Efficient Fraud Detection for Smart Meter Security in AMI Using Isolation Forest Algorithm

*Note: Sub-titles are not captured in Xplore and should not be used

1st Md Mehedi Hasan
dept. Computer Science and Engineering
American International University-Bangladesh
Dhaka, Bangladesh
mmhasan@aiub.edu

2nd Mo. Mushfiq Us Saleheen
dept. Computer Science and Engineering
American International University-Bangladesh
Dhaka, Bangladesh
aurpon10@gmail.com

3rd Asif Mahmud
dept. Computer Science and Engineering
American International University-Bangladesh
Dhaka, Bangladesh
asifmahmud001@gmail.com

4th Mo. Golam Faruk Ovi
dept. Computer Science and Engineering
American International University-Bangladesh
Dhaka, Bangladesh
golamfarukovi@gmail.com

5th Nafees Fuad Rahman
dept. Computer Science and Engineering
American International University-Bangladesh
Dhaka, Bangladesh
nfuad2310@gmail.com

***Abstract*—The Advanced Metering Infrastructure (AMI) is essential in modern energy systems to monitor and regulate energy consumption. Smart meters have become popular, but securing the data collected by these devices is a challenge. Fraud detection is vital for smart meter security to detect irregularities and prevent unauthorized system access. An efficient fraud detection system is proposed for AMI smart meter security using the isolation forest algorithm, a machine learning method for detecting abnormalities here. This system focuses on the point-to-multipoint topology and a smart communication system as the data collector to improve precision, efficacy, and data integrity. The performance of the isolation forest algorithm is compared with XGBoost, another machine learning method. The isolation forest algorithm is less prone to overfitting and can detect anomalies in both small and large datasets, making it more suitable for AMI smart meter security. The proposed system can detect anomalies and prevent unauthorized system access, ensuring the security and integrity of data collected from smart meters. The isolation forest algorithm outperforms 3 ML-based models such as the Decision Tree Classifier, Random Forest Classifier, and XGBoost, making it an ideal candidate for fraud detection in AMI smart meter security.**

***Index Terms*—AMI, XGBoost, Isolation Forest, smart meter, machine learning, point-to-multipoint topology.**

I. Introduction

The Advanced Metering Infrastructure (AMI) is an important component of modern energy systems since it allows utilities to monitor and regulate energy consumption in real time [1]. As smart meters become more popular, there is a greater need to protect the security and integrity of the data collected by these devices. Fraud detection is an important part of smart meter security since it may detect irregularities and prevent unauthorized system access [2]. In this research, we propose an efficient fraud detection system for AMI smart meter security based on the isolation forest algorithm, a well-known machine learning method for detecting abnormalities in a variety of applications [3]. We will concentrate on network design and topology, particularly point-to- multipoint topology, as well as the smart communication system between the smart meter and the Meter Data Management System (MDMS) [4]. We chose the point-to-multipoint architecture for our proposed fraud detection system after carefully considering the various possibilities. This design uses dedicated communication lines that are less vulnerable to data loss, corruption, or manipulation than shared communication paths utilized in other topologies, ensuring data integrity [4]. Furthermore, by decreasing noise and streamlining data transport, the point-to-multipoint topology can improve the precision and efficacy of the isolation forest method [4]. The smart communication system will also be used as a data collector in our suggested system. By doing so, we can ensure that all data collected is clean and accurate, limiting the possibility of false positives and false negatives caused by noisy data [4]. We will compare the performance of the isolation forest algorithm with XGBoost, another popular machine learning method, in addition to proposing an efficient fraud detection

system for smart meter security in AMI using the point-to-multipoint topology and a smart communication system as the data collector. Moreover, the isolation forest algorithm is less prone to overfitting, which is a common problem in machine learning algorithms [3]. Overfitting occurs when a model becomes too complex and starts fitting the noise in the data instead of the underlying patterns. The isolation forest algorithm is less susceptible to overfitting because it randomly selects features and splits the data, making it more resilient to noisy data [3]. While XGBoost is a sophisticated machine learning algorithm that has been utilized in different applications for fraud detection, the isolation forest approach has some features that make it better ideal for smart meter security in AMI [5]. The isolation forest algorithm's capacity to detect

abnormalities in high-dimensional datasets efficiently is one of its primary features. This is especially true for smart meter data, which might contain a significant number of attributes [3]. Furthermore, the isolation forest algorithm is capable of detecting anomalies in both small and large datasets, making it perfect for detecting even little thefts that may go unnoticed. This is critical in the context of AMI smart meter security, since even minor anomalies can have a major impact on system performance [5]. Furthermore, the isolation forest algorithm is capable of detecting anomalies in both small and large datasets, making it perfect for detecting even little thefts that may go unnoticed. This is critical in the context of AMI smart meter security, since even minor anomalies can have a major impact on system performance [5]. To summarize, while XGBoost is a powerful machine learning algorithm, the isolation forest technique has various features that make it more suited for AMI smart meter security. Its capacity to detect abnormalities in high-dimensional datasets effectively, resistance to overfitting, and ability to detect even little thefts make it an attractive candidate for fraud detection in AMI smart meter security [3] [5]. Overall, our suggested method for smart meter security in AMI employing the isolation forest algorithm and the point-to-multipoint topology with a smart communication system as the data collector can detect anomalies and prevent unauthorized system access [3] [4]. The network architecture, topology, and smart communication system will determine the system's effectiveness and reliability, making them critical parts of the proposed system [4]. We can design an efficient and trustworthy AMI system that ensures the security and integrity of data collected from smart meters by maximizing these factors.

II. Literature Review

The security of smart meters used in modern metering infrastructure has come under increased scrutiny in recent years (AMI). These tools have developed into a top target for manipulation by dishonest consumers trying to avoid paying for power use. These tools are used to monitor energy consumption and allow communication between the utility companies and customers. A recent occurrence recorded in 2022, [6] when 51 million Americans Were Affected by Healthcare Data Breaches owing to Poor Security, illustrated the possible consequences of meter manipulation. This incident emphasizes the critical necessity for strong security measures to guard against tampering with smart meters and unauthorized access while maintaining the accuracy of meter data.Dishonest consumers may falsify meter data by taking advantage of security flaws in smart meters, costing energy providers money. To counter these risks and preserve the dependability and sustainability of the electricity grid, it is crucial to build efficient security measures and fraud detection tools. Researchers have started working on this matter since 2013 [7], whereimplementation of anomaly detection techniques in smart meters to detect any unusual theft caused by physical-cyber-attack on smart meter. Raciti gives a thorough description of embedded systems' application in anomaly detection as well as cyber-physical systems' function in smart meters. The application of machine learning methods, particularly unsupervised learning, for smart meter anomaly detection is also covered in the book. Later that year in 2014 [8] there was a unified framework proposed. The framework integrates multiple privacy-protecting strategies, such anonymization and differential privacy, with security methods like encryption, access restriction, and intrusion detection. In 2018 research was performed focused on [9] developing an open-source smart meter platform that enables data sharing, security, and traffic analysis.The author offers an open-source, message broker-based system for communicating data from smart meters. Real-time traffic analysis is possible with this technology, which is safe and scale-able and enables the detection of potential security risks. By the next year in 2019, a handful of research was published regarding this matter. One of them was [10] a thorough analysis of the privacy and security risks raised by smart grid metering networks.The paper also discusses other types of attacks that may be made against smart grid metering networks, such as replay assaults, denial of service attacks, and meter manipulation. While another research [11] focused on smart meter security in relation to advanced metering infrastructure (AMI). The article analyzes the weaknesses and potential risks that smart meters confront and offers a few solutions to deal with these problems. The author talks about how different assaults, such tampering with meters, affect both the energy company and its consumers. In order to protect the privacy and security of the data exchanged between smart meters and the utility provider, the article also examines the significance of secure communication routes and encryption techniques.In order to identify and stop fraudulent actions, the author also recommends using intrusion detection systems and machine learning techniques. As in 2020 in this new decade the research on this sector significantly increased due to world evolution to the advance modern age. In this regard there was research based on the [12] significance of smart meter security in relation to managing smart grids. The authors emphasize the flaws in smart meters as well as the threats that might result from them, such power theft and illegal access to private information. A significant amount of sensitive data is being passed using AMI. And in order to

safeguard these data numerous only best possible approach technique is machine learning based solution. Later that year in 2020 [13] A new technique for identifying electricity theft utilizing the K-Nearest Neighbors (KNN) algorithm and empirical mode decomposition (EMD). The suggested technique entails utilizing EMD to decompose the load profile data from smart meters to get Intrinsic Mode Functions (IMFs). Finally, using KNN to evaluate the collected IMFs, the patterns of energy usage are examined. The suggested system is tested using actual data from a Pakistani distribution business, and the findings demonstrate that it can accurately and efficiently detect energy theft. Two research projects were carried out in the area of smart grid technologies in 2022. [14] One research examined the cybersecurity of the infrastructure for smart meters and suggested using the Median Absolute Deviation (MAD) approach to identify and stop malicious assaults. The second [15] research put forth an attribute-based encryption (ABE) and safe data aggregation-based approach to preserving data privacy and security in smart metering systems. In order to protect against cyber-attacks and maintain data privacy, both studies addressed important concerns in smart grid technology and stressed the significance of appropriate security measures. A new article from 2023 [16] emphasizes the rising worries about privacy in smart metering systems. It draws attention to the need to safeguard personal data and talks about technologies that improve privacy, including anonymization, encryption, and differential privacy.

III. Network Architecture

The network design is critical to the overall performance and efficiency of the isolation forest algorithm-based fraud detection system for smart meter security in AMI [1]. The selection of network architecture, especially the point-to-multipoint topology and smart communication system designed between the smart meter and the Meter Data Management System (MDMS), is essential to delivering efficient and accurate fraud detection. There are two most used topologies for designing the architecture for AMI.

A. Point to Multi-point Topology

Data may be delivered from many sources to a central system thanks to the point-to-multipoint topology's effective connectivity between several smart meters and the MDMS. This architecture can make it easier to gather, aggregate, and analyze real-time data, which is necessary for efficient fraud detection in AMI. Smart meters can interact with a single data collector whereas data collectors can connect with numerous meters in a point-to-multipoint network structure, as shown in Fig. 1 [11]. When opposed to the unlicensed ISM band, the licensed RF band has less noise. As a result, data collectors and meters may communicate in both directions across great distances.

In a one-to-many or many-to-one communication pattern, data is transported from a central hub to several smart meters in a point-to-multipoint topology. This design reduces interference from other devices or sources by providing dedicated communication lines between the central hub and each individual smart meter.

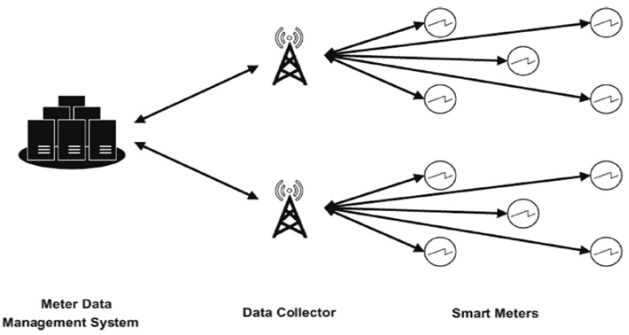

Fig. 1. Point to Multi-point Network Topology

AMI systems must maintain data integrity since precise and trustworthy data is necessary for fraud detection and other system functions. By employing dedicated communication pathways, which are less susceptible to data loss, corruption, or manipulation than shared communication paths used in other topologies, point-to-multipoint topology assures data integrity. The speedy data transfer capabilities of point-to-multipoint topology are another benefit. Several smart meters may easily collect and send data in parallel to the central hub or vice versa, enabling real-time or almost real-time data gathering and processing. By enabling quicker processing and anomaly identification, it can significantly improve the timeliness and efficiency of fraud detection algorithms, such as the isolation forest method.

B. Mesh Network Topology

Smart meters can use various communication methods such as the RF mesh network, the wireless star network, and the PLC method. Combining the most suitable method for each case enables a high area coverage rate and fast expansion into new areas. In Advanced Metering Infrastructure (AMI) systems, the mesh network topology is a common communication architecture [17].

Smart meters can connect in between and with a central hub through a network known as an RF mesh network utilizing radio frequency (RF) transmissions. Smart meters work together to create a self-healing mesh network in this architecture, where each meter serves as a relay for the others, providing reliable communication even in the event of individual meter failures. Instead of passing information through other meters, a wireless star network allows for direct contact between each smart meter and a central hub. This architecture is appropriate for instances when the smart meters are reasonably near to the hub or where direct connection is preferred since each one has a direct communication link with the central hub. With PLC (Power Line Communication), smart meters may talk with one another across the current power distribution lines. This technique eliminates the need for extra communication

infrastructure by allowing smart meters and other devices to communicate over the power lines.

The topology of the mesh network, whether it be RF mesh, wireless star, or PLC, is determined by many variables, including the location of the region, the density of smart meters, the need for communication, and the cost. Each architecture offers benefits and drawbacks in terms of security, scalability, coverage, and dependability. To create an effective and trustworthy AMI system, proper analysis and optimization of the mesh network architecture are crucial. The success of fraud detection algorithms like the isolation forest method that you are utilizing in your research, as well as the overall performance and effectiveness of the AMI system, can be considerably impacted by the mesh network topology decision.

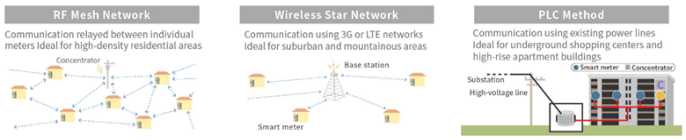

Fig. 2. RF Mesh, Wireless Star and PLC Network Methods

We will be considering point-to-multipoint topology over mesh topology because utilizing the isolation forest method to conduct research on effective fraud detection in smart meter security in AMI, point-to-multipoint topology can be favorable. The dedicated communication pathways provided by point-to-multipoint architecture can reduce noise and enhance data transfer, producing cleaner and more trustworthy data for analysis. This is especially crucial when using a machine learning technique for anomaly detection because the algorithm's performance evaluation and training depend on precise and clean data. Point-to-multipoint topology can enhance the precision and efficacy of your machine-learning technique by minimizing noise and optimizing data transfer, as well as the possibility of false positives or false negatives brought on by noisy data. To guarantee the integrity and dependability of your study findings in the context of smart meter security in AMI, using point-to-multipoint topology might be a wise decision.

C. Smart Communication System

In the network, every home has a Home Area Network (HAN). Inside the HAN there are multiple IoT devices and a Home Energy Management System (HEMS) which is connected to the HAN. HEMS usually manages household energy consumption and improves energy efficiency. Every HAN relates to a distinct smart meter. All the smart meters are then either connected to the Field Area Network (FAN) or Wide Area Network (WAN). Meter Data management systems (MDMS) are made up of several subsystems with various applications, such as managing utility costs for service providers, optimizing power grids, and interacting with consumers. For instance, customers can communicate with one another via a personalized website to monitor and manage their electricity use. The consumption data may be accessed by other utilities for improved analysis. This communication system sits between the smart meter which is HAN side device and the MDMS which is provider side control system. This thing work as a data collector in the point- to-multipoint topology shown in fig. 3.

Methodology

D. Data Collection

Due to the complexity and volume of data involved, as well as problems with data quality and accessibility, gathering and evaluating energy data can be difficult. By investigating the use of publicly accessible datasets, such as the Energy Anomaly Detection competition dataset on Kaggle, we hope to overcome some of these difficulties in this research project. This dataset includes a sizable amount of data approximately 200 commercial buildings on energy use, including details on energy use, meter readings, square-feet, temperature, and humidity, etc. The dataset's original size is around 1,048,575 samples.

E. Data Pre-processing

1) Feature Engineering: We conducted data preprocessing by first performing feature engineering to take only important features. By examining the data and comprehending the domain, the important features were initially identified. With the use of a visualization tool like Seaborn, generate a heatmap of the correlation matrix. To determine whether or not they are correlated feature, we saw there is no highly correlated features in the dataset to remove. we construct a heatmap using Seaborn's heatmap() tool.

2) Missing Values Handle: We made the decision to remove any missing values and outliers from the dataset because we were working with a huge dataset, which would ensure the accuracy of the data. Instead of imputing the missing values and outliers, the outcome might not change if the missing values and outliers are dropped.

3) Data Splicing: Data splitting is a crucial step. The process of splitting the available data into a training set and a testing set is known as data splicing. The testing set is used to assess the performance of the machine learning model, whereas the training set is used to train the model. we split the data into training and testing sets with a ratio of 70:30 to allow for model training and evaluation.

4) Standardization: Finally, we performed standardization on the numerical variables to normalize their values, ensuring that all variables have the same scale and preventing any variable from dominating the model training.

$$\mathrm{x}_s caled = ((x - mean))/(standard_d eviation)$$

where x represents the input feature's initial value, mean represents its mean value in the training set, and standard deviation represents its standard deviation. To make the input features more appropriate, the Standard Scaler scales them to have a mean of 0 and a standard deviation of 1.

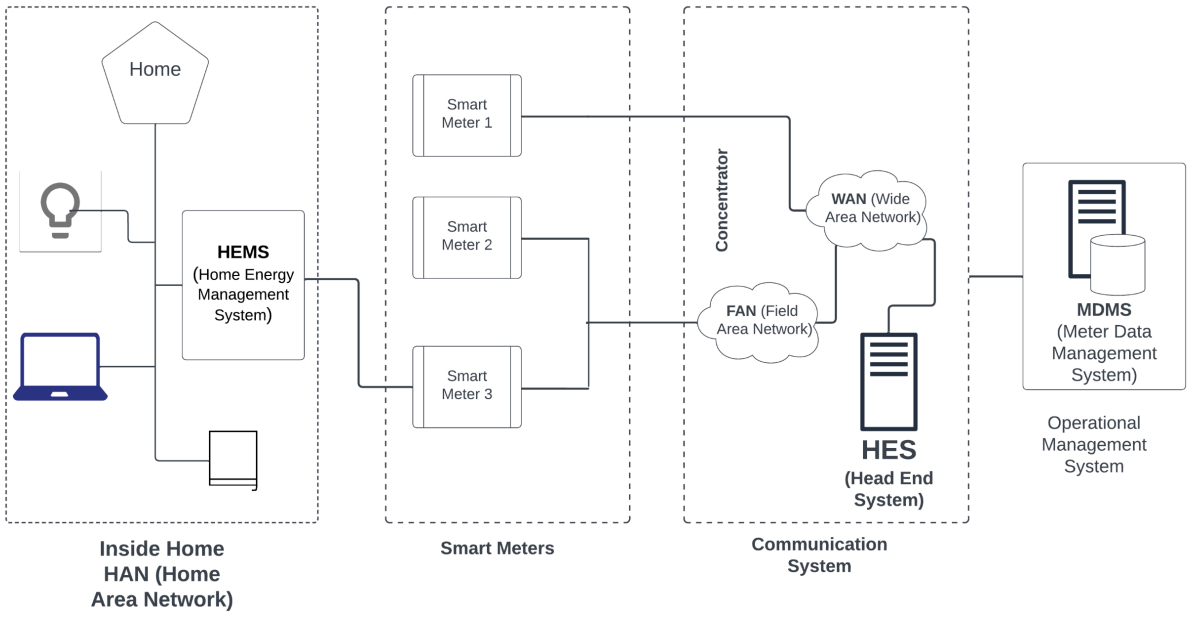

Fig. 3. Smart Communication System

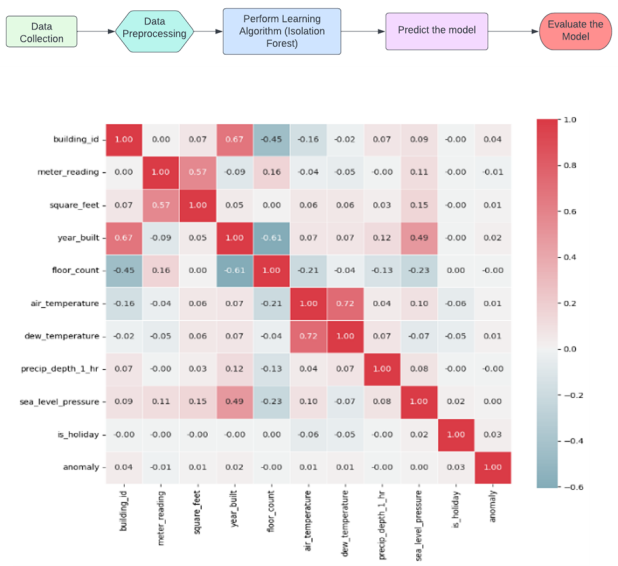

Fig. 4. Feature Engineering Heat-map

F. Perform Isolation Forest Algorithm

The main tenets of the Isolation Forest algorithm are reviewed in this section. A forest of isolation trees that are constructed randomly is used to find outliers in the data. Each tree is cultivated using a separately collected sample that is a subset of the examined dataset. The trees are known as isolating because they repeatedly divide the space being studied until either all of the distinct points are contained inside their leaves or the depth limit is achieved. By submitting the data to the growing forest and generating an anomaly measure for each data point, outlier detection is carried out. The formula is used to determine the measurement.[1]

s(x,n)=2$^{(}$ $-$ $(E(h(x)))/(C(n)))$.....(1)

$Where$

C(n)=2H(n-1)- (2(n-1))/n .....(2)

H(n)=ln(n)+0.57772156649 .....(3)

In the formulae (1), (2), (3) the E(x) term stands for the average lengths of all the paths that have to be gone through in the isolating trees in order to isolate the point x. C(n) is theoretical average path length of unsuccessful search in the binary search tree. The original publication can be consulted by a reader who is particularly interested because it contains the original Isolation Forest building procedure and the formula for determining the anomalous score. However, the addition in this study focuses on the algorithm for creating an isolation tree. As a result, we remember it as pseudocode; for more information, see Algorithm 1 [18].

Algorithm 1 : iTree(X,e,l) Inputs: X – input data, e – current tree height, 1 – height limit Output: iTree 1. if el or —X—1 then

2. return exNode size ←—X—
3. else
4. let Q be a list of attributes in X
5. randomly select an attribute qQ
6. p←uniformRandomReal(form←min(X_q), to ¸$max(X_q)$
7. X_l ¸$filter(X_q p)$
8. X_r ¸$filter(X_q > p)$
9. return inNode Left ←iTree($X_l, e+1, l$),
10. Right ←($X_r, e+1, l$),
11. SplitAtt ←q,SpilitValue ←p

12. end if

Result Analysis

In this article, all experiments are carried out on a computer equipped with an Intel(R) Core (TM) i7-8550U processor running at 1.80 GHz or 1.99 GHz and 8 GB of RAM, and all programming is done on a Jupyter Notebook. Precision, FPR, recall, and F1 Score are the evaluation measures used to gauge the effectiveness of the suggested methodology. The Isolation Forest Algorithm's parameters are set according to necessity.

DTC [19] A Decision Tree Classifier is a classification problem-solving supervised machine learning method. It operates by dividing the data into subgroups depending on a set of rules generated from the data's characteristics. These rules are depicted as a tree structure, with each node representing a feature-based judgment and each leaf node reflecting a final classification result. A DTC is used to properly categorize fresh data based on patterns in the training data. It is widely utilized in industries such as banking, marketing, and healthcare to detect fraud, segment customers, and diagnose diseases.

RFC c [20] The machine learning algorithm known as RFC, or Random Forest Classifier, is a member of the family of ensemble learning techniques. To increase accuracy and lessen the overfitting issue, it generates a lot of decision trees and integrates their predictions. To boost variety and decrease correlation between the trees, each tree in the forest is trained using a random subset of the training data and characteristics. Where precise and reliable categorization is required, such as in banking, healthcare, and image recognition, RFC is frequently utilized.

XGBOOST [21] Extreme Gradient Boosting, often known as XGBOOST, is a potent machine learning approach for regression and classification issues. Due to its capability to handle unbalanced datasets and good accuracy performance, it is designed to be extremely scalable, efficient, and has gained popularity in anomaly detection jobs. By creating an ensemble of decision trees and repeatedly learning from the mistakes made in earlier rounds, XGBOOST is used to increase the detection rate of abnormalities in anomaly detection.

G. Evaluation Metrics

The aim of classifying the energy consumption data into two categories—normal and abnormal (signifying probable theft)—can be stated as a binary classification issue to solve the challenge of detecting smart meter fraud. The performance of proposed methods is evaluated by precision (Pre), Recall (Re) and F1 Scores (F1). The confusion matrix has been used to evaluate anomaly detection.

Here is the metric that we use to determine the performance from table II and the definition of precision, Recall and F1 Scores can be written as,

Precision = True Positives / (True Positives + False Positives) [22]

Recall = True Positives / (True Positives + False Negatives) [23]

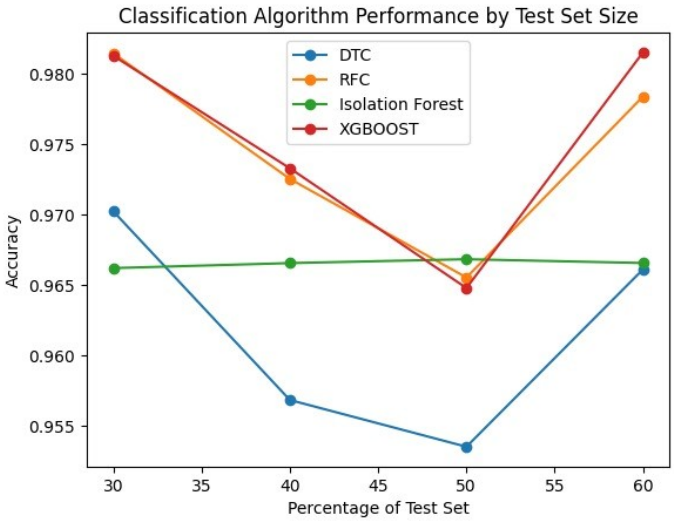

Fig. 5. Accuracy % graph

F1 Score = 2 * (Precision * Recall) / (Precision + Recall) [24]

From the above evaluation we can conclude that Isolation Forest is the better algorithm compared to other methods like DTC, RFC and XGBOOST. Here in accuracy sometimes other methods win but for anomaly detection we must care about a high false negative value. For that the Precision with which we can determine that isolation forest has a high precision on detection anomaly. The most important thing is Recall percentage. It is considering anomaly detection capabilities so here in our results we can see Isolation Forest has the highest value in every performance in Recall.

H. Performance of proposed Method

In our experiment we train our model with smart meter data. The performance of the proposed method is evaluated in 4 different evaluation metrics which are accuracy, precision, recall and F1 Score. It is worth mentioning that we have randomly chosen data from our dataset with respect to 30%, 40%,50% and 60% of training ratios (Fig. 6). From around 971273 dataset (after data cleaning) our algorithm performed linearly (Fig. 5) where others fluctuate frequently. Though the accuracy is not highest with 96.65% in 40% and 60% evaluation. But rather than that, the other metric is a far better score than the other methods. With the highest 95.33% in precision, 96.65% in recall and 95.97% in F1 score. Our method has the best performance when the evaluation considering 60% training ratio (Fig. 7). The more data it will get trained the better performance we can get.

I. Performance Comparison

So, with the above equation and from table I (considering only 50% as the base for comparison) we can see that precision which is also known as predictive value. FN and TN are not included in the formula, and this may produce skewed results for our unbalanced classes. Where our algorithm gives a far

		Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Considering 30% test set	Our Model	96.61	95.25	96.61	95.92
	DTC	97.02	41.67	52.12	46.31
	RFC	98.14	98.14	61.96	62.90
	XGBOOST	98.12	66.56	48.12	55.86
Considering 40% test set	Our Model	**96.65**	95.26	**96.65**	95.94
	DTC	95.68	26.71	43.88	33.11
	RFC	97.25	45.15	57.58	50.61
	XGBOOST	97.32	45.19	43.65	44.41
Considering 50% test set	Our Model	96.61	95.25	96.61	95.92
	DTC	95.35	27.20	53.74	36.12
	RFC	96.55	36.88	57.56	44.96
	XGBOOST	96.47	33.30	43.89	37.87
Considering 60% test set	Our Model	**96.65**	**95.33**	**96.65**	**95.97**
	DTC	96.60	34.63	43.60	38.60
	RFC	97.8 3	54.97	63.51	58.93
	XGBOOST	98.14	64.68	53.52	58.57

TABLE I

RESULTS OF CONSIDERING DIFFERENT PERFORMANCE EVALUATION WITH DIFFERENT TEST SETS.

TABLE II

CONFUSION MATRIX APPLIED IN SMART METER THEFT DETECTION.

Users	Detected as a Theft User	Detected as a Normal Use
Theft users	TP (true positive)	FN (false negative)
Normal users	FP (false positive)	TN (true negative)

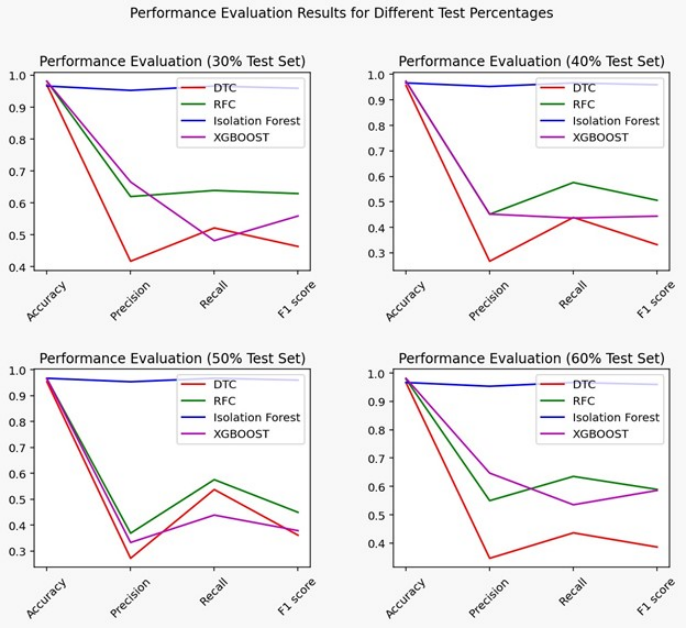

Fig. 6. Performance Evaluation results for different test percentages

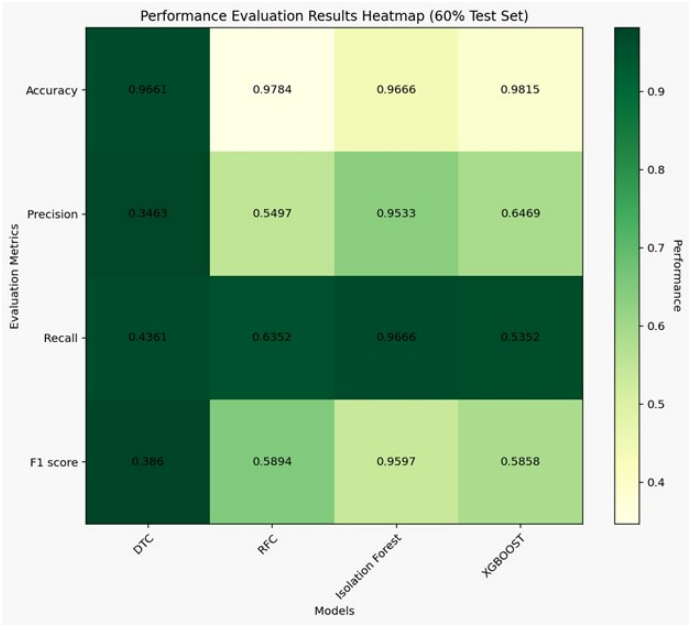

Fig. 7. Performance evaluations results heatmap of 60% Test Set

better result than the other in %. Recall is a likelihood that a fraud detection will lead to a positive test outcome (true-positive rate). This method, as we can see, does not include FP and TN; as a result, sensitivity may result in a biased outcome. The classifier has a sensitivity of 96.61% when reporting 50% of the performance evaluation data set. In comparison to the previous 2 measures, the F1 score contains both Recall and Precision and offers a more balanced perspective, although it may produce biased results in some situations since it excludes TN. But in our case the F1 score of our method for any evaluation test set is way better than the others.

CONCLUSION

Advanced metering infrastructure (AMI) is an integrated system of communication technologies that enables two-way communication between a utility and its customers' energy meters. The system uses smart meters to collect and transmit real-time data on energy consumption and allows utilities to remotely monitor and manage the flow of electricity, gas, and water. This technology enables utilities to optimize their distribution networks, reduce peak demand, and improve energy efficiency, leading to cost savings and environmental benefits. Smart meters, which are a crucial component of AMI, are digital devices that measure and transmit energy consumption data to the utility in real-time. They are important because they enable utilities to accurately track and bill customers for

their energy usage, which promotes conservation and helps to reduce energy waste. Smart meters also provide customers with access to their energy consumption data, which empowers them to make informed decisions about their energy usage and reduce their energy bills. However, with the increasing use of smart meters, security concerns have arisen. Smart meters are vulnerable to cyber attacks, which could result in unauthorized access to sensitive customer information or the disruption of energy supply. To mitigate these risks, security measures such as encryption and access controls need to be implemented. Isolation forest is a machine learning algorithm that can be used to detect anomalies in data. It works by partitioning data points into smaller and smaller subsets until anomalies can be easily identified as data points that are isolated from the rest. XGBoost is another machine learning algorithm that is commonly used for classification and regression tasks. It works by building a sequence of decision trees, each of which is trained to correct the errors of the previous tree. This iterative process allows XGBoost to accurately model complex relationships between variables, making it a powerful tool for predicting energy consumption and optimizing energy management systems. The study compared the performance of Isolation Forest with three other algorithms, DTC, RFC, and XGBOOST, using evaluation metrics such as precision, recall, and F1 score. The results showed that Isolation Forest outperformed the other algorithms in detecting anomalies with a high recall value. The study also evaluated the proposed method's performance in different evaluation metrics such as accuracy, precision, recall, and F1 score, with the best performance achieved at a 50% training ratio. The F1 score was used as a more balanced perspective, taking into account both recall and precision. The study used a confusion matrix to evaluate the performance of the algorithms. Overall, the results suggest that Isolation Forest is a better algorithm for detecting energy theft. In conclusion, AMI and smart meters are important technologies for improving energy efficiency and reducing waste. However, security measures need to be implemented to protect against cyber threats. Isolation forest can be used to detect anomalies and optimize energy management systems, leading to further cost savings and environmental benefits.

REFERENCES

[1] P. D. Halle and S. Shiyamala, "Ami and its wireless communication security aspects with qos: A review," *Advances in Smart System Technologies: Select Proceedings of ICFSST 2019*, pp. 1–13, 2021.

[2] F. Shehzad, N. Javaid, A. Almogren, A. Ahmed, S. M. Gulfam, and A. Radwan, "A robust hybrid deep learning model for detection of non-technical losses to secure smart grids," *IEEE Access*, vol. 9, pp. 128 663–128 678, 2021.

[3] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422.

[4] V. Ford, A. Siraj, and W. Eberle, "Smart grid energy fraud detection using artificial neural networks," vol. 2015, 12 2014.

[5] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of xgboost," 11 2019.

[6] MarketScreener, "Smart meter, llc - 51 million americans were affected by healthcare data breaches in 2022 due to weak security: Marketscreener," Mar 2023. [Online]. Available: https://www.marketscreener.com/quote/stock/AT-T-INC-14324/news/Smart-Meter-LLC-51-Million-Americans-Were-Affected-by-Healthcare-Data-Breaches-in-2022-Due-to-Wea-43312174/

[7] M. Raciti and S. Nadjm-Tehrani, "Embedded cyber-physical anomaly detection in smart meters," in *Critical Information Infrastructures Security: 7th International Workshop, CRITIS 2012, Lillehammer, Norway, September 17-18, 2012, Revised Selected Papers*. Springer, 2013, pp. 34–45.

[8] G. Kalogridis, M. Sooriyabandara, Z. Fan, and M. A. Mustafa, "Toward unified security and privacy protection for smart meter networks," *IEEE Systems Journal*, vol. 8, no. 2, pp. 641–654, 2013.

[9] R. d. T. Caropreso, R. A. Fernandes, D. P. Osorio, and I. N. Silva, "An open-source framework for smart meters: Data communication and security traffic analysis," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 2, pp. 1638–1647, 2018.

[10] P. Kumar, Y. Lin, G. Bai, A. Paverd, J. S. Dong, and A. Martin, "Smart grid metering networks: A survey on security, privacy and open research issues," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2886–2927, 2019.

[11] A. M. Khattak, S. I. Khanji, and W. A. Khan, "Smart meter security: Vulnerabilities, threat impacts, and countermeasures," in *Proceedings of the 13th International Conference on Ubiquitous Information Management and Communication (IMCOM) 2019 13*. Springer, 2019, pp. 554–562.

[12] R. Marah, I. El Gabassi, S. Larioui, and H. Yatimi, "Security of smart grid management of smart meter protection," in *2020 1st international conference on innovative research in applied science, engineering and technology (IRASET)*. IEEE, 2020, pp. 1–5.

[13] S. Aziz, S. Z. H. Naqvi, M. U. Khan, and T. Aslam, "Electricity theft detection using empirical mode decomposition and k-nearest neighbors," in *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*. IEEE, 2020, pp. 1–5.

[14] P. Prabhakar, S. Arora, A. Khosla, R. K. Beniwal, M. N. Arthur, J. L. Arias-Gonzáles, F. O. Areche *et al.*, "Cyber security of smart metering infrastructure using median absolute deviation methodology," *Security and Communication Networks*, vol. 2022, 2022.

[15] M. S. Abdalzaher, M. M. Fouda, and M. I. Ibrahem, "Data privacy preservation and security in smart metering systems," *Energies*, vol. 15, no. 19, p. 7419, 2022.

[16] L. Polčák, "Responsible and safe home metering: How to design a privacy-friendly metering system," in *Information Security and Privacy in Smart Devices: Tools, Methods, and Applications*. IGI Global, 2023, pp. 1–40.

[17] "Smart meter project." [Online]. Available: https://www.tepco.co.jp/en/pg/development/domestic/smartmeter-e.html

[18] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422.

[19] "1.10. decision trees." [Online]. Available: https://scikit-learn.org/stable/modules/tree.html

[20] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[21] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[22] S. S. S. R. Depuru, L. Wang, V. Devabhaktuni, and P. Nelapati, "A hybrid neural network model and encoding technique for enhanced classification of energy consumption data," in *2011 IEEE power and energy society general meeting*. IEEE, 2011, pp. 1–8.

[23] P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity theft detection in ami using customers' consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, 2015.

[24] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*. Springer, 2005, pp. 345–359.