International Research Journal of Mathematics, Engineering and IT
ISSN: (2349-0322)
Association of Academic Researchers and Faculties (AARF)
Impact Factor- 8.132 Volume 11, Issue 08, Aug 2024
Website- www.aarf.asia, Email : editoraarf@gmail.com

ANOMALY DETECTION IN SMART GRIDS: A MACHINE LEARNING PERSPECTIVE ON ELECTRICITY THEFT PREVENTION

Shakeel Ahmad Najar[1], Dr. Ajit Kumar[2]
Research Scholar, Computer Science, Shri JJT University, Rajasthan, India [1]
Professor, Computer Science, Shri JJT University, Rajasthan, India [1]

Abstract: *Data from smart grids can be examined to find anomalies in a variety of fields, including cybersecurity, fault finding, energy theft, etc. There is a compelling case to be made for anomaly detection using machine learning. We need to extract features from the raw grid data. Any occurrences or modifications in the smart grid data that deviate from the typical pattern are referred to as anomalies. The outcomes of a common grid layout can differ greatly based on patterns or modifications in power, voltage, current, or consumption. In this research, an anomaly detection model is developed for a hardware-based testbed implementation of a real-world smart grid system. It is possible to enhance system behaviour in data communication flow by identifying anomalous activity. Additionally, it will detect any changes in parameters that can point to the existence of cyberattacks. Our suggested anomaly detection methodology uses several decision trees to separate outliers from typical observations, based on the Isolation Forest (IF) paradigm. The simulation findings were used on a hardware-based testbed to validate the effectiveness of the suggested detection strategy. Principal component analysis was utilised to optimise feature selection, and the dickey-fuller test was employed to assess the model's performance.*

Keywords—Machine Learning, Anomaly detection, Smart Grid, PLC, SCADA, Sensor Data, Power Systems.

I. INTRODUCTION

By combining conventional power data and communication technology, smart metres send real-time data from end users to the smart grid. Because of their intricate architecture and the sensitive nature of the data they convey, smart grids require cybersecurity. The cybersecurity section systematically monitors the data flow from each grid parameter for irregularities. We shall talk about the different kinds of cyberattacks that can affect the grid in this study. After doing a thorough examination of the data and taking into account the underlying attacks, we will create a machine learning algorithm that uses individual data production and data flow points to identify anomalies.

In the electrical industry, compromised smart grid devices are one of the biggest security risks. False information from a sensor could lead a control device to increase voltage, overloading the grid. There could be security flaws in smart equipment that measure and regulate the grid, like sensors and controllers. Malicious behaviour on a control device can do the same thing, stopping the flow of electricity. The grid's smart grid components have to function as intended and provide a steady supply of this vital resource. This research attempts to analyse smart grids from a data-driven standpoint by examining real-world data streams from a hardware-based testbed accessible within the Energy Systems Research Laboratory at Florida International University. This study develops a machine learning model that identifies anomalies in data and marks them for examination.

The machine learning model in this study, which is built on isolation forests, makes use of historical data to forecast grid parameter values in the future. Additionally, it keeps an eye on grid activities in case an attacker manages to breach the system or if another problem modifies the grid's specifications. To test if the built model can identify system changes, a few fictitious attacks were introduced. This approach allows for the detection of attack locations with a 98.74% accuracy rate by allowing anomalous behaviour at each node to be observed.

The remainder of the document is organised as follows: The relevant work is presented in Section II. The architecture, philosophy, procedures, and specifics of the machine learning model and anomaly detection for the smart grid testbed are presented in Section III. The graphs and findings are shown in Section IV, along with an explanation of the anomaly point that the machine learning model identified. The paper is finally concluded in Section V, which also addresses future work.

II. CONSISTING WORK

The critical infrastructure organisations are now vulnerable to a range of threats, such as physical assaults, virus releases, and information theft, due to their rising reliance on CyberPhysical Systems (CPS). Machine learning-based techniques can be used for a variety of security-related tasks, such as virus detection, access control, anomaly detection, intrusion detection, and classification.

The anomaly detection method is used in this work to improve security. Agrawal et al. claim that an evaluation of current data mining techniques has taken place. Here is an illustration of a malicious voltage control computation using a neural network model. A rule-based detection method is used by academics to identify attacks on smart grids.

Distinctiveness in relation to current practices:

While anomalies can be found in smart grids, this is mostly useful in the corporate and industrial domains. This work investigates anomaly detection during the generation stage. In our model, an anomaly is found at each grid point. We go over data engineering in other parts. This method shortens the time it takes to locate an attack in the grid in addition to cutting down on action time. Python machine learning is used for the implementation.

III. STEPS ADOPTED, ARCHITECTURE, MODEL, AND IDEOLOGY

Cooperation between electricity suppliers and distributors is facilitated by the smart grid. To find irregularities and provide an extra degree of protection, a machine learning system has been put into place. Energy is created in a smart grid by utilising cutting-edge technology, cutting carbon dioxide emissions, and optimising power for effective resources through the application of new technologies. Supervisory control and data acquisition, or SCADA, is a crucial part of a high-level management system that monitors and controls electricity grids. It automates and controls these systems. They have a variety of auxiliary devices and controllers that allow them to communicate over the Internet. PLCs, RTUs, sensors, metres, embedded computers, Intelligent Electronic Devices (IEDs), HMIs, and Remote Terminal Units (RTUs) are just a few examples of the many devices that fall under the category of control and peripheral devices. The smart grid is constructed by connecting these devices (fig (1)). The difficulty of creating a smart grid is clarified.

Figure 1: Smart Grid Testbed in Energy Systems Research Laboratory

A.Anomaly Detection: what is it?

A phenomena that emerges unexpectedly in a data set and deviates from the expected data is called an anomaly. Unsupervised anomaly detection, sometimes referred to as anomaly detection, is frequently applied to data that lacks a label. A few presumptions form its foundation.

- The frequency at which anomalies occur is low.
- Anomalies differ from normal data in a significant and recognizable way.

B.Smart grids need to detect anomalies, but why?

A smart grid can be exposed to the following types of attacks and machine learning can help mitigate them:

- **Device attack:** There will be an effort to take control of or compromise a device. This involves committing malevolent physical acts, data theft, and network attacks against the smart grid (if the hacked devices may operate as control elements). By disrupting a circuit, compromised IEDs, like circuit breakers, have the potential to intentionally cause power outages. Device attacks must be avoided by implementing access control.
- **Data attack:** an attempt to manipulate commands or add, remove, or modify data in network traffic in order to influence the smart grid's decision-making or behaviour. Customers frequently compromise their smart metres in order to lower their electricity bills. •**Privacy attack:** examines data on electricity use to find out or deduce personal information about consumers. In order to increase the efficiency of grid operation and gather precise information about the state of the grid, smart metres gather data on power use multiple times an hour. Data with such delicate privacy features needs to be shielded from unwanted access.

- **Network availability attack:** seems as a DoS (denial of service). As a result, when they are exhausted or overloaded, data communications to the smart grid slow down or stop altogether. An attacker with malevolent intent may constantly bombard a control centre with false information, causing it to spend most of its effort confirming the information's veracity. As a result, the control centre is unable to respond promptly to valid traffic. In smart grid control and communications, time is of the essence. Effectively handling network availability assaults is crucial.
- A few more complex threats that might exist in smart grids include:
- **Attacks involving data integrity:** Manipulate the data so the signals indicate spurious values, which could either force the actuator to make incorrect changes to the device or force the control center to make wrong decisions
- **Denial of Service (DoS) Attacks:** Delay control actions. As a result of a DoS attack on a physical system whose corrective control is time-constrained, the entire system could become unstable.
- **Replay attacks:** Retransmit legitimate control packets, incorrect decisions can be made as a result. Having a networked power-controlled system impacted by this kind of attack is problematic.
- **Timing attacks:** A type of DoS attack. Adversaries will delay the transmission of signals instead of completely cutting off communication between the system and control. The Controller will be affected by this delay, and it could even cause the controlled system to become unstable.
- **Desynchronization Attacks:** This type of attack targets controls that are difficult to synchronize.
- **Sniffing attacks:** such attacks can expose sensitive information about users and the internal operation of power companies.
- **Reconfigure attack:** This involves installing malicious firmware on Smart Grid devices and using the firmware to perform different kinds of attacks.

C.Isolation Forest

Isolation forests focus on separating anomalies in order to identify them, as opposed to standard profiling techniques that isolate normal points. Since anomalies are distinct and uncommon data points, outlier detection can be accomplished by applying the isolation forest algorithm, which yields the anomaly score for each sample. Using the Isolation Forest technique, Isolation Forest provides the anomalous score for every sample. For additional comparison, this anomaly score is returned to the data as a new column.

The process of building an isolation tree involves the following cases:

- Select a random subset of the data.
- Until every point in the dataset is isolated select one feature at a time and partition the feature at a random point in its range.

The prediction process involves:

- For Each I-tree in the forest Perform binary search for the new point across the Itree, traversing till a leaf and compute an anomaly score based on the depth of the path to the leaf.
- Aggregate the anomaly score obtained from the individual I-trees to come up with an overall anomaly score for the point.

Separation By first selecting a feature at random and then selecting a split value between the maximum and minimum values for that feature, Forest recursively creates partitions on the dataset. In a tree, an anomaly's route length—which is the number of edges travelled from the root node—will be shorter. In Binary Search Tree (BST), which resembles an I-tree in structure, the path length of a failed search is computed as:

$$C(N) = 2H(N-1) - (2(N-1)/N) \quad (1)$$

where H(i) is the harmonic number and it can be estimated by ln(i) + 0.5772156649 (Euler's constant). n is the number of instances in the dataset. This is best explained in.

D.Architecture and data

Our lab has created a remote control for the SCADA system using a local IP address. We can retrieve data from the grid at the desired spots with the aid of the SCADA control. Figure 2 displays the architecture of the testbed. The other three generators are auxiliary generators, while generator G1 is a slack generator. All four generators in the architecture follow the same trend of data flow, which is from G1 to L1 (load). In order to operate synchronously, generators are connected via transmission lines and a three-phase bus.

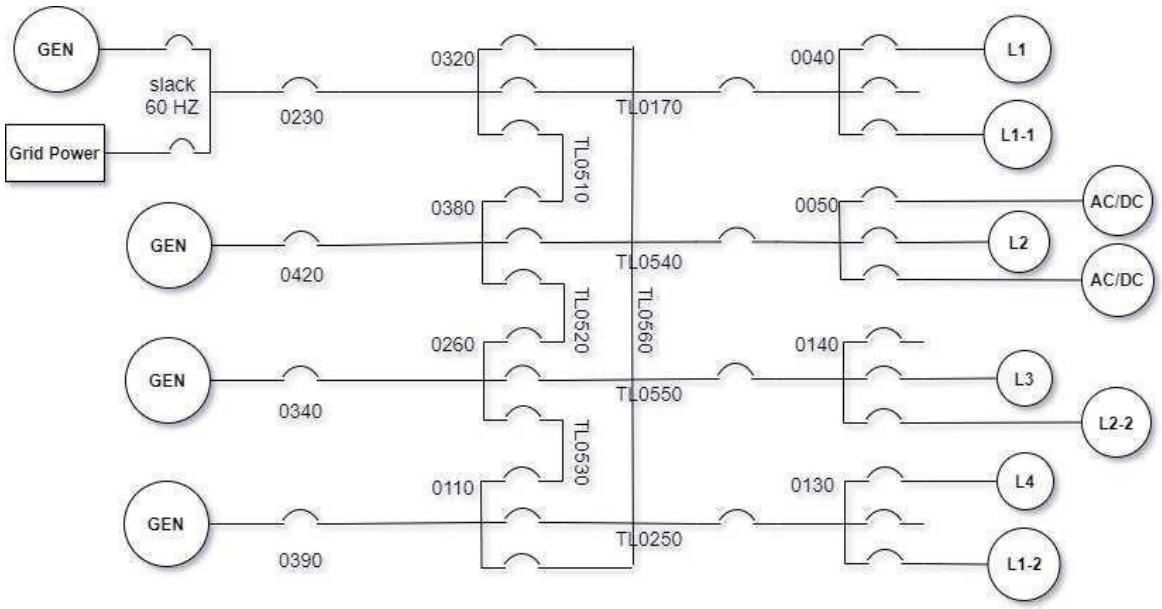
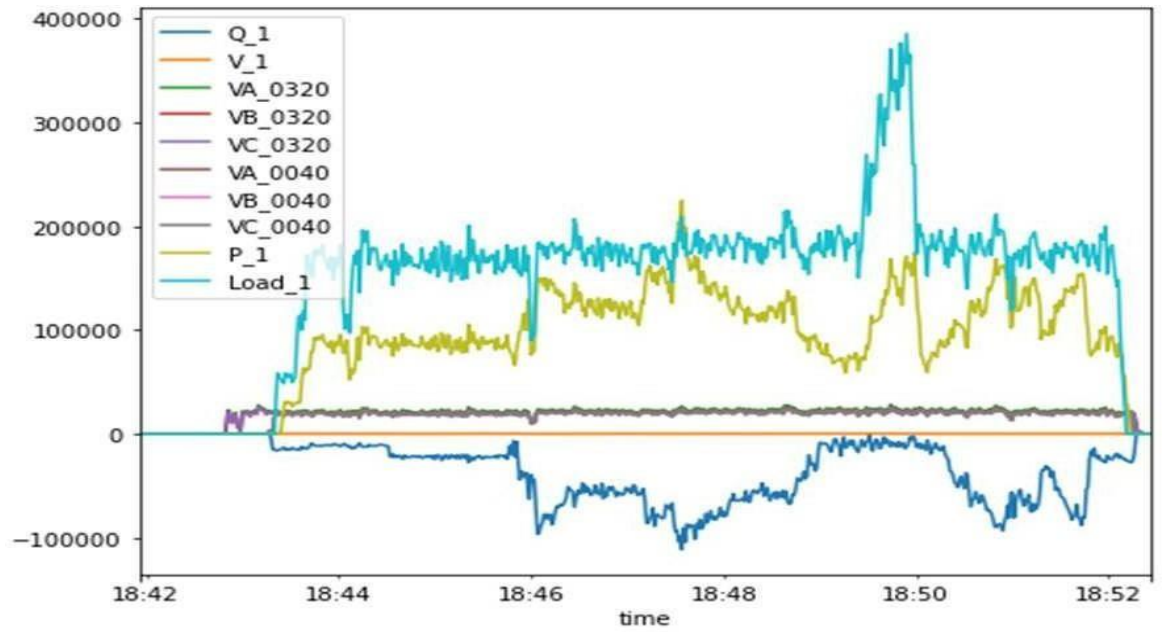


Figure 2: Smart grid testbed Network Architecture

Every point in the architecture is where the data is gathered. Table 1 displays data from the tested architecture. This data predates both feature engineering and feature extraction. Each bus, generator, load, and circuit breaker provides us with information on voltage, power, reactive power, synchronisation status, time, and load values. The following buses are included in the architecture: 0320, 0380, 0260, 0110, 0040, 0050, 0140, 0130. The loads are denoted by the letters L1, L2, L3, and L4. The generators are denoted by the letters G1, G2, G3, and G4.



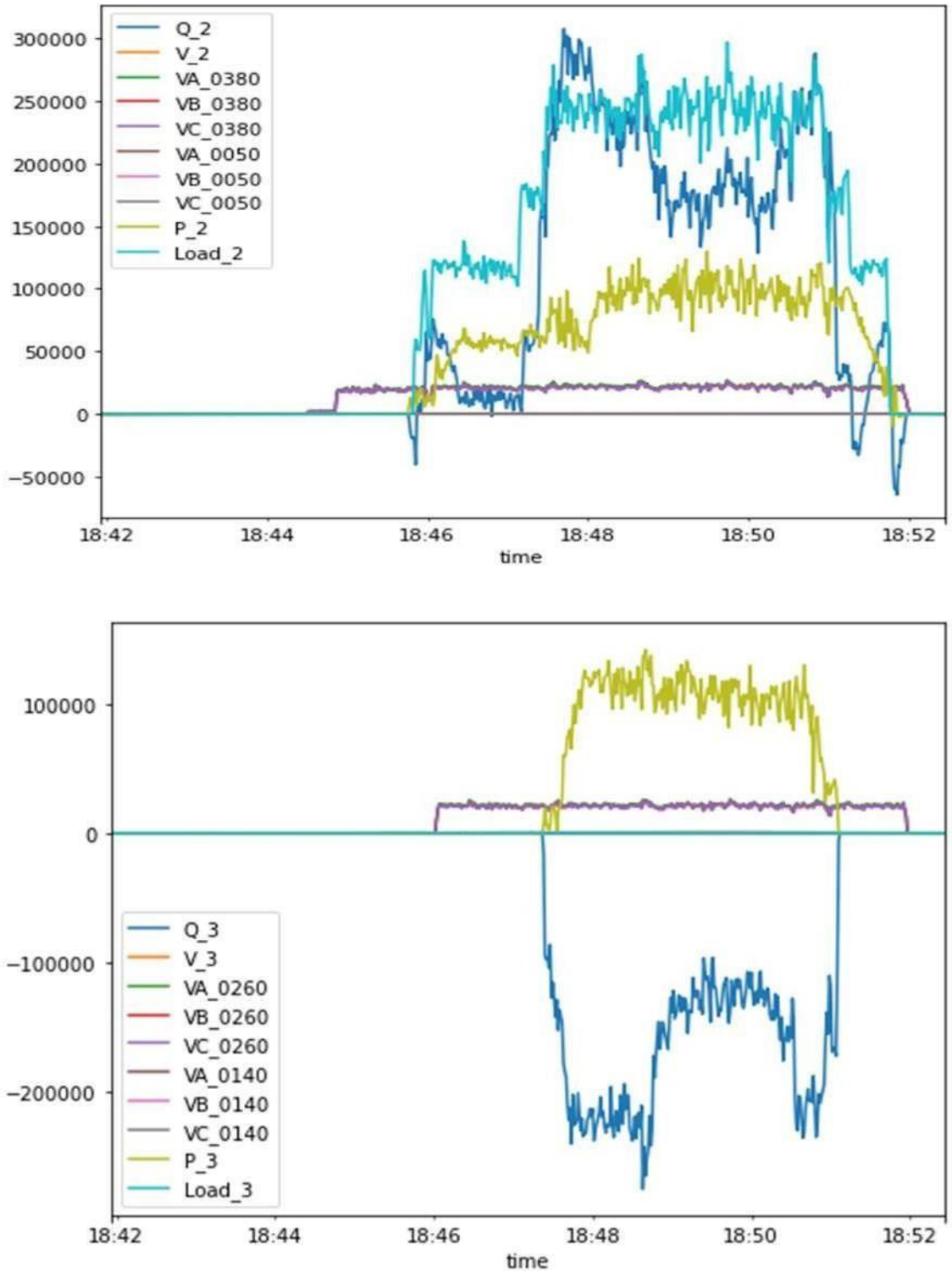


Figure 3: comparison plot for time and grid parameters

Table 1: Data Types and Parameters

Parameter	Data Type	Parameter	Data Type
Time	Int64	VA_0140	Float64
Load_1	Float64	VA_0260	Float64
Load_2	Float64	VA_0320	Float64
Load_3	Float64	VA_0380	Float64
Load_4	Float64	VB_0040	Float64
P_1	Float64	VB_0050	Float64
P_2	Float64	VB_0140	Float64
P_3	Float64	VB_0260	Float64
P_4	Float64	VB_0320	Float64
Q_1	Float64	VB_0380	Float64
Q_2	Float64	VC_0040	Float64
Q_3	Float64	VC_0050	Float64
Q_4	Float64	VC_0140	Float64
Sync_1	Float64	VC_0260	Float64
Sync_2	Float64	VC_0320	Float64
Sync_3	Float64	VC_0380	Float64
Sync_4	Float64	V_1	Float64
Sync_Status_2	Float64	V_2	Float64
VA_0040	Float64	V_3	Float64
VA_0050	Float64	V_4	Float64

The grid parameters relation with time is given in figure

3. The goal is to detect anomalies in every grid parameter separately without consuming a lot of time, which is why we adopted Isolation Forest as an anomaly detection algorithm.

E.Feature extraction, and feature engineering

Now that we own all the data required for developing and evaluating models. Because of zeros and irregular patterns in the data, it is necessary to clean the data. There are more zeros than we anticipated because the data comes from a smart grid that has numerous buses, circuit breakers, four generators, and other components. Every grid parameter has a unique timestamp associated with its definition. This is the feature extraction portion.

In this instance, the initial step was to use Python to create a data frame with the information from each grid parameter. In order to combine the files over mean and sum, we built a merging code.

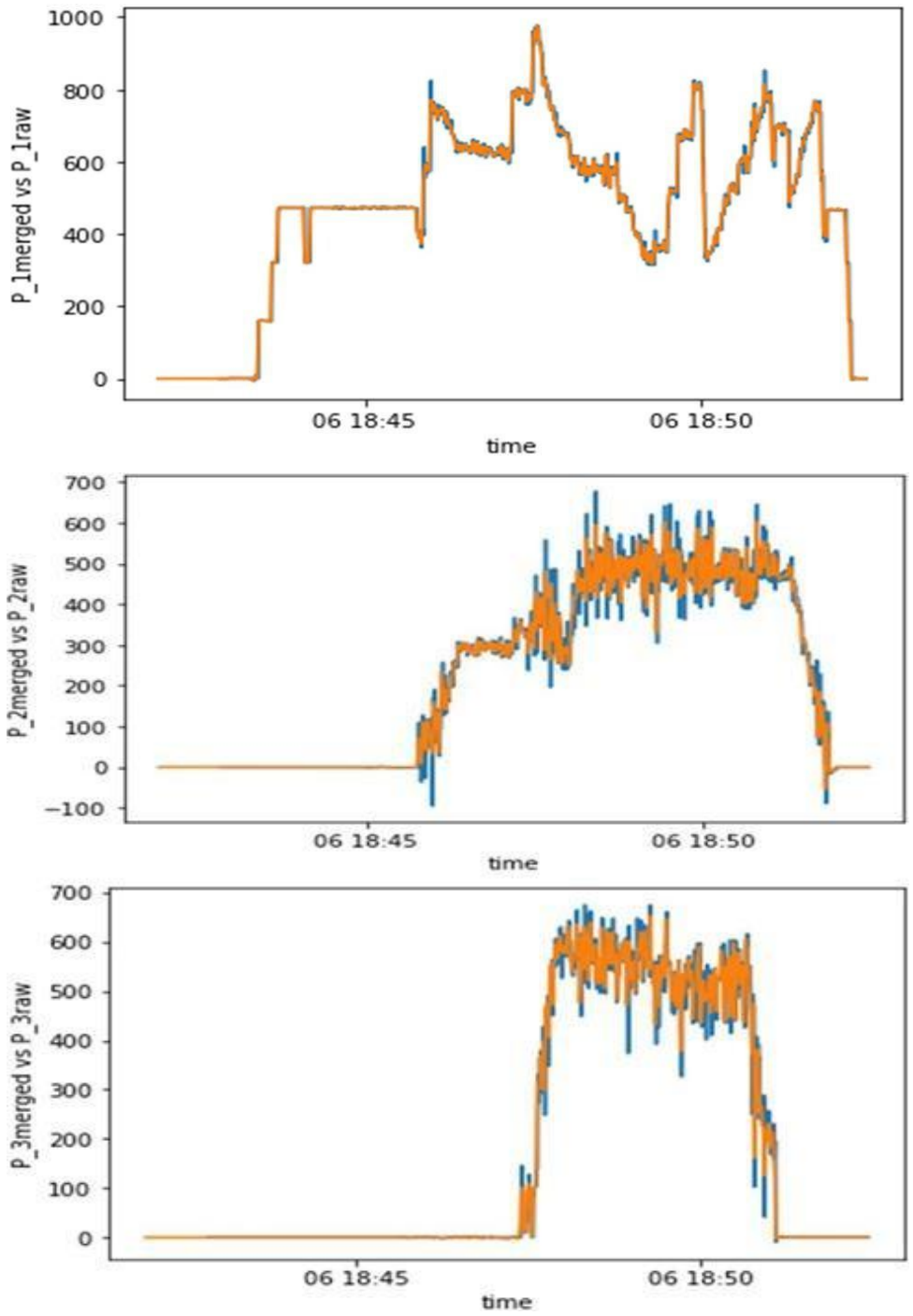


Figure 4: Similarity plot and cases (P_1, P_2, P_3)

The merged file featured a lot of zeros, which was the same issue we had. We experimented with mixing the files over mean, sum, forward fill, and backward fill for a few tries before deciding to alter the frequency of data collection. In order to decrease the number of zeros, we consolidated the time column; as a result, the time-frequency was changed from milliseconds to seconds. Following this procedure, we compared the cleaned and combined file to the original raw dataset for a single grid parameter. The outcome is displayed in figure 6. Next, we determined which traits are required for anomaly identification by examining

Figure 6 and comparing them. The unnecessary cluster was then eliminated. For example, when a value is present at bus 0320, the generator's circuit breaker is closed, allowing current to flow. Therefore, we can disregard that circuit breaker's condition.

To verify all the conditions, the next step is to examine the association between each feature. Figure 4 displays the correlation plot for each parameter taken into consideration. Building the Isolation Forest for anomaly detection comes next, after the data is clean and the required findings are obtained. Based on historical data and trends, Isolation Forest projects future values and contrasts them with the present values. The Isolation Forest labels points as anomalies if the actual recorded values are out of order or do not make sense.

F.Principal Component Analysis

A conventional scaler is used to fit the data to the machine learning model. Principal Component Analysis (PCA) is a more popular technique for accelerating machine learning algorithms [15][24]. Using PCA to accelerate the learning algorithm may make sense if the input dimension is the reason it is operating too slowly. Data visualisation is one more frequent use of PCA. Two-component PCA was employed. Figure 5 illustrates the significance of the primary components based on inertia.

The Augmented Dickey-Fuller Test is one of the statistical tools that we use to determine whether the time series data is stationary or non-stationary (has some time-dependent structure).

- **Null Hypothesis (H0):** If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has some time-dependent structure.
- **Alternate Hypothesis (H1):** The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have a time-dependent structure.

We interpret this result using the p-value from the test. A p-value below a threshold (such as 5% or 1%) suggests we reject the null hypothesis (stationary). Otherwise, a p-value above the threshold means we fail to reject the null hypothesis (non-stationary).

- **p-value > 0.05:** Fail to reject the null hypothesis (H0), the data has a unit root and is non-stationary.
- **p-value <= 0.05:** Reject the null hypothesis (H0), the data does not have a unit root and is stationary.

The PCA value we got from the test is 0.19758759589640856, this specifically means that the data has a time-dependent structure. That is wonderful since all the grid parameters are time dependent.

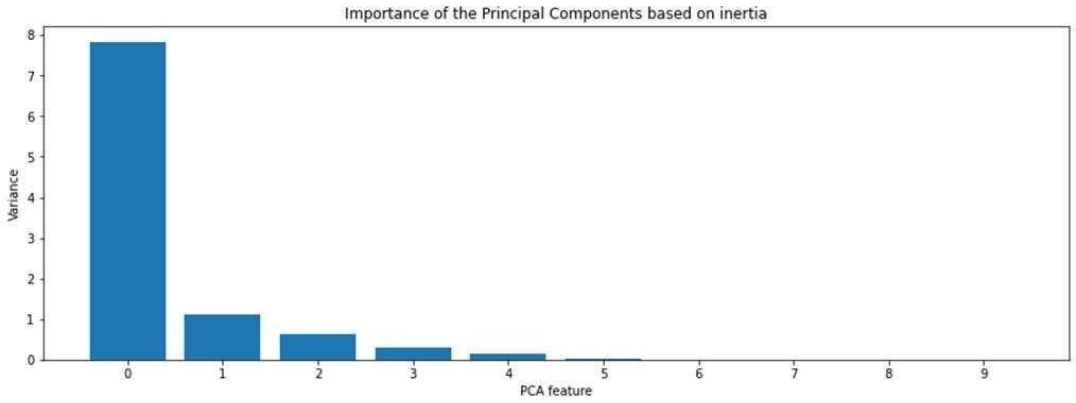


Figure 5: PCA based on inertia

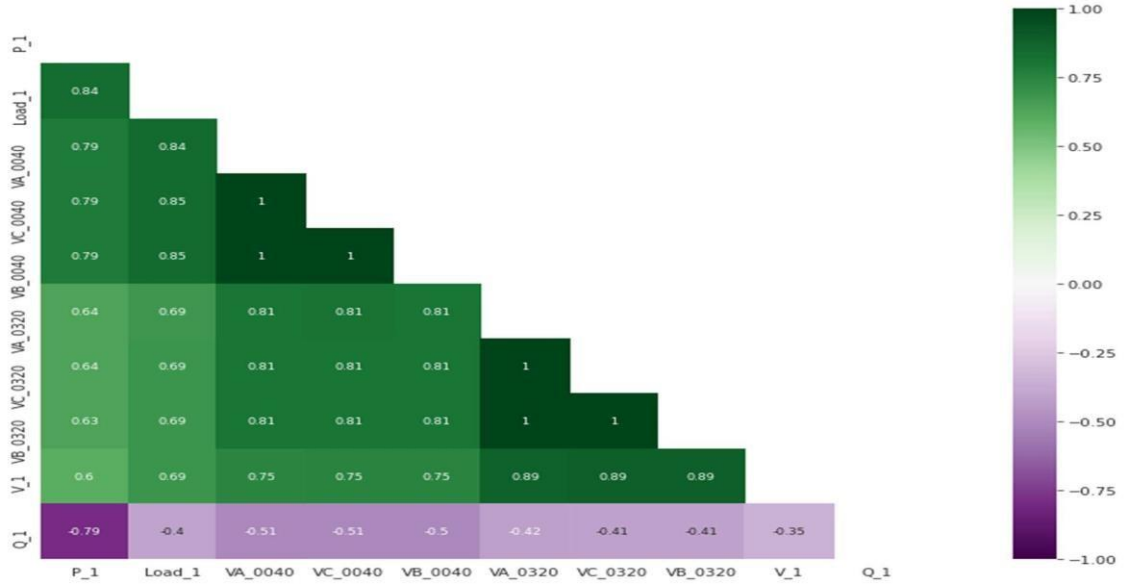


Figure 6: Correlation plot for grid parameters

G.Isolation Forest Model

One common method for creating the model is to use the Isolation Forest, which is covered in section C. There are a few steps in the process of building the tree, and some techniques help to speed up the work. To create an isolation forest, take a sample of each tree's data and choose a dimension at random from the sample. Once the dimension has been chosen, a random value is taken from that dimension, and the data is split at that value by drawing a straight line through it. Until the tree is finished, the aforementioned procedure is repeated. The word "isolation forest" comes from the several trees that are created to make it a forest. The line becomes an N-1 dimensional hyperplane for multi-dimensional data. Principal component analysis, feature extraction, and feature engineering are some of the techniques that help with the process of choosing the dimension, choosing values within the dimension, and splitting the data.

This study considers isolation forests for anomaly detection. In sklearn ensemble modules, the isolation forest algorithm is made available as a class. Design options include bootstrap, n-jobs, max-features, contamination, n-estimators, random-state, verbose, and warm-state. The

number of base estimators in the ensemble is denoted by N-estimators. The number of samples to take from x in order to train each base estimator is called max-samples. The percentage of outliers in the data set is known as contamination. Usually, the remaining parameters are left at their default settings, which are best described. Our concept, which is described in section F, processes max-samples and n-estimators based on data from the PCA. The anomaly ratio is set to 3% because the contamination level is set at 0.03.

We required to train the model as we defined the parameters of the aforementioned model, and we accomplish this by using the data from the real-time hardware setup. Fit the target column for our model is what we are using for this. The primary purpose of anomaly detection is to modify the target each time the parameter is supplied. Every parameter is focused on anomaly inspection with this method. The target column in our model is the PCA analysed column, which is the primary distinction between the fit approach and the standard fit method. In other words, just the columns that are significant or required for detection and prediction are taught to the model.

We add the anomaly column to the dataset at this point, since the model will output the Isolation Forest instance after it has been correctly trained. By using the trained models predict function, we may determine the anomaly column values once the model has been fitted. Observations of anomalies are noted as follows: Anomalies are represented by -1, and normal data is represented by 1. This method has the advantage of labelling every data point or row with an anomaly identifier. By establishing a threshold, we may adjust the anomaly function. If the threshold is set too closely to the normal data points, the model will be tight with the data, and if the threshold is set with a huge gap, the model will be forgiving. The anomalous points in the graphs displayed in Figure 7 are those that have a red cross next to them. The results section explains the cause of the unusual spots in the data.

Evaluating the model's accuracy in classifying a data point as anomalous is the final step in the model building process. Accuracy was the hyperparameter we utilised for this. The model's accuracy percentage is 98.74%, a respectable result given the volume and variety of data. Other metrics, such as mean absolute error (MAE) and mean squared error (MSE), are available to take into account while evaluating a model. But when your data is balanced, accuracy is a useful strategy to take into account. Using a unique testing method known as the Dickey-Fuller test, we were able to identify if the data is balanced or unbalanced. This demonstrates the balance of the data.

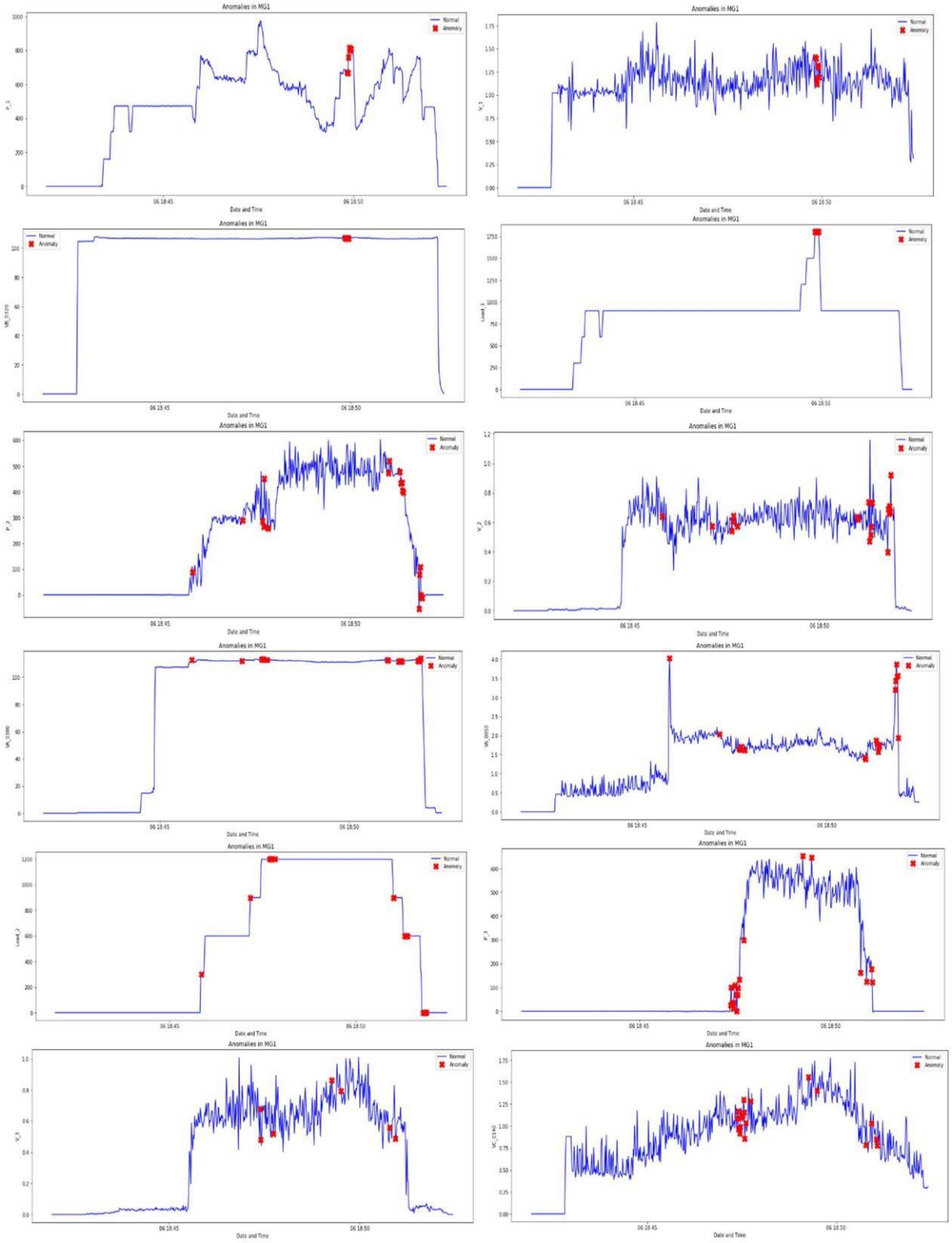


Figure 7: Anomaly plot for individual grid parameters: P_1, V_1, P_2, V_2, V_3, Load_1, Load_2, Load_3, VA_0380, VC_0140, VA_0050, VB_0320.

IV. RESULTS

Assuming that the anomaly ratio is 3%, we ran the model script we wrote (Isolation Forest) for the results section and fitted the model using the PCA analysed values. As may be seen in picture 6 and described in section III (E), we created graphs for each individual parameter

based on how similar they were to one another. The anomalies will now be visualised. To do this, we will annotate the original graph with the anomalies' precise location and timing. Figure 7 displays the results.

All of the grid parameter values and time-plotted graphs are included in the findings and graphs section. The identified anomalies are displayed alongside each individual grid parameter. The voltage, power, and current variations we made to test the model's functionality are the anomalies in the system. The accuracy and results indicate that the model appears to have performed effectively.

The abrupt changes in values for each grid parameter are the cause of the anomalous locations in the graph, which are shown by red cross marks. For instance, the power in the system spikes at 6:18:45, which is the moment when we manually increased the load value to an unanticipated value, as seen by the anomalous points noted in (a)P_1 anomalous chart in Figure 7. In millisecond intervals, anomalies are identified. Microgrid 1's voltage ought to be abnormal based on the relationship between power and voltage. The anomalous spots are labelled in (g)V_1, and the markings occur simultaneously at 6:18:45. The manually increased load levels are displayed in (h)Load_1, where they are also indicated as anomalies. This is an effort to demonstrate how the machine learning model can identify an attack at the utility point that could result in a sharp spike or decrease in load.

We varied the voltage at the generating point in our second attempt to demonstrate the model's performance. This is an attempt to demonstrate whether the generator is having issues producing an erratic load. Improper power generation due to irregular load values might harm household appliances or the Grid. This test case can be found in (e)Load_2, (c)P_2, and (i)V_2. The voltage is altered in the aforementioned anomaly charts from a typical generating point between 6:18:45 and 6:18:50. Two seconds later, at 6:18:50, we adjusted the voltage once more after that. The system notices these modifications and simultaneously marks them in V_2, P_2, and Load_2. This means that every bus's data update is recorded for the entire system.

Let's examine the hardware structure as depicted in figure 2 for a clearer explanation. Since every transmission line is a three-phase line, voltage and current in phases A, B, and C will always be present at every point. Section III provides an explanation of architecture (D).

Taking into account microgrid 1, which is comprised of load 1, bus 0320, bus 0040, and generator 1 (G1). While power generation and voltage can be adjusted to suit the needs of the load, the frequency, 60 Hz, remains fixed. Figure 7 displays graphs that illustrate how load affects voltage and current at each phase. The primary changes in voltage and current that happen during a grid attack also impact power. We have implemented a number of actions in

the smart grid, which we have divided into cases. These instances include power consumption during faults, load sharing for synchronisation, normal operation, maximum load operation, transmission line, and bus failure cases. If we closely examine the values in each of the aforementioned scenarios, voltage, power, and current will help us identify any irregularities. Figure 2's first section shows P_1, Q_1, VA_0320, VB_0320, VC_0320, VA_0040, VB_0040, VC_0040, and Load_1. Let's examine part 1, where all of the aforementioned values should indicate anomalies simultaneously when we unintentionally increased the load. To demonstrate that there is a systemic attack taking place and that the grid is intended to behave abnormally, we abruptly raised the load. We tried this in an effort to test the model we created, and as figure 7 illustrates, at time 06:18:50, all of the parameters from microgrid 1 would exhibit abnormalities.

The individual microgrids that comprise the entire smart grid go through this procedure again. Additional values to examine include P_2, Q_2, VA_0380, VB_0380, VC_0380, VA_0050, VB_0050, VC_0050, and others, similar to part 1. Table [1] displays these values along with the data types for Load_2, P_3, Q_3, VA_0260, VB_0260, VC_0260, VA_0140,

VB_0140, VC_0140, Load_3, P_4, Q_4, VA_0110, VB_0110, VC_0110, VA_0130, VB_0130, VC_0130, and Load_4. Based on the PCA and feature extraction described in III (E), we assessed the significance of each element and chose only the most significant features.

V. CONCLUSION AND FUTURE WORK

Using a machine learning anomaly detection approach, we have carried out anomaly detection and demonstrated the isolation forest's performance. At a specific moment, anomalies in each grid parameter were found. Maximum accuracy was achieved by applying the algorithm with optimised hyperparameters. To test the model, attack scenarios were added, and the outcomes were confirmed. Additionally, we used principal component analysis to optimise feature selection and the Dickey-Fuller test to test the model's performance.

In the future, this procedure might be packaged using Pickle and integrated with the SCADA system to allow action on abnormalities at each grid parameter point.

We combined the two tests listed above to create the third exam. We manually adjusted the voltage and the load. This is an attempt to highlight Grid-damaging anomalies in the Grid settings. The system notices the identical modifications, which are flagged as abnormalities.

REFERENCES

[1] K. Amarasinghe, C. Wickramasinghe, D. Marino, C. Rieger, and M. Manicl, "Framework for data-driven health monitoring of cyber- physical systems," in 2018 Resilience Week (RWS), Aug 2018, pp. 25– 30.

[2] C. S. Wickramasinghe, D. L. Marino, K. Amarasinghe, and M. Manic, "Generalization of deep learning for cyber-physical system security: A survey," in IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society, Oct 2018, pp. 745–751. [3]S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," Procedia Computer Science, vol. 60, no. Supplement C, pp. 708 – 713, 2015, knowledge-Based and Intelligent Information & Engineering Systems 19th Annual Conference, KES2015, Singapore, September 2015 Proceedings. [Online].
Available:http://www.sciencedirect.com/science/article/pii/S1877050915023479

[4] A. M. Kosek, "Contextual anomaly detection for cyber-physical security in smart grids based on an artificial neural network model," in 2016 Joint Workshop on Cyber-Physical Security and Resilience in Smart Grids (CPSR-SG), April 2016, pp. 1–6.

[5] Y. Sun, X. Guan, T. Liu, and Y. Liu,"A cyber-physical monitoring system for attack detection in smart grid," in 2013 IEEE Conference on
Computer Communications Workshops (INFOCOM WKSHPS), April 2013, pp. 33–34. [6]F. T. Liu, K. M. Ting and Z. Zhou, "Isolation Forest," *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17.

[7]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html[8]Pedregosa, Fabian & Varoquaux, Gael & Gramfort, Alexandre & Michel, Vincent & Thirion, Bertrand& Grisel, Olivier & Blondel, Mathieu & Prettenhofer, Peter & Weiss, Ron & Dubourg, Vincent & Vanderplas, Jake & Passos, Alexandre & Cournapeau, David & Brucher, Matthieu & Perrot, Matthieu & Duchesnay, Edouard & Louppe, Gilles. (2012).
Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12.

[9] P. Louridas and C. Ebert, "Machine Learning," in IEEE Software, vol. 33, no. 5, pp. 110-115, Sept.-Oct. 2016, doi: 10.1109/MS.2016.114.

[10] A. Bensalma, "New fractional Dickey Fuller test," 2015 6th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO), 2015, pp. 1-6, doi: 10.1109/ICMSAO.2015.7152263.

[11] S. Liu, Z. Ji and Y. Wang, "Improving Anomaly Detection Fusion Method of Rotating Machinery Based on ANN and Isolation Forest," 2020 International Conference on Computer
Vision, Image and Deep Learning (CVIDL), 2020, pp. 581-584, doi:

10.1109/CVIDL51233.2020.00-23.

[12] S. Luo, L. Luan, Y. Cui, X. Chai, Z. Wang and Y. Kong, "An Attribute Associated Isolation Forest Algorithm for Detecting Anomalous Electro-data," 2019 Chinese Control Conference (CCC), 2019, pp. 3788-3792, doi: 10.23919/ChiCC.2019.8866495.

[13] K. Kittidachanan, W. Minsan, D. Pornnopparath and P. Taninpong, "Anomaly Detection based on GS-OCSVM Classification," 2020 12th International Conference on Knowledge and Smart Technology (KST), 2020, pp. 64-69, doi: 10.1109/KST48564.2020.9059326. [14]https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e

[15] T. Chin and D. Suter, "Incremental Kernel Principal Component Analysis," in IEEE Transactions on Image Processing, vol. 16, no. 6, pp. 1662-1674, June 2007, doi: 10.1109/TIP.2007.896668.

[16] M. Vikram, R. Pavan, N. D. Dineshbhai and B. Mohan, "Performance Evaluation of Dimensionality Reduction Techniques on High Dimensional Data," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 1169-1174, doi: 10.1109/ICOEI.2019.8862526.

[17] C. Guo and W. Luk, "Quantisation-aware Dimensionality Reduction," 2020 International Conference on Field-Programmable Technology (ICFPT), 2020, pp. 237-240, doi: 10.1109/ICFPT51103.2020.00041.

[18] D. Xu, Y. Wang, Y. Meng and Z. Zhang, "An Improved Data Anomaly Detection Method Based on Isolation Forest," *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, 2017, pp. 287-291, doi: 10.1109/ISCID.2017.202.

[19] M. Ozay, I. Esnaola, F. T. Yarman Vural, S. R. Kulkarni and H. V. Poor, "Machine Learning Methods for Attack Detection in the Smart Grid," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 8, pp. 1773-1786, Aug. 2016, doi: 10.1109/TNNLS.2015.2404803.

[20] C. Kaygusuz, L. Babun, H. Aksu and A. S. Uluagac, "Detection of Compromised Smart Grid Devices with Machine Learning and Convolution Techniques," *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1-6, doi: 10.1109/ICC.2018.8423022.

[21] M. D. Levine, "Feature extraction: A survey," in *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1391-1407, Aug. 1969, doi: 10.1109/PROC.1969.7277.

[22] Y. Xie, G. Liu, R. Cao, Z. Li, C. Yan and C. Jiang, "A Feature Extraction Method for Credit Card Fraud Detection," *2019 2nd International Conference on Intelligent Autonomous Systems (ICoIAS)*, 2019, pp. 70-75, doi: 10.1109/ICoIAS.2019.00019.

[23] A. M. Kosek and O. Gehrke, "Ensemble regression model-based anomaly detection for cyber-physical intrusion detection in smart grids," *2016 IEEE Electrical Power and Energy Conference (EPEC)*, 2016, pp. 1-7, doi: 10.1109/EPEC.2016.7771704.

[24] T. Chin and D. Suter, "Incremental Kernel Principal Component Analysis," in *IEEE Transactions on Image Processing*, vol. 16, no. 6, pp. 1662-1674, June 2007, doi: 10.1109/TIP.2007.896668.

[25] R. Punmiya and S. Choe, "Energy Theft Detection Using Gradient Boosting Theft Detector With Feature Engineering-Based Preprocessing," in *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2326-2329, March 2019, doi: 10.1109/TSG.2019.2892595.

[26] A. Stefanov and C. Liu, "Cyber-power system security in a smart grid environment," *2012 IEEE PES Innovative Smart Grid Technologies (ISGT)*, 2012, pp. 1-3, doi: 10.1109/ISGT.2012.6175560.

[27] Z. Yu and W. Chin, "Blind False Data Injection Attack Using PCA Approximation Method in Smart Grid," in *IEEE Transactions on Smart Grid*, vol. 6, no. 3, pp. 1219-1226, May 2015, doi: 10.1109/TSG.2014.2382714.

[28] S. Liu, X. P. Liu and A. El Saddik, "Denial-of-Service (dos) attacks on load frequency control in smart grids," *2013 IEEE PES Innovative Smart Grid Technologies Conference (ISGT)*, 2013, pp. 1-6, doi: 10.1109/ISGT.2013.6497846.

[29] T. R. Sharafeev, O. V. Ju and A. L. Kulikov, "Cyber-Security Problems in Smart Grid Cyber Attacks Detecting Methods and Modelling Attack Scenarios on Electric Power Systems," *2018 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, 2018, pp. 1-6, doi: 10.1109/ICIEAM.2018.8728654.

[30] C. P. Vineetha and C. A. Babu, "Smart grid challenges, issues and solutions," *2014 International Conference on Intelligent Green Building and Smart Grid (IGBSG)*, 2014, pp. 1-4, doi: 10.1109/IGBSG.2014.6835208.

[31] Sook-Chin Yip, Wooi-Nee Tan, ChiaKwang Tan, Ming-Tao Gan, KokSheik Wong, "An anomaly detection framework for identifying energy theft and defective meters in smart grids", International Journal of Electrical Power & Energy Systems, Volume 101, 2018, Pages 189- 203, ISSN 0142-0615,https://doi.org/10.1016/j.ijepes.2018.03.025. (https://www.sciencedirect.com/science/article/pii/S0142061517318719)