% respectively. Moreover, for PD, the same median accuracy is also obtained when 10 or 5 CMs are used. Additionally, there is also a visible decrease in the median accuracy at CM quantities less than 5 when LCL and PD are used as features in SVM, ANN, and DT. This significant decrease suggests that using 5 CMs in the given test network provides a good trade-off since the marginal increase in accuracy reduces greatly at higher CM quantities.

Based on Figures 2 to 5, DT generally performed worse than ANN and SVM in terms of median accuracy except for GD with 24 CMs. On the other hand, ANN and SVM showed close to similar results in terms of median accuracy and robustness to PV and NM penetration for every feature and CM configuration combination.

The highest median accuracy was attained using ANN- LCL and 24 CMs with a median accuracy of 100%. It also had the smallest range with a span of 0.48%.

B. Check Meter Quantity vs Accuracy Correlation

Table III shows the Kendall's B Tau correlation table which contains the tau correlation coefficient and p value for each feature-algorithm pair. Looking at the column for the p-value, there is a correlation between the CM quantity and theft detection accuracy when LCL and PD are used as features

Fig. 2. GD Boxplot

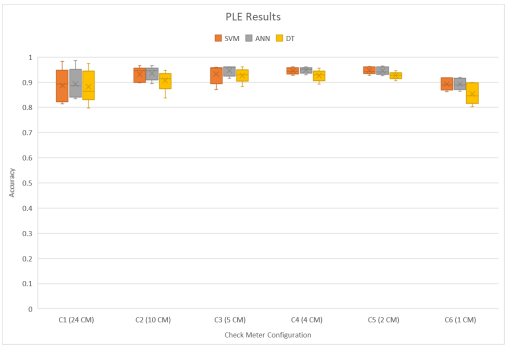

Fig. 3. PLE Boxplot

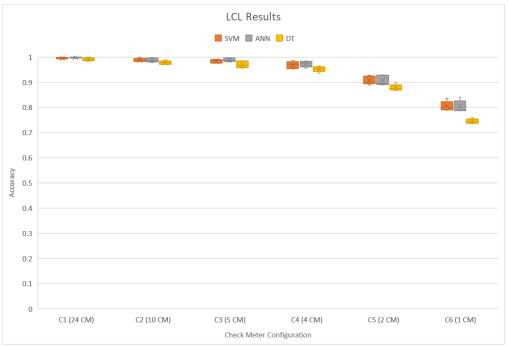

Fig. 4. LCL Boxplot

Fig. 5. PD Boxplot

for SVM, ANN, and DT since they all resulted in a p-value less than 0.001. Furthermore, LCL and PD exhibit a moderate to high positive correlation for all algorithms with tau values ranging from 0.714 to 0.859. On the other hand, there is no correlation between the CM quantity and theft detection accuracy when the features used are GD and PLE for all algorithms since all their p-values are greater than 0.01.

C. General Discussions

For the test network used in this study, determining the "best" classifier-feature pair depends on the actual situation the electricity theft detection system will be used. If there are no constraints on the number of CMs used, any classifier paired with LCL or PD will perform well with at least 5 CMs. However, if there are constraints in the CM quantity such that only one CM (most economical) is allowed, any classifier paired with GD or PLE is better because the CM quantity and theft detection accuracy with these features are not correlated.

TABLE III
KENDALL'S B - TAU CORRELATION

ML Algo	Feature	Tau	p-value
SVM	GD	0.007	0.957
	LCL	0.829	<0.001
	PLE	0.015	0.913
	PD	0.755	<0.001
ANN	GD	0.065	0.637
	LCL	0.831	<0.001
	PLE	0.057	0.676
	PD	0.714	<0.001
DT	GD	0.135	0.327
	LCL	0.859	<0.001
	PLE	0.065	0.637
	PD	0.756	<0.001

The the trend among the features may be visualized through Figure 6 which presents the histograms of the four features used at 100% PV and NM penetration with 24 CMs (left) and 1 CM (right) in the network. For both C1 and C6 only the features of the CM with theft was used.

For C1, since the benign and malicious curves in Figures 6a and 6b have a greater overlap with respect to the total area of the benign curve, there is a less prominent difference between the benign and malicious features for GD and PLE. This makes it harder for their respective models to determine whether that difference is due to technical losses during distribution or due to electricity theft. On the other hand, the benign and malicious curves in Figures 6c and 6d do not overlap as much for C1. This aids the classifiers in detecting the occurrence of theft when there is a high number of CMs.

When the number of CMs decreased, it can be seen that the area covered by the benign curves in in Figures 6a and 6b increased, thereby reducing the overlap percentage. As such, the electricity theft detection accuracy using GD and PLE increased when the CM quantity decreased. Meanwhile, decreasing the CM quantity for Figures 6c and 6d resulted in a greater overlap for the benign and malicious curves. This entails that benign and malicious values coexist for a greater range of LCL and PD values which negatively affects the theft detection capabilities of their respective models. Further discussions on the effect of PV, NM, and CM quantity on the magnitudes of the features can be found in Appendix D.

V. CONCLUSION

In this study, the performance of three algorithms (SVM, ANN, and DT) and four features (GD, LCL, PLE, and PD) on six different check meter configurations with varying PV and NM penetrations were evaluated through their accuracy. The Kendall's B Tau correlation test was also used to further quantify the relationship between check meter quantity and theft detection accuracy.

The results show that when PD or LCL are used as features, the theft detection accuracy increases as the number of CMs in the network increases. Additionally, configuration C3 (with

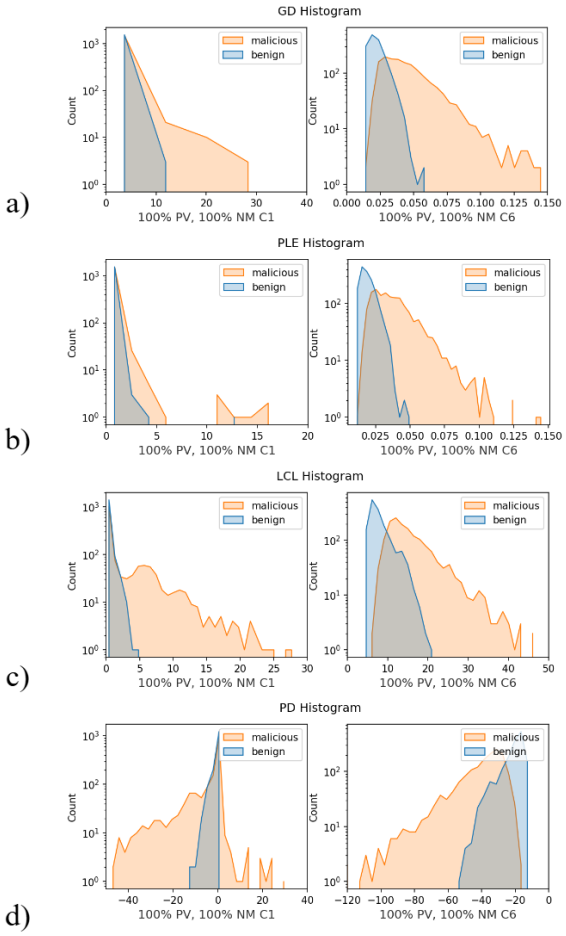

Fig. 6. Histograms for a) Gamma Deviance b) Percent Loss Error c) Log Cosh Loss and d) Poisson Deviance

5 check meters) can be considered as the most practical configuration for any of the tested algorithms paired with either PD or LCL as it is able to create an accurate and robust theft detection system with the least amount of CMs. On the other hand, theft detection accuracy and check meter quantity are not correlated when using GD or PLE as features. The use of GD or PLE is suggested for the chosen test feeder when there are economical constraints because accuracy will not be affected by lower quantity of check meters. In terms of the performance of algorithms, SVM and ANN did not vary much in their results and DT performed worst among all the algorithms used.

Expanding on this topic, future studies could explore using other test systems that have different network topologies, check meter quantities, and theft representation. Furthermore, the optimal placement of check meters can also be studied to improve the theft detection system.

REFERENCES

[1] Department Of Energy. in *Philippine Energy Plan 2020-2040*, page 66, June 2022.

[2] M. -M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero and A. Gomez-Exposito, "Hybrid Deep Neural Networks for Detection of Non-Technical Losses in Electricity Smart Meters," in *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1254-1263, March 2020, doi: 10.1109/TPWRS.2019.2943115.

[3] L. G. Arango, E. Deccache, B. D. Bonatto, H. Arango, P. F. Ribeiro and P. M. Silveira, "Impact of electricity theft on power quality," *2016 17th International Conference on Harmonics and Quality of Power (ICHQP)*, Belo Horizonte, Brazil, 2016, pp. 557-562, doi: 10.1109/ICHQP.2016.7783346.

[4] I. Petrlik et al., "Electricity Theft Detection using Machine Learning,"*International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, pp. 420-425, 2022.

[5] R. M. P. Maala, A. M. B. Rebamba and A. E. D. Tio, "Classification-Based Electricity Theft Detection on Households with Photovoltaic Generation and Net Metering," *TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON)*, Chiang Mai, Thailand, 2023, pp. 1094-1099, doi: 10.1109/TENCON58879.2023.10322383.

[6] C. Lavilla, M. Osorio, and Z. Restituto, "Effect of Net Metering and Rooftop Photovoltaics on Electricity Theft Detection," Capstone Project, EEEI, University of the Philippines, Diliman, 2021.

[7] R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, and X. Shen. Energy-theft detection issues for advanced metering infrastructure in smart grid. Tsinghua Science and Technology, 19(2):105-120, 2014.

[8] R. A. Walling, R. Saint, R. C. Dugan, J. Burke and L. A. Kojovic, "Summary of Distributed Resources Impact on Power Delivery Systems," in *IEEE Transactions on Power Delivery*, vol. 23, no. 3, pp. 1636-1644, July 2008, doi: 10.1109/TPWRD.2007.909115.

[9] M. Shaaban, U. Tariq, M. Ismail, N. A. Almadani and M. Mokhtar, "Data-Driven Detection of Electricity Theft Cyberattacks in PV Generation," in *IEEE Systems Journal*, vol. 16, no. 2, pp. 3349-3359, June 2022, doi: 10.1109/JSYST.2021.3103272.

[10] M. Ismail, M. F. Shaaban, M. Naidu and E. Serpedin, "Deep Learning Detection of Electricity Theft Cyber-Attacks in Renewable Distributed Generation," in *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3428-3437, July 2020, doi: 10.1109/TSG.2020.2973681.

[11] Ausgrid - Solar home electricity data. https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Solar-home-electricity-data.

[12] "The weather year round anywhere on Earth," Weather Spark, https://weatherspark.com/ (accessed Jan. 7, 2024).

[13] IEEE. IEEE PES Test Feeder. https://cmte.ieee.org/pes-testfeeders/resources/ accessed Jan. 7, 2024).

[14] "OpenDSS," EPRI Home, https://www.epri.com/pages/sa/opendss (accessed Jun. 6, 2024).

[15] "Support Vector Machine (SVM) algorithm ," JavaTPoint,https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm. (accessed Jul. 28, 2023).

[16] "SVC," Scikit-learn, https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[17] "Artificial Neural Network (ANN) Tutorial," JavaTPoint,https://www.javatpoint.com/artificial-neural-network, (accessed Jul. 28, 2023)

[18] "MLPClassifier," Scikit-learn, https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

[19] "Decision Tree (DT) Classification Algorithm," JavaTPoint,https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm, (accessed Jul. 28, 2023)

[20] "DecisionTreeClassifier," Scikit-learn, https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[21] "Kendall's tau-B using SPSS statistics," Laerd Statistics, https://statistics.laerd.com/spss-tutorials/kendalls-tau-b-using-spss-statistics.php (accessed May 31, 2024).

[22] "SPSS software," IBM, https://www.ibm.com/spss (accessed Jun. 6, 2024).

[23] "Kendall Tau-b Correlation Coefficient," Penn State, https://online.stat.psu.edu/stat509/lesson/18/18.3 (accessed Jun. 20, 2024).

APPENDIX A
VERIFICATION WITH IEEE LV TEST FEEDER DOCUMENTATION

The original 55 load shapes from [13] were simulated in the test feeder that was used in the power flow simulation of this study using OpenDSS to verify the correctness of the test feeder used. The results of this simulation were then compared to the results that were published in the IEEE European LV Test Feeder documentation. Both the result of the simulation and the documentation provides 1 day worth of readings from a meter placed in Line 1 with 30-minute intervals. Since the measurements were done per phase, the sum of all three phases was summed for simplicity of comparison. Only the real power component was observed since it was the relevant parameter used in the study. Figure 7 shows the comparison between the theoretical measurements from the documentation and the actual measurements from the simulation.

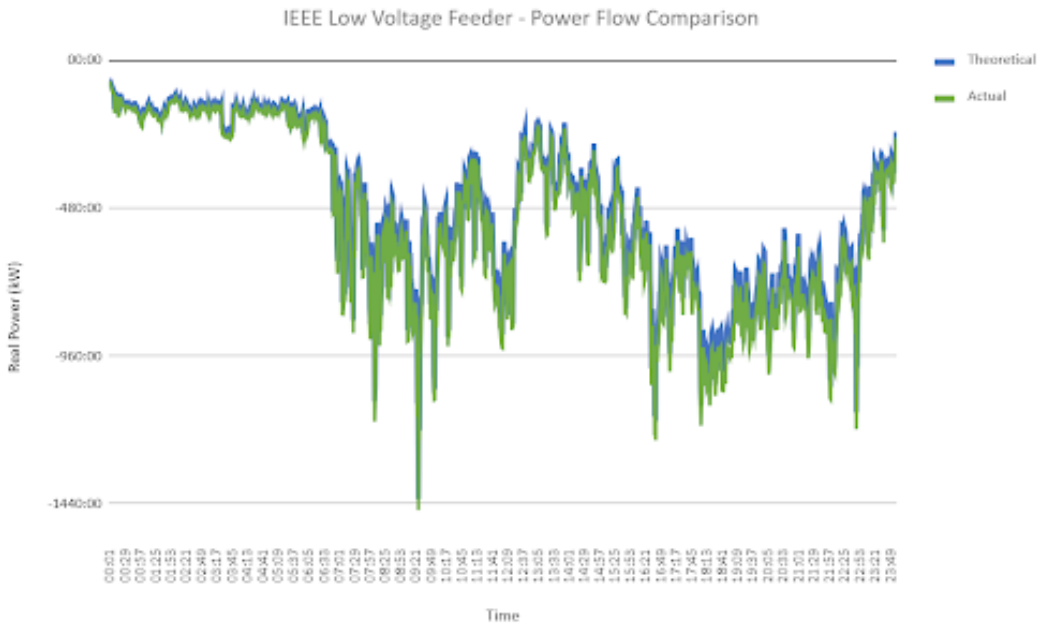

Fig. 7. Comparison of Simulation and Documentation

It can be seen from Figure 7 that the shape of the measurements were identical. The calculated average error was approximately 1.4kW for each measurement interval. This accounts to about 7.41% average error. With minimal error, it was concluded that OpenDSS simulation was similar to the documentation provided by [13]

APPENDIX B
BOX AND WHISKERS PLOTS

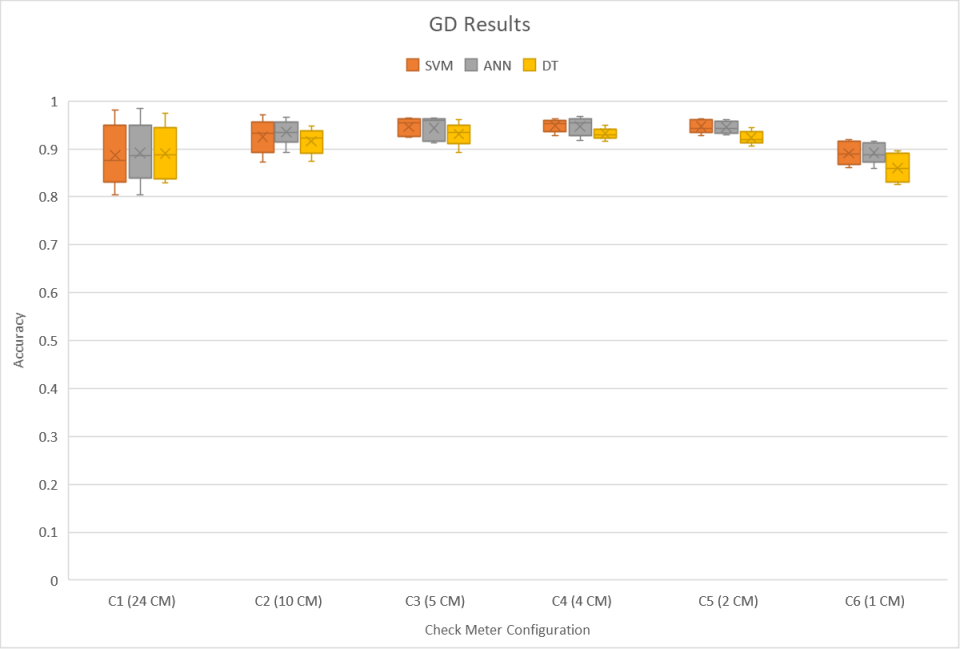

Fig. 8. GD Boxplot

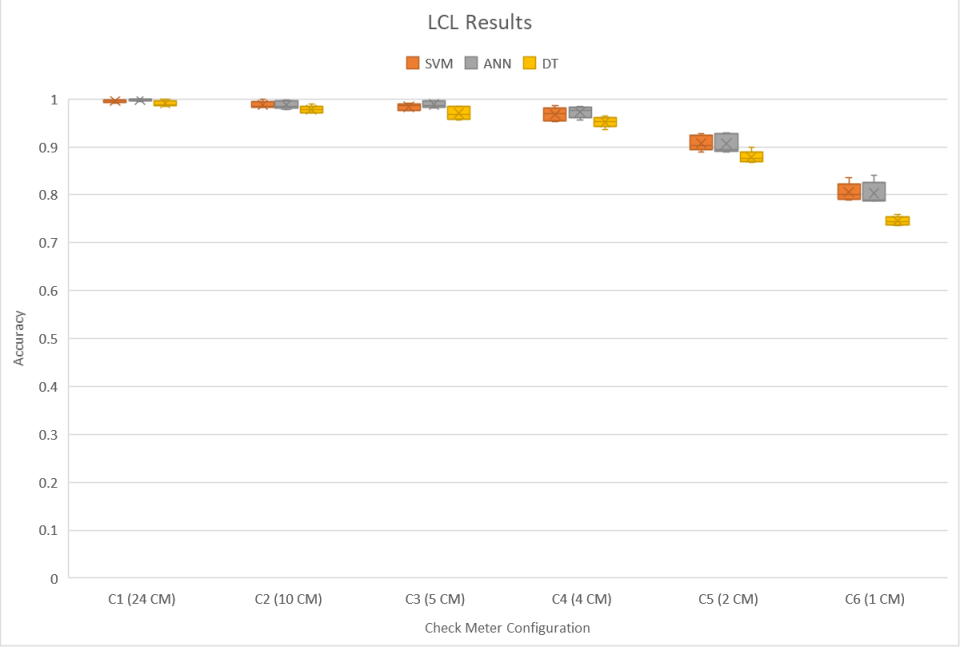

Fig. 9. LCL Boxplot

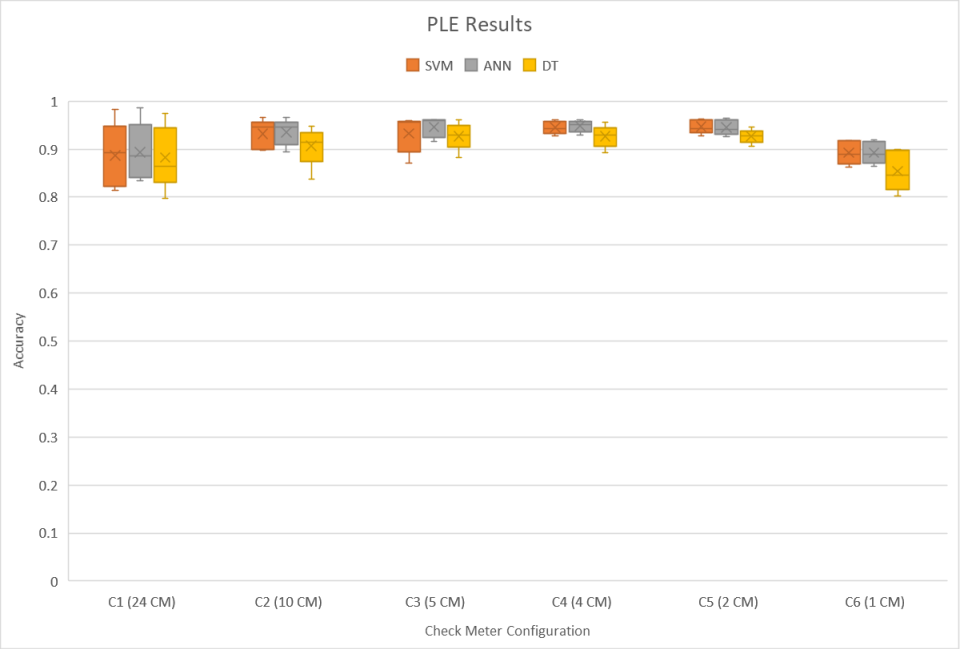

Fig. 10. PLE Boxplot

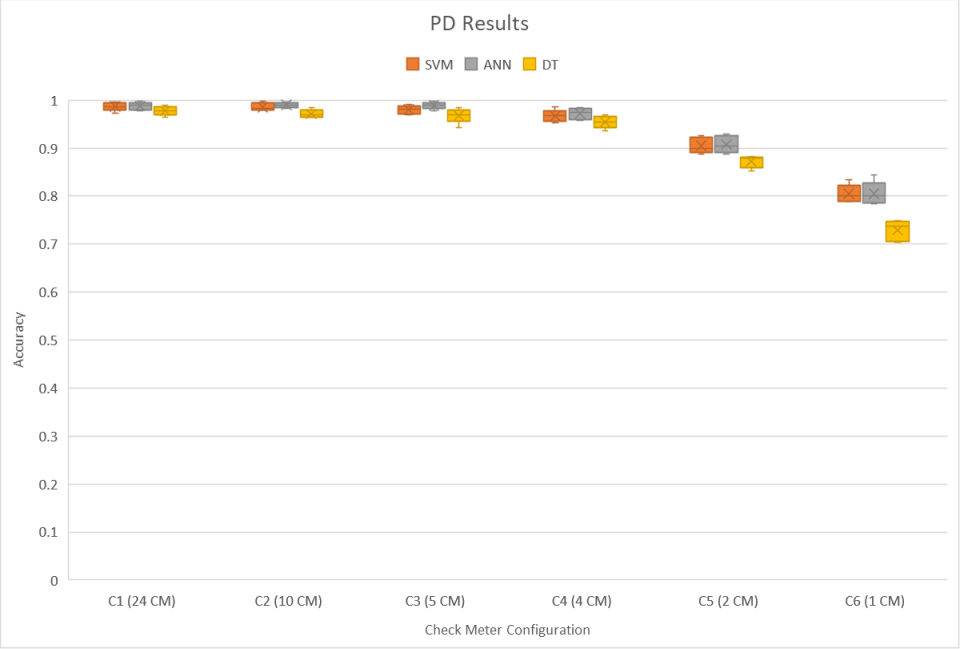

Fig. 11. PD Boxplot

APPENDIX C
BAR GRAPH OF PV AND NM PENETRATION

In order to better visualize the performance of each feature-classifier pair with varying check meter quantity, the following figures shows the accuracy of each feature-algorithm pair for each PV and NM penetrations.

A. SVM

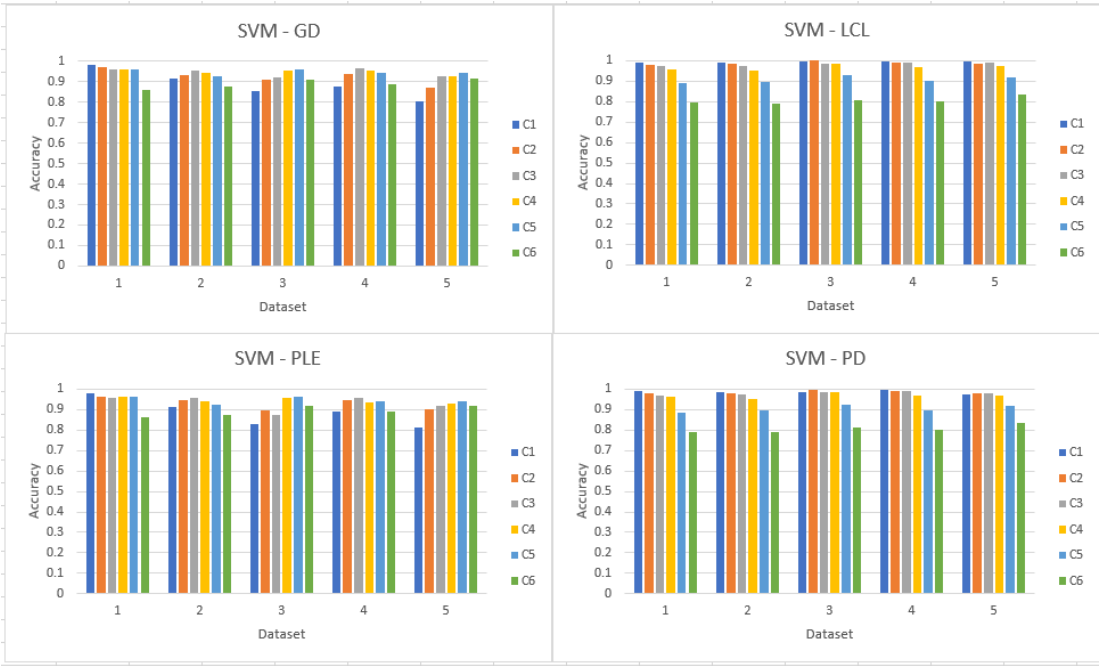

Fig. 12. SVM Bar Graph

B. ANN

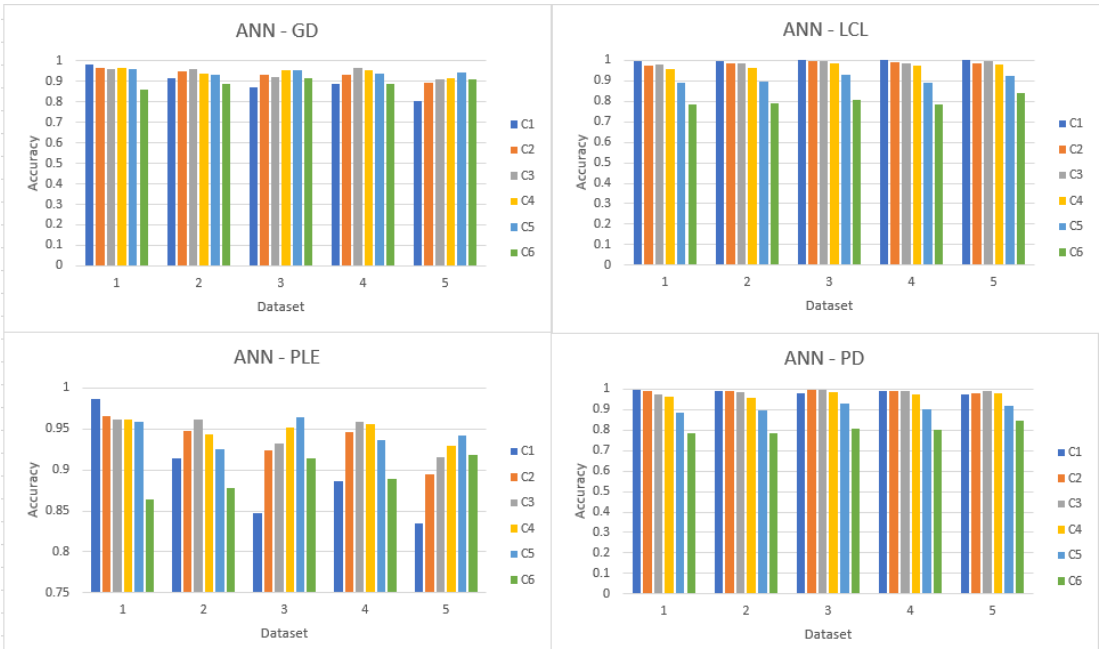

Fig. 13. ANN Bar Graph

C. DT

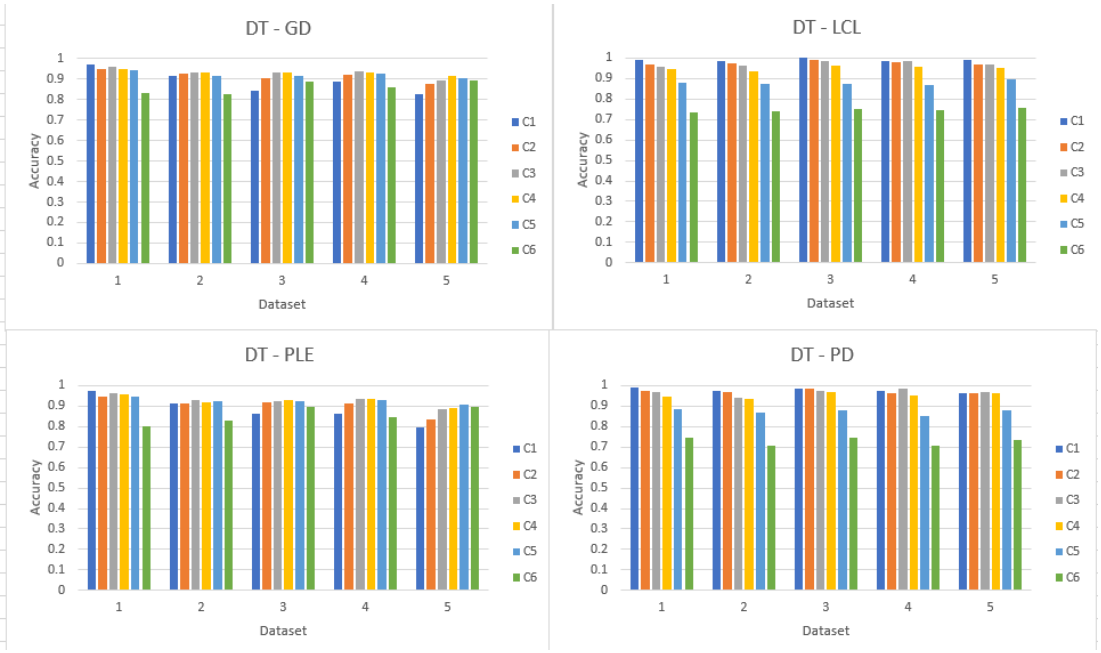

Fig. 14. DT Bar Graph

APPENDIX D
SCATTER PLOTS

A scatter plot from the data of the histogram was made to further analyze the trends shown by the results. Figures 15 to 18 shows the scatter plots for both C1 and C6 per feature. The reference points (orange) represents a linear relationship between the benign and malicious values. Closer points to the reference points mean that benign and malicious points are close to one another.

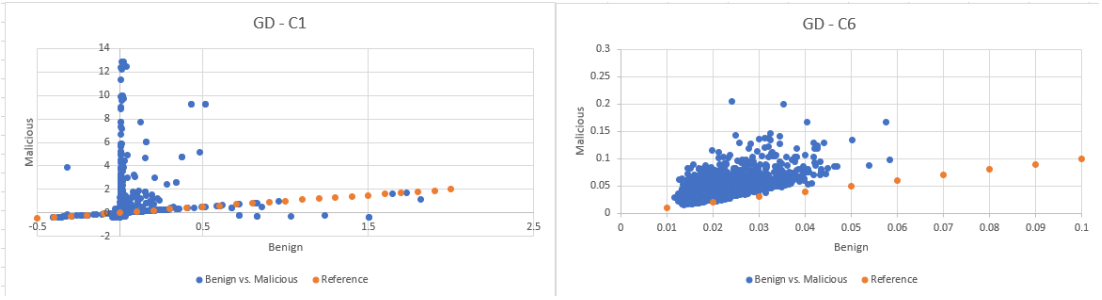

Fig. 15. GD Scatter Plots

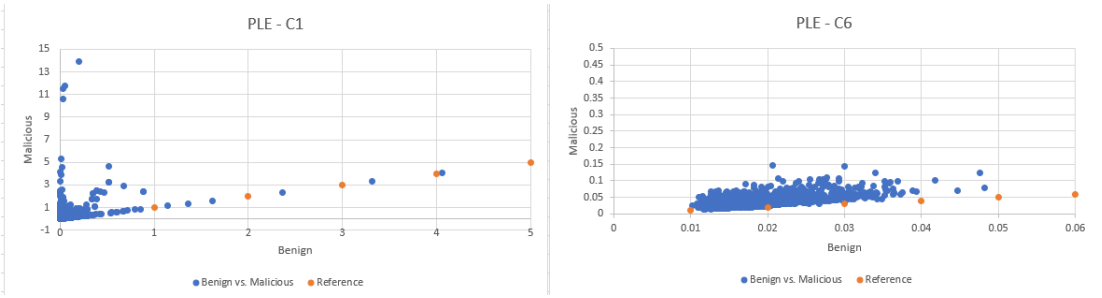

Fig. 16. PLE Scatter Plots

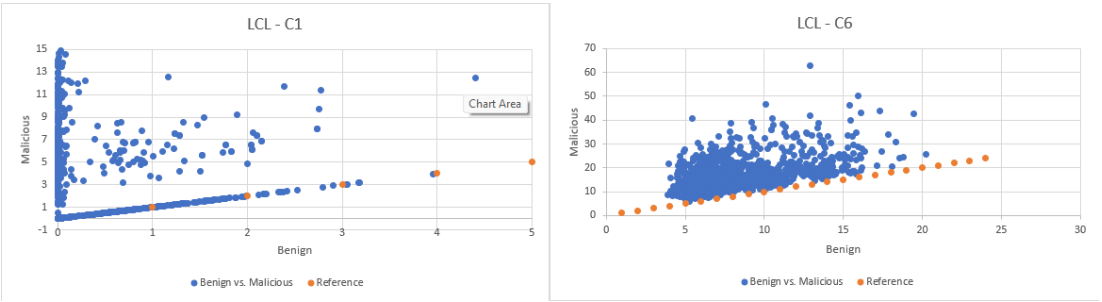

Fig. 17. LCL Scatter Plots

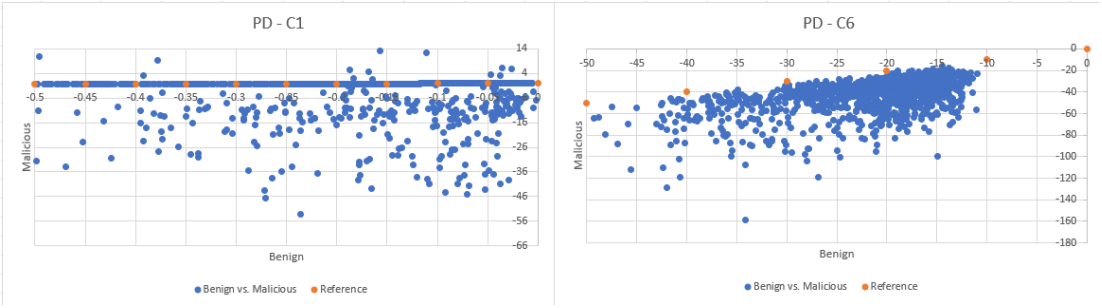

Fig. 18. PD Scatter Plots

For all the features with configuration C1, there are points that lie close to the reference points. However, majority of the points are concentrated towards the y-axis except for PD-C1. This shows that at lower values of the benign features there are high values for malicious features. This observation is due to the high PV and NM penetration and the low number of household connected in a check meter. Based on the theft representation used, there is a greater difference between the malicious and benign meter readings malicious value from the benign when there is a negative net consumption. This means that there are more instances wherein the malicious household meter readings total to a very negative number (eg. -40), while benign sum is only a small positive value (eg. 0.02). Moreover, the small number of households connected to a check meter entails that the technical losses are enough to offset the effect of theft.

Following the explanation above, the ratio between the sum of household readings and the check meter readings becomes larger in C1 as compared to C6. Since the equations for GD and PLE rely on this ratio, the C1 plots for 6a and 6b have a wider domain as compared to their C6 counterparts.

APPENDIX E
PROJECT REPOSITORY

Table IV shows the Python and OpenDSS scripts used in the study, together with their descriptions. A GitHub repository containing these scripts along with their generated files can be accessed through this link: https://github.com/kmvg-upd/ee199-sgrc1

TABLE IV
TABLE OF SCRIPTS USED

clean_ausgrid_dataset.py	Extracts the 161 load profiles and generation profiles from the Ausgrid data
sim_util.py	Functions that set the PV and NM penetration levels
startup.py	Generates the PV and NM combinations
time_series_python.dss	OpenDSS file that setups the LV feeder
simulate_datasets.py	Iterates the OpenDSS simulation for each household combination
adjust_cm_values.py	Adjust CM readings to its designated area and then sum to daily readings
apply_theft_C1.py	Apply random theft multiplier to each household four times for C1 and then sum to daily readings
duplicate_datasets.py	Duplicates households with theft to other CM configurations
sort_days.py	Sort benign and malicious datasets into seven days
merge_dataset.py	Merges benign and malicious datasets
gamma_deviance.py	Extracts the gamma deviance feature from the datasets
log__cosh_loss.py	Extracts the log cosh loss feature from the datasets
percent_loss_error.py	Extracts the percent loss error feature from the datasets
poisson_deviance.py	Extracts the poisson deviance feature from the datasets
tune_SVM.py	Optimizes the hyperparameters for SVM
tune_ANN.py	Optimizes the hyperparameters for ANN
tune_DT.py	Optimizes the hyperparameters for DT
SVM.py	Performs the SVM classification
ANN.py	Performs the ANN classification
DT.py	Performs the DT classification