

Machine Learning and Explainable AI for Robust Smart Grid Security: A New Dataset Study

Md Tanjil Islam Aronno¹, Md. Sabbir Hossain Molla²,
Md Tanvir Rahman Sahed³, Md Nasif Mahmud Zishan⁴

¹Department of Electrical and Electronic Engineering, Bangladesh University of Engineering & Technology
Dhaka, Bangladesh

^{2,3,4}Department of Electrical, Electronic & Communication Engineering, Military Institute of Science & Technology
Dhaka, Bangladesh

¹aronnondc@gmail.com, ²sabbir.eece@gmail.com, ³tanvir.sahed00@gmail.com, ⁴nasif.jisan@gmail.com

Abstract—A smart grid is the combination of networking equipment to capture parameters related to the cyber domain, as well as Advanced Metering Infrastructure (AMI) to obtain necessary consumption data and physical domain parameters. However, today's cyber-physical grid systems are exposed to numerous threats, which have increased in frequency due to the integration of low-cost networking equipment and Internet of Things (IoT) devices. If we know that any attack has occurred in the grid, the operators can disconnect the utilities directly, and knowing the type of cyberattack will help the operators investigate the reason and solve it immediately. In this paper, we have done an analysis of machine learning classifiers to detect an attack case in the smart grid, as well as identify the type of cyberattack using the Smart Grid Cyber-Physical Attack dataset to help the operators disconnect the utilities and to investigate. It has been found that Random Forest achieves the highest accuracy of 98% while detecting the attack scenario, whereas LightGBM achieves an accuracy of 99% while identifying the type of attack. This result has been validated using 5-fold cross-validation, which proves the robustness of Random Forest and LightGBM in performing the task.

Index Terms—Smart Grid, Cyber-Physical Attack, Machine Learning, Random Forest, 5 Fold Cross-validation, LightGBM

I. INTRODUCTION

With the course of time, the world is adopting smart grid (SG) technology, as it offers various functionalities to energy users as well as to regulatory authorities. Moreover, integration of conventional nonrenewable sources with renewable sources ensures bidirectional generation and demand response. An SG incorporates various advanced electronic devices, including smart meters in the AMI, to deliver perfect energy consumption information and enable demand response at the consumer level. At the transmission and distribution levels, various types of devices are utilized to gather synchronized data, automate protection mechanisms, and perform localized control operations. Additionally, IoT-enabled devices contribute to enhanced situational awareness by providing crucial grid-related insights.

However, this integration of IoT and networking technologies has introduced vulnerabilities, exposing the entire grid system to potential cyberattacks. The primary cybersecurity risks that allow attackers to infiltrate devices or control networks include internet access, misconfigured firewalls, un-

secured wireless routers, and wired modems. Attacks can impact both physical and cyber domain parameters. Physical domain parameters include real and reactive power, three-phase voltages and currents, system frequency, and others. Anomalies in these values may indicate the existence of an attack.

Due to the tendency of SG systems to be targeted for security breaches, it is important for national energy regulatory bodies to implement proactive defense measures and take swift actions aimed at detecting and countering possible threats to SG systems. These threats include both physical and cyberattacks, such as backdoor, brute force, false data injection (FDI), ransomware, reverse shell, and others. These attacks can occur either as a single attack or as multiple attacks simultaneously. For this reason, there has long been a scarcity of databases containing data on various attack types. Most of the available databases contained only FDI data [1]. The recently published Smart Grid Cyber-Physical Attack Dataset [2] is a vital contribution to the research community, as it incorporates various features of data for the aforementioned attacks. So, we decided to work with this dataset, and we have:

(i) Pre-processed the dataset, taking the benign class and the attack classes comprising the five attack types. Then applied machine learning (ML) models with 5 fold cross-validation analysis to determine the performance of the models and their robustness. The result will help the operators identify an attack scenario and take immediate steps.

(ii) Declared classes for the various attack types (benign-0, backdoor-1, brute force-2, etc.). Then performed correlation heatmap analysis and principal component analysis to determine the most significant parameters of the dataset that contribute to detecting the various attack classes. In fact, this part of our work can be seen as a multiclass detection so that the operators can directly investigate the attack type and act immediately.

II. LITERATURE REVIEW

In recent times, the protection of SG networks from various cyber and physical domain attacks has become a much-discussed topic. Sridhar et al. [3] emphasized the need for

securing smart grids against cyber threats that can disrupt operations. A notable real-world example is the 2015 Ukraine cyberattack [4], initially mistaken for a system fault but later identified as a coordinated attack causing a nationwide blackout. Since then, attack types such as FDI, brute force, ransomware, backdoors, and reverse shells have been documented, among which FDI is the most prevalent.

SG systems continuously monitor generation and distribution parameters to enable responsive control. Various studies have examined how cyberattacks impact these physical signals. Linertz et al. [5] analyzed effects on grid frequency and voltage, while Bragatto et al. [6] and Liang et al. [7] focused on FDI's influence on voltage regulation and system stability. Real-world incidents like the Ukraine breach showed how backdoors and reverse shells were exploited to disrupt breaker operations, methods also found in ransomware attacks.

In terms of detection, Ten et al. [8] developed early intrusion detection based on cyber-domain event timing. PMU-based techniques that monitor physical parameters like current and voltage have also been explored, such as hierarchical temporal memory (HTM) learning [9], though it lacks anomaly classification capabilities.

ML is increasingly adopted in grid monitoring [10], yet most methods focus narrowly on FDI attacks [11]. Recent surveys [12] highlight growing interest in learning-based detection using physical-domain metrics. Classic ML classifiers like SVM, kNN, and ensemble models are widely used. For instance, Himeur et al. [13] improved kNN for abnormal consumption detection, while Rajalakshmi et al. [14] used isolation forests to detect anomalies in current, voltage, and power.

But most of these approaches were focused on binary classification of FDI injection. Alani et al. [15] proposed a two-stage attack classification scheme: stage 1 detects anomalies, and stage 2 classifies them into multiple attack categories like FDI, DoS, replay, and jamming using machine learning classifiers. But these attacks were focused on cyber domain parameters. Therefore, a classification of various attacks pertaining to the physical domain of smart grids will bridge the gap and contribute to taking necessary precautions by integrating the best classifier into AMIs.

III. METHODOLOGY

The overall workflow consists of dataset preprocessing and normalization, followed by the application of ML classifiers. The complete workflow is illustrated in Figure 1.

A. Dataset Description

The Smart Grid Cyber-Physical Attack Dataset gives numerical data of cyber and physical measurements under multiple scenarios, including normal operation and various cyberattacks such as backdoor, brute force, FDI, ransomware, and reverse shell attacks. Each scenario contains paired cyber and physical data files, allowing joint analysis across domains. Table I gives a clear notion of the data types included in the dataset.

TABLE I: Smart Grid Cyber-Physical Attack Dataset

Scenario	Description	Class
Benign	Normal grid operation	0
Backdoor	Hidden entry point exploited	1
Brute Force	Repeated login attempts	2
FDI	Tampering with sensor data	3
Ransomware	Data locked by encryption	4
Reverse Shell	Remote control session	5

B. Dataset Pre-processing

1) *Preprocessing for Binary Classification:* For the purpose of binary classification of attack, all attack categories were consolidated into a single class (Attack = 1), while benign instances were retained as label 0. To achieve balanced class representation, benign samples (99,900) were matched with an equal number of attack samples by randomly down-sampling 19,980 instances from each attack category. This process yielded a final dataset of 199,800 records, evenly distributed between benign and attack classes. This dataset pre-processing stage ensures a higher classification accuracy. Figure 1 illustrates the overall preprocessing of the dataset along with the distribution of data between the benign and attack classes.

2) *Preprocessing for Multiclass Classification:* Over the same dataset, we did the class assignment task so that the machine learning model could detect various classes.

C. Dataset Analysis

1) *Correlation Analysis:* We calculated Pearson correlation coefficients in order to assess feature relevance. ActivePower and ReactivePower had the highest correlation with attack classes, as shown in figure 2, as expected by their high discriminative power. Phase voltages and phase currents had significant intercorrelation, suggesting that they contain redundant information removed later by PCA. The correlation heatmap highlights that ActivePower and ReactivePower show the strongest correlation with the class label, indicating their high discriminative power.

2) *Feature Normalization:* The Min-Max normalization was used in normalizing the numerical features in the dataset. The transformation is given by the equation 1.

$$x'_{i,j} = \frac{x_{i,j} - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (1)$$

Where, $\min(x_i)$ and $\max(x_i)$ denote the minimum values and maximum values of feature i , respectively. This scaling constrains all numerical variables to the range [0,1].

3) *Feature Distribution Analysis:* To understand feature-level discriminability, we visualized the normalized distributions of the top correlated features across eight classes, shown in Figure 3. The x-axis represents normalized feature values obtained through Min-Max scaling, while the y-axis indicates the probability density. Instead of raw counts, it rescales the histogram, which gives a probability density estimate that allows comparison across features and classes.

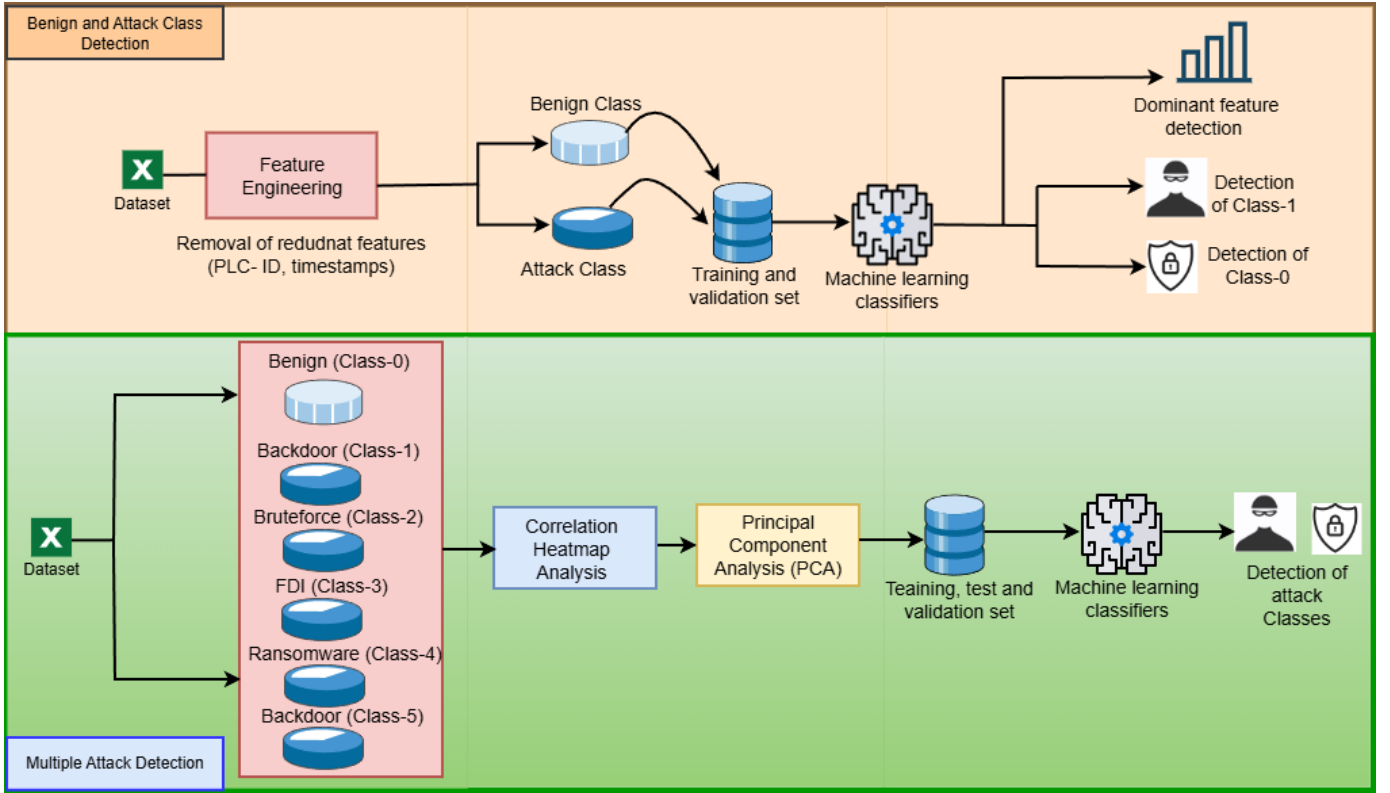


Fig. 1: End-to-end workflow of the cyber-physical attack detection task, from dataset preparation to model evaluation.

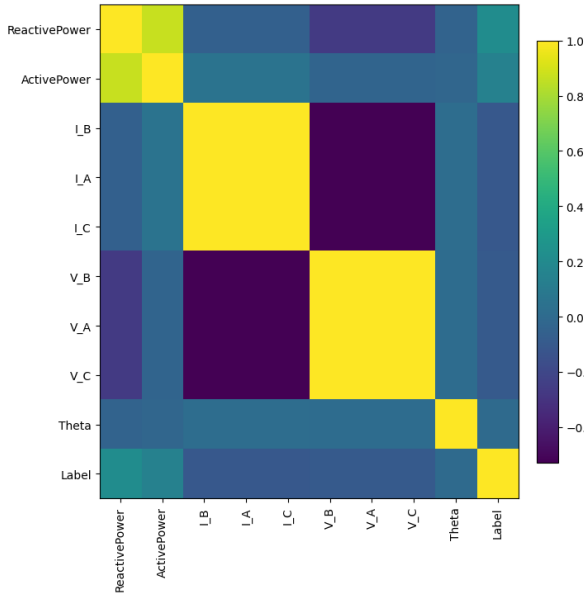


Fig. 2: Correlation Heatmap

D. Training Procedures

1) *Train, Test, and Validation set Selection:* The dataset was divided into training, validation, and testing sets with an 80%, 10%, and 10% proportion, respectively, using Python's Scikit-learn's `train_test_split` function.

2) *Batch Size Selection Rationale:* A batch size of 2048 was used during training to ensure stable gradient updates and efficient utilization of GPU resources. Instead of feeding all the samples at once, the data were processed in mini-batches, which balances memory efficiency with convergence stability.

E. 5-fold Cross-Validation

To ensure reliable evaluation and minimize bias from dataset partitioning, a stratified 5-fold cross-validation strategy was adopted. Here, the training dataset was divided into 5 equal subsets, with 4 folds used for model training and the remaining fold reserved for validation in each iteration. The performance metrics obtained from all folds were then averaged to provide a stable estimate of the model's generalization ability. This was implemented using the `StratifiedKfold` class from the scikit-learn library in Python, which preserves the class distribution across folds, thereby ensuring balanced validation splits for multi-class attack detection.

F. Machine Learning-Based Classification Methods

Several ML classifiers were implemented to compare their effectiveness in detecting cyberattacks from grid data.

1) *Random Forest:* It integrates the predictions of various decision trees via majority voting. Averaging the results of the trees reduces overfitting and achieves strong generalization. Its robustness and ability to capture nonlinear relationships made it the best-performing model in this study.

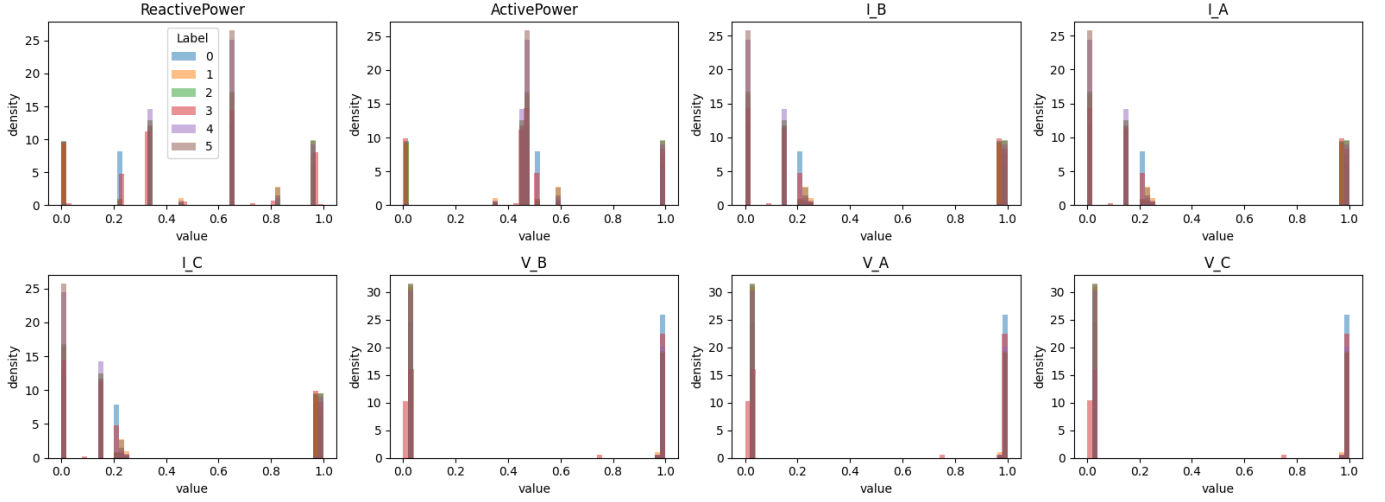


Fig. 3: Histograms split by label for top-discriminative features

2) *XGBoost*: It is a gradient boosting framework that constructs trees iteratively, where each subsequent tree refines the previous ones' errors. It is used for being very accurate in prediction, especially during the processing of structured tabular data.

3) *LightGBM*: It is also a boosting algorithm that is optimized for speed and scalability. Employing histogram-based feature evaluation and leaf-wise tree growth, it is very fast on big datasets while keeping competitive accuracy.

4) *k-Nearest Neighbors*: It is a local-based method that gives the class labels based on the majority vote among the nearest neighbors. It is an easy and local data structure-based, so it is appropriate to identify attacks that gather around some local areas of the feature space.

G. Evaluation Metrics

1) *Accuracy*: It refers to the total proportion of properly classified cases across all samples. It is measured by equation 2.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where TP , FP , TN , and FN are true positives, false positives, true negatives, and false negatives, respectively.

2) *Precision*: It indicates the reliability of positive predictions, i.e., the proportion of predicted attacks that are truly attacks. It can be measured by equation 3.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

3) *Recall*: It indicates the model's ability to properly recognize actual attacks. It is calculated by equation 4.

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

4) *F1-score*: It is the harmonic mean of precision (P) and recall (R), which balances the risk of false positives and negatives. It is measured by equation 5.

$$F1 - score = 2 \times \frac{P \times R}{P + R} \quad (5)$$

5) *ROC-AUC*: It compresses this curve into one number, where larger numbers reflect the stronger discriminatory power.

IV. RESULTS AND DISCUSSION

A. Overall Model Performance

Applying different classifiers, we observed that, for the binary classification, RF has shown the best classification, with an accuracy of 0.98, a precision of 0.99, a recall of 0.97, an F1-score of 0.98, and a 0.99 ROC-AUC, as supported by the confusion matrix shown in Figure 4. As the dataset was cleaned and the attack classes and benign class were equal in number, these models have shown a perfect classification. The 5 fold cross-validation also validates this outstanding performance of RF. Table II shows the performance metrics of classifiers employed to detect attack scenarios in the smart grid cyber-physical attack dataset. Regarding the multiple attack

TABLE II: Performance comparison of Binary classification

Models	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.61	0.59	0.73	0.65	0.67
SVM	0.62	0.60	0.74	0.66	0.67
kNN	0.95	0.95	0.94	0.95	0.97
XGBoost	0.96	0.96	0.95	0.95	0.98
LightGBM	0.96	0.98	0.94	0.96	0.98
Random Forest	0.98	0.99	0.97	0.98	0.99

classification, we saw that LightGBM has achieved the highest accuracy of 0.99, with similar precision, recall, and F1-score. This proves that LightGBM is capable of performing the multiclass detection task in an almost perfect manner. Table

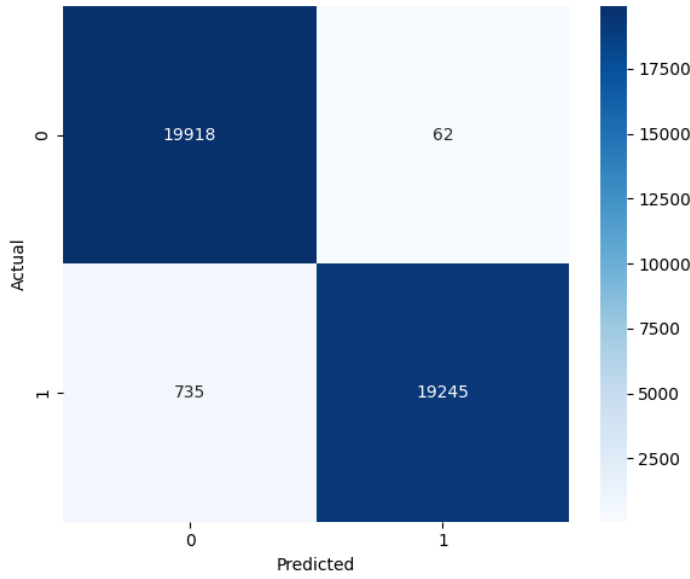


Fig. 4: Confusion matrix of the random forest

TABLE III: Performance comparison of Multi-class classification

Models	Accuracy	Precision	Recall	F1-Score	ROC-AUC
kNN	0.93	0.92	0.93	0.93	0.97
CatBoost	0.98	0.97	0.97	0.97	1.00
XGBoost	0.98	0.98	0.98	0.98	1.00
LightGBM	0.99	0.99	0.99	0.99	1.00

III shows the evaluation metrics of the four classifiers. These results are also validated by the confusion matrix as shown in Figure 5.

B. ROC-AUC Curve Analysis

The combined ROC–AUC analysis of the attack and benign classification, as shown in Figure 6, shows discrimination ($AUC = 0.999$), followed closely by XGBoost (0.987) and LightGBM (0.982). kNN also performed reasonably (0.976), while logistic regression (0.672) and SVM (0.671) showed weak separation. While in the case of the multiple attack detection, LightGBM showed an ROC-AUC of 1.00, which highlights the strong discriminative capacity of boosting-based classifiers.

C. Feature Importance Analysis

Figure 7 shows the feature importance ranking from the LightGBM model. Reactive power and active power are the most influential features, while phase currents (I_A , I_B , I_C) provide a moderate contribution. In contrast, voltages (V_A , V_B , V_C) and the phase angle (Θ) have a lower impact. This indicates that power-related attributes carry the strongest discriminative information for separating benign and attack classes.

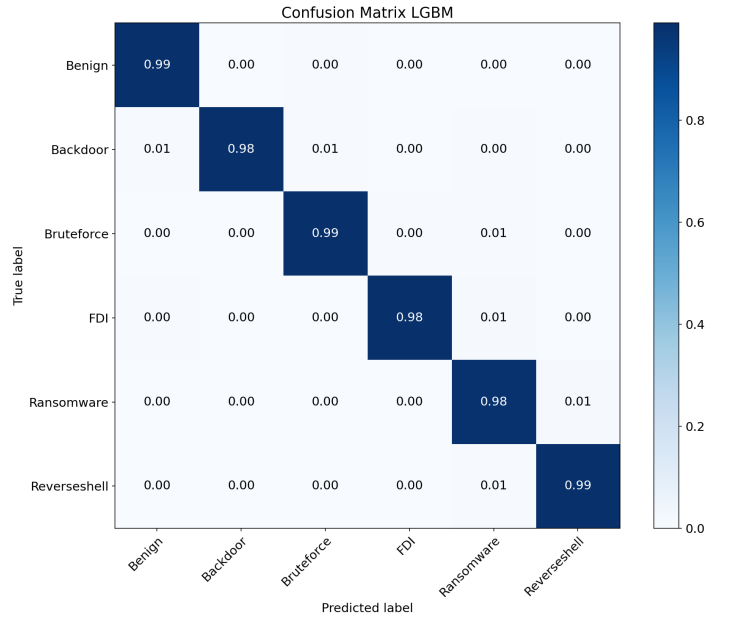


Fig. 5: Confusion matrix for multiclass classification

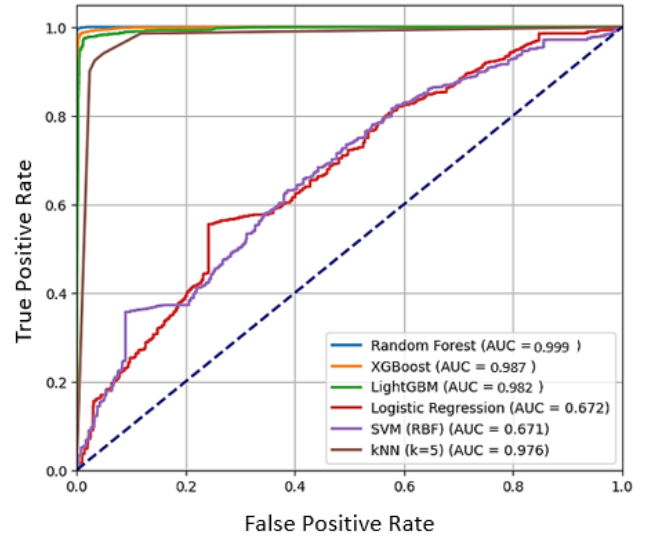


Fig. 6: Combined ROC-AUC curve

D. Generalization and Cross-Validation Analysis

To determine how well the top-performing RF model and LightGBM generalize and to look for potential overfitting, a learning curve was graphed, as in Figure 8. When the training data size is small, validation accuracy is low compared to training accuracy. When it is large enough, the two accuracies improve and converge to each other, and the space between the two curves is repeatedly small (less than 0.5%). This convergence indicates how well it generalizes through unseen data without being an overfitting model. A similar pattern was seen during the multi-class classification scenario as well.

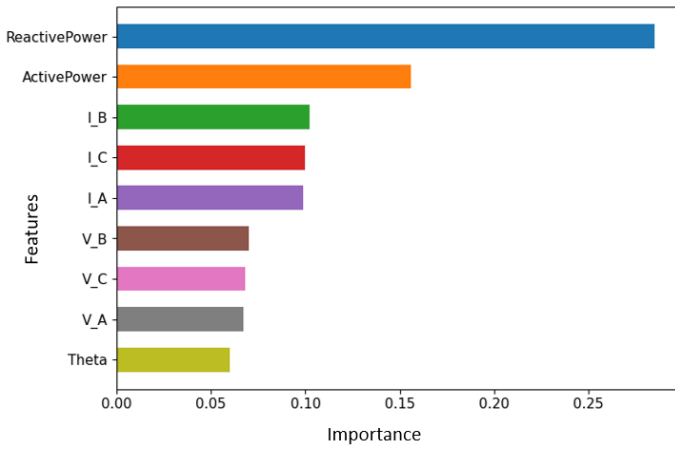


Fig. 7: Feature importance analysis

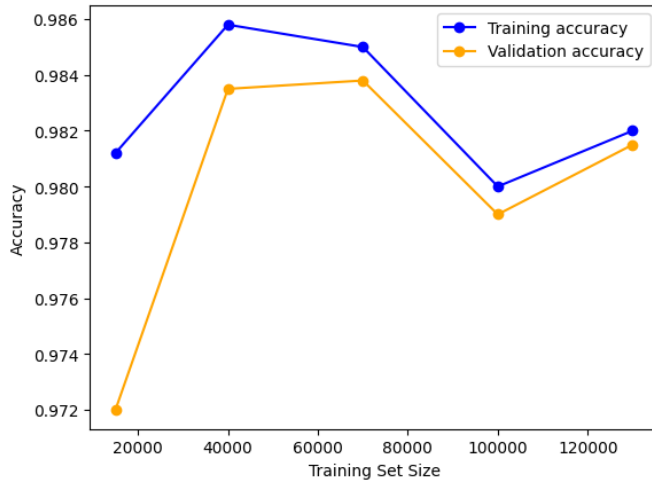


Fig. 8: Learning curve

E. Discussion of Findings

Model performance metrics show that RF, XGBoost, LightGBM, and kNN effectively detected attacks, with RF outperforming others due to its ability to capture nonlinear patterns. While for the multiple attack detection scenario, LightGBM performs the best. This supports the use of RF and LightGBM in AMI systems to build reliable detection mechanisms for various physical domain attacks in a smart grid.

Feature importance analysis revealed that reactive power is the most critical indicator of attacks, likely because it's easily manipulated through power factor correction devices.

V. CONCLUSION AND FUTURE WORKS

Our work has presented an evaluation of a recently published dataset comprising five categories of cyberattack data, both in the physical and cyber domains. The RF model achieves an accuracy of 98%, an F1-score of 98%, and 99% ROC-AUC, whereas LightGBM showed 99% accuracy and other metrics while identifying various types of attacks. The high metrics prove the high quality and less imbalance of the

training set. Analyzing the most contributing features to the attack detection, we found that reactive power contributed the most. This result indicates that reactive power values are very easy to manipulate by adjusting power factor improvement devices.

As this dataset was recently released and combines five attack categories in both domains, there is a significant opportunity to expand this research work in some other directions. As an example, future research could be devoted to anomaly detection exclusively within the cyber domain as an extension of the physical domain analysis. Some recent advances, including graph neural networks, can be employed over this dataset to see how they perform.

REFERENCES

- [1] Xuefei Yin. Datasets for Stealthy False Data Injection Attacks (SFDIA) against Smart Grid, 2024.
- [2] Jacob Sweeten, Amr Elshazly, Abdulrahman Takiddin, Muhammad Ismail, Shady Refaat, and Rachad Atat. Smartgrid-cyberphysical-attack-dataset, 2025.
- [3] Siddharth Sridhar, Adam Hahn, and Manimaran Govindarasu. Cyber-physical system security for the electric power grid. *Proceedings of the IEEE*, 100(1):210–224, 2011.
- [4] Cybersecurity & Infrastructure Security Agency (CISA). Cyber-attack against ukrainian critical infrastructure. ICS Alert IR-ALERT-H-16-056-01, July 2021. Last revised July 20, 2021.
- [5] Philipp Linnartz, Alexander Winkens, and Andreas Ulbig. Assessing the impact of cyber attacks manipulating distributed energy resources on power system operation. *arXiv preprint arXiv:2207.07968*, 2022.
- [6] Tommaso Bragatto, Marco Antonio Bucarelli, Maria Sofia Bucarelli, Federico Carere, Alberto Geri, and Marco Maccioni. False data injection impact on high res power systems with centralized voltage regulation architecture. *Sensors*, 23(5):2557, 2023.
- [7] Jingwen Liang, Lalitha Sankar, and Oliver Kosut. Vulnerability analysis and consequences of false data injection attack on power system state estimation. *IEEE Transactions on Power Systems*, 31(5):3864–3872, 2015.
- [8] Chee-Wooi Ten, Junho Hong, and Chen-Ching Liu. Anomaly detection for cybersecurity of the substations. *IEEE Transactions on Smart Grid*, 2(4):865–873, 2011.
- [9] Anomadarshi Barua, Deepan Muthirayan, Pramod P Khargonekar, and Mohammad Abdullah Al Faruque. Hierarchical temporal memory-based one-pass learning for real-time anomaly detection and simultaneous data prediction in smart grids. *IEEE Transactions on Dependable and Secure Computing*, 19(3):1770–1782, 2020.
- [10] MD Jainul Abudin, Surmila Thokchom, RT Naayagi, and Gayadhar Panda. Detecting false data injection attacks using machine learning-based approaches for smart grid networks. *Applied Sciences*, 14(11):4764, 2024.
- [11] Marcelo Fabian Guato Burgos, Jorge Morato, and Fernanda Paulina Vizcaino Imacaña. A review of smart grid anomaly detection approaches pertaining to artificial intelligence. *Applied Sciences*, 14(3):1194, 2024.
- [12] Jiuqi Elise Zhang, Di Wu, and Benoit Boulet. Time series anomaly detection for smart grids: A survey. In *2021 IEEE electrical power and energy conference (EPEC)*, pages 125–130. IEEE, 2021.
- [13] Yassine Himeur, Abdullah Alsalemi, Faycal Bensaali, and Abbes Amira. Smart power consumption abnormality detection in buildings using micromoments and improved k-nearest neighbors. *International Journal of Intelligent Systems*, 36(6):2865–2894, 2021.
- [14] D Rajalakshmi, K Sudharson, K Rajesh Kambattan, A Suresh Kumar, R Arulkumar, and MA Starlin. Enhancing real-time fault detection in electrical grids using hybrid ensembleboost over wireless networks. *SSRG International Journal of Electronics and Communication Engineering*, 11(4):24–33, 2024.
- [15] Sahar Yousif Mohammed, Mohammad Aljanabi, Maad M Mijwil, Ali J Ramadhan, Mostafa Abotaleb, Hussein Alkattan, and Zainalabideen Albadran. A two-stage hybrid approach for phishing attack detection using url and content analysis in iot. In *BIO Web of Conferences*, volume 97, page 00059. EDP Sciences, 2024.