

Micro-PMU Data-driven Anomalous Voltage Event Detection for the Power Distribution System [★]

Saththiyan Parameswaran¹, Maitreyee Dey¹, and Preeti Patel¹

Department of Computer Science and Applied Computing, London Metropolitan University, 166-220 Holloway Rd, London N7 8DB, UK
sap1039@my.londonmet.ac.uk, m.dey@londonmet.ac.uk

Abstract. As Distributed Energy Resources become more prevalent in the power grid, the challenges facing grid operators have intensified. The integration of micro-synchrophasors (micro-PMUs) into the grid infrastructure has provided access to high-resolution data, yet analyzing such vast amounts of information presents its own set of obstacles. This study examines thirty days of micro-PMU data from April 2023, employing statistical analysis and unsupervised learning techniques to identify various voltage events over time. Real-world data from grid-connected solar farms in Norfolk, England, is utilized to detect and analyze these voltage events, as well as to explore their distinct patterns.

Keywords: Power Grid · Unsupervised Learning · Voltage Event Detection · Micro-synchrophasor data.

1 Introduction

The fight against climate change necessitates a transition to renewable energy sources, notably solar power. However, inconsistencies and anomalies in solar power generation, such as power failures, can pose significant challenges for power distribution systems and the energy market. The power distribution cycle encompasses various interconnected events spanning from power generation to end-use. The power generation sector has undergone substantial evolution and encountered significant transformations that require a novel approach to monitoring and control. Detecting any real-time abnormalities, such as voltage fluctuations and irregularities, is crucial as they may serve as early warning signals for potential faults or disruptions in the power grid.

It is crucial to identify any abnormalities to maintain a stable power supply capable of meeting demand at all times. Therefore, the integration of micro-PMUs is the recent addition to the power grid aimed at capturing high-precision measurements designed to record voltage, phase angle, frequency, and other

[★] The authors would like to extend their sincere appreciation to Neuville Grid Data for their generous provision of the data, which greatly contributed to the successful completion of this research project.

power-related factors. The objective of micro-PMU-based applications is to improve visibility and situational awareness by offering higher resolution and precise time synchronization, enabling the identification of subtle changes and dynamics across the time dimension. As part of a comprehensive study, abnormal voltage data points, captured in real-time with micro-PMUs, are the primary focus of research.

Clustering, a powerful tool in the realm of power grid analysis, plays a crucial role in unravelling the intricate patterns of events and pinpointing anomalies within the data. By grouping data points with similar characteristics into clusters, this technique allows for a deeper understanding of the underlying structures within a dataset. Moreover, it enables the identification of outliers, which may signify irregularities or potential faults in the system. The widespread application of clustering in fault detection underscores its efficacy as a method for identifying abnormalities within power grid data. By leveraging clustering algorithms, researchers and operators can gain insights into potential issues within the system, paving the way for proactive measures to maintain grid stability and reliability [17].

Therefore, this study employs a clustering algorithm that incorporates the Euclidean distance metric to develop a sophisticated machine-learning model tailored for identifying anomalies within high-resolution voltage data. By utilizing this approach, the model is designed to discern abnormal voltage events, such as voltage sag and voltage swell, within a 33kV solar distribution grid located in England.

2 Literature Review

2.1 Fault Detection and Classification

The contemporary transition towards smart grids, characterized by decentralized generation, dynamic loads, and complex interactions, has spurred the exploration of innovative solutions to ensure stability, reliability, and efficiency in the power system. AI and ML are powerful tools for addressing challenges [6]. In [8], research proposes a novel method for pinpoint source localization of events in distribution feeders using micro-PMUs. The approach leverages the compensation theorem and equivalent circuits to generate pseudo-measurements at unmonitored locations. K-means clustering is then applied to identify event clusters based on voltage and current phasor features. The method demonstrates accurate source localization with low error rates [5]. A comprehensive review of ML-based fault detection techniques highlights the significance of feature selection and model selection strategies, demonstrating promising results with features such as voltage, current, and frequency data [15]. A study by [7], examines frequency events categorization in power distribution systems that can be detected through micro PMU measurements, demonstrating the versatility of machine learning in handling specific fault-related challenges.

2.2 Stability Assessment

Grid stability assessment, crucial for preventing blackouts and ensuring efficient power delivery, has been explored through AI-based methods, including EEMD-SVM, ANN, and PSO-SVM [12]. These techniques exhibit promise in capturing the complex dynamics of converter-dominated power systems, contributing to more accurate and proactive stability assessment [12]. Furthermore, deep learning in online frequency stability assessment and control, utilizing a CNN-LSTM network to extract spatial and temporal features from power system data, has demonstrated high accuracy in predicting system frequency nadir [20]. The literature also addresses forecasting challenges in renewable energy sources and load demand for optimizing grid operation and resource allocation [10] [19]. Another research proposes a novel method for pinpointing source localization of events in distribution feeders using micro-PMUs. The approach leverages the compensation theorem and equivalent circuits to generate pseudo-measurements at unmonitored locations [5]. K-means clustering is then applied to identify event clusters based on voltage and current phasor features. The method demonstrates accurate source localization with low error rates (MSE 0.0007 pu, RMSE 0.0011 pu) [8]. Another work focused on detecting and classifying events in both event logs and raw PMU data. An ensemble of K-means clusters captures different event patterns in the data. Statistical, temporal, and spatial features are extracted for event characterization. The proposed method achieves high event detection accuracy (93 %), particularly for categories like fault and switching events [11]. A generative adversarial network (GAN) for unsupervised event detection in micro-PMU measurements had been used to distinguish normal and abnormal data patterns, enabling anomaly detection as potential events [1].

2.3 Supervised Learning for Forecasting

While unsupervised learning excels at anomaly detection and pattern recognition, supervised learning offers valuable tools for tasks like regression and forecasting. A study employs feature selection techniques to identify relevant event detection and classification features. Logistic regression and decision trees are then utilized for event identification and categorization. The proposed system provides insights into grid state and event characteristics, contributing to improved situational awareness for operators [16]. LSTM recurrent neural networks have proven effective in forecasting solar and wind power in isolated microgrids, leading to improved load frequency control [10]. Additionally, a study investigating ML algorithms for short-term forecasting of peak load hours in regional power systems identifies gradient boosting machines as the most effective method [19]. Challenges and future directions include the development of robust and generalizable AI models, seamless integration of AI with grid technologies, addressing data privacy and security concerns, fostering explainable AI, and exploring novel AI architectures such as neuromorphic computing and reinforcement learning [18]. In [9], the history of frequency prediction models in power systems examines the limitations of traditional model-driven methods such as time-domain

simulation and equivalent models due to their high computational cost and accuracy constraints had been explained. A vertical power flow forecasting using Long Short-Term Memory (LSTM) neural networks has been studied in [3]. The study investigates the effectiveness of different regular training update strategies for improving forecast accuracy. The proposed approach with regular updates achieves significant reductions in forecast errors (MAE 0.014 pu, RMSE 0.017 pu) compared to traditional LSTM models. Anomaly detection in power generation plants using various machine learning algorithms, including SVM, KNN, LR, and MLP presented in [13]. The study evaluates anomaly detection performance based on fuel consumption data. MLP neural networks achieved the best results with an Area Under Curve (AUC) of 0.98, demonstrating their effectiveness for anomaly detection in power generation applications.

2.4 Research without Micro-PMU Data

While the previously reviewed studies primarily focus on micro-PMU data, the following research explores integrating additional data sources for enhanced power system analysis. A study investigates outlier detection and removal algorithms in K-means and hierarchical clustering for datasets like Wholesale Customers. The work demonstrates the effectiveness of outlier removal in improving clustering performance and highlights the importance of data pre-processing for unsupervised learning tasks [2]. A novel unsupervised K-means clustering algorithm with an automatic cluster number determination mechanism was proposed in [17]. The proposed algorithm outperforms existing K-means methods in terms of accuracy. Another study delves into the topic of anomaly detection in power generation plants, using a range of machine learning algorithms like SVM, KNN, and MLP. The results show that MLP achieved the best performance, which highlights the effectiveness of deep learning techniques for detecting anomalies in power systems [4]. Additionally, the study compared the performance of K-means, Local Outlier Factor (LOF), and One-Class SVM (OC-SVM) for anomaly detection. The results showed that OC-SVM demonstrated superior performance, which indicates its potential for detecting anomalies in power systems.

The primary objective of this research is to analyze voltage data obtained from an actual 33kV solar site, diverging from the prevalent focus on synthetic datasets in this field [14]. Using clustering techniques, the study endeavours to automatically detect various voltage events within high-resolution micro-PMU data collected from a solar farm situated in Norfolk, England. Statistical methods, particularly clustering algorithms integrated with the Euclidean distance metric and standard deviation, are utilized to identify abnormal voltage occurrences, such as voltage sag and voltage swell. The overarching objective of this model is to enhance the reliability of the grid by enabling the timely detection of deviations from normal voltage levels. The research offers valuable insights into the patterns and characteristics of these voltage events by using real-world data obtained from a grid-connected solar farm in Norfolk, England. This departure from synthetic datasets enhances the applicability and reliability of the

findings, contributing to a deeper understanding of voltage dynamics in solar energy systems.

3 Research Methodology

The methodology proposed for detecting anomalies in micro-synchrophasor voltage data employs an unsupervised learning algorithm, namely K-means clustering. This approach is tailored to target anomalies within datasets, following a comprehensive five-step process.

- Firstly, the dataset is accumulated and pre-processed to ensure its suitability for analysis. This step involves cleaning the data, handling missing values, and organizing it in a format conducive to subsequent analysis.
- Secondly, the optimal number of clusters is determined using the Within-Cluster Sum of Squares (WCSS) method. By identifying the ideal cluster count, the analysis can effectively partition the data into meaningful groups.
- Thirdly, the K-means clustering algorithm is applied to segment the voltage data into clusters, thereby delineating distinct operational states. This enables the identification of patterns and similarities within the dataset, aiding in the detection of anomalies.
- Fourthly, the Euclidean distances between data points and the centroid within each cluster are computed. These distances serve as a measure of deviation from the cluster’s central tendency, helping to identify potential anomalies.
- Lastly, the mean and standard deviation of these distances are calculated to establish a dynamic threshold for anomaly detection. By determining a threshold based on statistical measures, the methodology can effectively differentiate between normal variations and anomalous events in the voltage data.

The process concludes with visualizing anomaly data using appropriate graphs that illustrate voltage readings over time, offering comprehension of anomalies. Through this systematic approach, the proposed methodology aims to improve anomaly detection capabilities in micro-synchrophasor voltage data, thereby advancing the detection and prediction of anomalies within solar power generation plants.

3.1 Data Collection

This study examines data collected from a utility-scale solar farm situated in Norfolk, United Kingdom. Utilizing a micro-Phasor Measurement Unit (μ PMU), the data captures 8.64 million voltage data points every 10 milliseconds per day. This extensive dataset forms a robust foundation for investigating voltage variations, facilitating the capture of both transient events and steady-state conditions.

The data obtained from this site includes measurements of 3-phase voltage magnitude, alongside other electrical parameters. This research focused on voltage-related occurrences, specifically voltage sag and voltage swell, conducting a comprehensive analysis of both phenomena. In Fig. 1 below, voltage data for a typical day (24 hours) is presented, encompassing a total of 8.64 million data points for each phase, with distinct colors representing the unique patterns of each phase’s data.

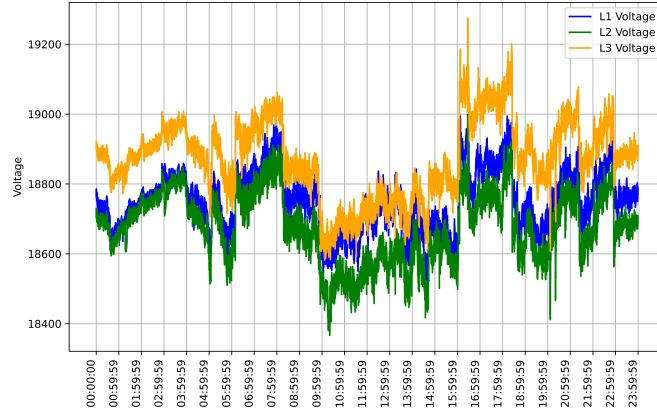


Fig. 1: A normal single day’s three-phase voltage data visualisation

While Fig. 2 illustrates the voltage data distribution for the specific day depicted in Fig. 1, it showcases the distribution of data points across the three phases. The distinct colors denote the data distribution for each phase, highlighting a well-balanced system. Moreover, the data falls within a range of ± 6

3.2 Within-Cluster Sum of Squares (WCSS)

In this research data, the elbow method is utilized to determine the ideal number of clusters. This method involves calculating the WCSS, which is a significant metric in clustering algorithms used to evaluate the similarity of data within each cluster. The process involves selecting several clusters (K), assigning data points to these clusters randomly, and then calculating the centroids and distances between data points and their assigned cluster centroids iteratively. WCSS is calculated by summing the squared distances for all data points within each cluster, which provides a measure of how tightly the data is grouped within clusters. The formula for WCSS considers the squared differences between each data point and its cluster centroid. By repeating this process for different values of k and selecting the one that minimizes WCSS, the algorithm can identify the optimal number of clusters that best represent the data. A lower WCSS

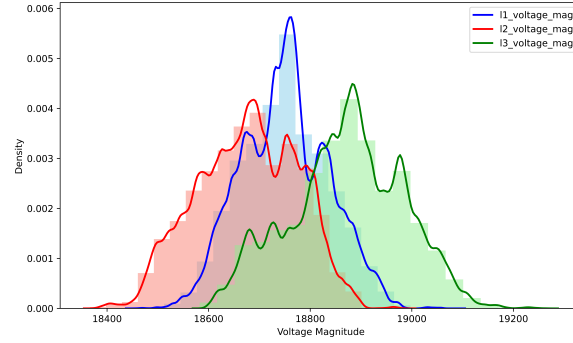


Fig. 2: A normal single day's three-phase voltage data distribution

indicates greater similarity among data points within clusters, which suggests that the clustering algorithm is more effective in grouping the data. [1] also used this method to find the optimum number of clusters in their research.

3.3 K-means Clustering

The K-means clustering algorithm is commonly used in machine learning. It operates under the assumption that the number of clusters, denoted as 'K', is known. K-means is a flat clustering algorithm designed to minimize the sum of squared distances between data points and their corresponding cluster centroids. Initially, K-means randomly selects K data points and assigns them to clusters. The algorithm then iteratively computes cluster centroids and updates data point assignments until convergence. This process involves calculating the sum of squared distances between data points and centroids, reassigning data points to the closest cluster centroid, and recalculating centroids based on the newly assigned data points. K-means employs the expectation-maximization strategy, utilizing the Expectation step to assign data points to clusters and the maximization step to compute the centroids. The iterative process continues until an optimal solution is achieved, reducing diversity within clusters and increasing similarity among data points within the same cluster.

3.4 The Euclidean distance

The Euclidean distances capture spatial relationships between data points and their respective cluster centroids, serving as a fundamental measure of dissimilarity. Such anomalies often manifest as data points located far away from their cluster centroids, indicating significant deviations from established patterns within the cluster.

The calculated distances play a crucial role in determining a dynamic anomaly threshold, which incorporates the mean and standard deviation of distances

within each cluster. This threshold facilitates the classification of data points as either normal or abnormal based on their Euclidean distances from the centroid. Thus, the Euclidean distance metric is a critical tool in discerning patterns and enabling the identification of anomalies in micro-synchrophasor voltage data within the framework of K-means clustering. This integrated approach enhances the precision of anomaly detection and contributes to the broader understanding of dynamic behaviors within the intricacies of voltage data. The following equation is used to calculate the distance to a data point from its cluster centroid.

$$d(x_i, centroid_k) = \sqrt{\sum_{i=1}^n (centroid_{ki} - x_{ii})^2} \quad (1)$$

3.5 Dynamic Thresholding

During the research study, anomalies were detected in the voltage data collected from the solar farms. The Euclidean distance was used to identify these anomalies. Initially, the data was clustered, and then the mean and standard deviation of the Euclidean distance between each data point and its centroid were calculated for each cluster. An outlier is identified by measuring its distance from the centroid and comparing it to a distance threshold determined by the mean and standard deviation. The statistical approach adopted in this study helped in detecting unusual voltage behaviors and anomalies within the data. The clustered data was also examined using the three-sigma rule to highlight anomalous events. If the distance of each data point exceeded the three standard deviations from its own centroid, the data point was considered an outlier. Therefore, 3 SD is the threshold for this research.

4 Result Analysis and Discussion

In this study, a vast dataset consisting of 259.2 million voltage data points was utilized. These data were collected over a comprehensive 30-day period, ranging from April 1st to April 30th, at the previously mentioned solar site. The experiment focused on daily data, aiming to comprehend daily events, their timing, and potential causes. Additionally, all outlier events captured by the method were validated using power quality data provided by Neuville Grid Data.

The substantial volume of data enabled robust statistical analyses, facilitating the identification of patterns, trends, and anomalies with a high level of accuracy and confidence. This comprehensive approach offers valuable insights into the dynamics of the voltage data, contributing to a deeper understanding of the phenomenon under investigation.

4.1 Clustering Analysis

The analysis of experimental results commences with determining the optimal number of clusters for the data. This investigation was conducted separately

for several days of data, and the results presented are for a single day of data. Fig. 3 illustrates the data of all voltage lines plotted to determine the number of clusters. As depicted below, $K = 3$ is considered to be the optimal number of clusters for this dataset.

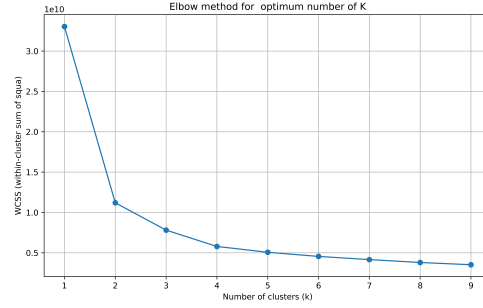


Fig. 3: Optimal number of cluster selection for a single day voltage data using WCSS Method

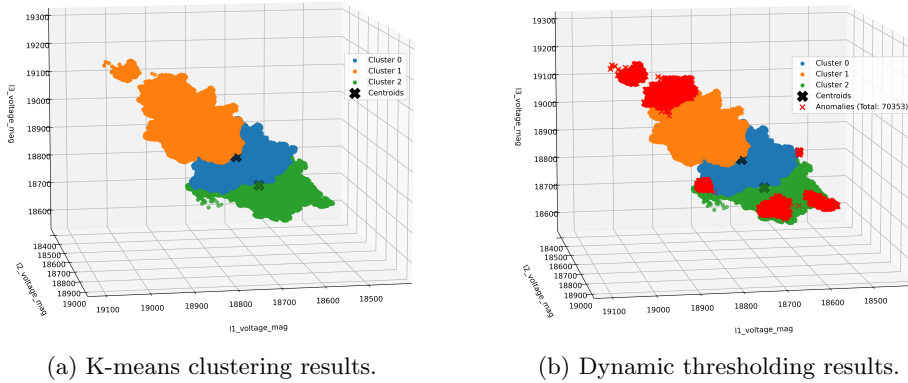


Fig. 4: The clustering and anomaly detection results for a single day (April 20th, 2023).

Upon determining the optimal number of clusters, the K-means algorithm was applied to identify distinct patterns within the voltage data. This process was conducted separately for each day's data. Fig. 4a illustrates the clustering outcome for April 20th, 2023. In the figure, three different colors represent three distinct clusters, with the random state selected as 1200.

Following the clustering process, dynamic thresholding was applied to each cluster to identify outlier data points within them. Fig. 4b illustrates the anomaly points highlighted in red, which were identified by the 3-standard deviation (3-SD) threshold.

4.2 Visualization of Detected Anomalies

A detailed visual exploration of detected anomalies across three distinct voltage phases is shown in Fig.5, each sub-figure showcasing the time series voltage data for April 20th, 2023. In each figure, voltage data points are depicted in three different colors to aid in distinguishing between the voltage phases. The anomalous voltage points are highlighted in red within each figure, suggesting potential voltage sags or swells, which are events of particular interest in this analysis. Each phase's voltage data is examined in its respective image, offering a clear and comprehensive view of anomalous events over time. Such visual representation provides a dynamic perspective on the temporal patterns and occurrences of anomalies within the voltage dataset.

All anomalous data points detected by the method undergo validation through power quality measurements conducted by the operators.

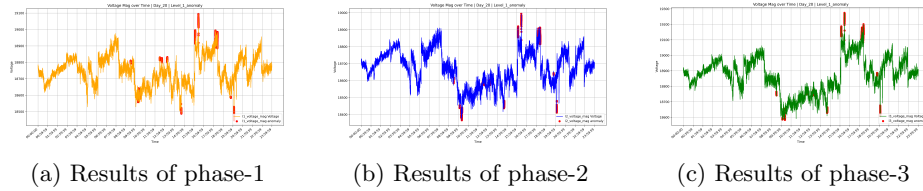


Fig. 5: The results of detected anomalies in the time-series voltage data for a single day.

5 Result Comparison

As part of this research on identifying the most effective anomaly detection method, a comparative analysis with two other prominent techniques: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Isolation Forest Model (IFM) has been conducted, shown in Fig. 6. While DBSCAN is excellent at identifying clusters based on density, IFM isolates anomalies by randomly partitioning feature space. However, after thorough evaluation, the K-Means algorithm emerged as the preferred choice for continued research work.

The comparisons are depicted in Fig.6. Fig.6a displays the results of DBSCAN, indicating that all three phases' data are selected as anomalies. Meanwhile, Fig. 6b illustrates the results of IFM, where it can be observed that although it detects some abnormal events, it also identifies many normal data points as anomalies.

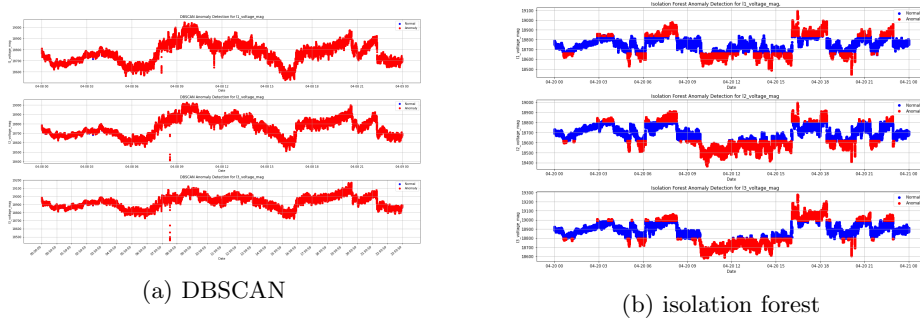


Fig. 6: A comparison of results obtained using DBSCAN and Isolation Forest for the given day.

6 Conclusions and Future Work

The findings of this study carry significant implications for enhancing anomaly detection within the power distribution grid. By effectively managing a large dataset of 8.64 million data points per day and utilizing machine-learning techniques to identify voltage anomalies, the study contributes to advancing grid monitoring and maintenance practices.

Moving forward, the team will delve deeper into investigating voltage anomalies and determining their root causes. Additionally, they will explore other anomalous events to gain a better understanding of their relationship with operating conditions. The objective is to pinpoint the underlying reasons for these outliers and leverage that information to enhance operational processes.

References

1. Armin Aligholian, Alireza Shahsavari, Emma M Stewart, Ed Cortez, and Hamed Mohsenian-Rad. Unsupervised event detection, clustering, and use case exposition in micro-pmu measurements. *IEEE Transactions on Smart Grid*, 12(4):3624–3636, 2021.
2. Anwesha Barai and Lopamudra Dey. Outlier detection and removal algorithm in k-means and hierarchical clustering. 2017.
3. Katharina Brauns, Christoph Scholz, Alexander Schultz, André Baier, and Dominik Jost. Vertical power flow forecast with lstms using regular training update strategies. *Energy and AI*, 8:100143, 2022.
4. Efrem Heri Budiarto, Adhistya Erna Permanasari, and Silmi Fauziati. Unsupervised anomaly detection using k-means, local outlier factor and one class svm. In *2019 5th international conference on science and technology (ICST)*, volume 1, pages 1–5. IEEE, 2019.
5. Maitreyee Dey, Soumya P Rana, James Wylie, C V Simmons, and Sandra Dudley. Detecting power grid frequency events from μ pmu voltage phasor data using machine learning. 2022.

6. Maitreyee Dey, Soumya Prakash Rana, Clarke V Simmons, and Sandra Dudley. Solar farm voltage anomaly detection using high-resolution μ pmu data-driven unsupervised machine learning. *Applied Energy*, 303:117656, 2021.
7. Nan Duan and Emma M Stewart. Frequency event categorization in power distribution systems using micro pmu measurements. *IEEE Transactions on Smart Grid*, 11(4):3043–3053, 2020.
8. Mohammad Farajollahi, Alireza Shahsavari, Emma M Stewart, and Hamed Mohsenian-Rad. Locating the source of events in power distribution systems using micro-pmu data. *IEEE Transactions on Power Systems*, 33(6):6343–6354, 2018.
9. Yongfei Hu, Huaiyuan Wang, Yang Zhang, and Buying Wen. Frequency prediction model combining isfr model and lstm network. *International Journal of Electrical Power & Energy Systems*, 139:108001, 2022.
10. Dhananjay Kumar, HD Mathur, S Bhanot, and Ramesh C Bansal. Forecasting of solar and wind power using lstm rnn for load frequency control in isolated microgrid. *International Journal of Modelling and Simulation*, 41(4):311–323, 2021.
11. Tu Lan, You Lin, Jianhui Wang, Bruno Leao, and Dmitriy Fradkin. Unsupervised power system event detection and classification using unlabeled pmu data. In *2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, pages 01–05. IEEE, 2021.
12. Wentao Liu, Tamas Kerekes, Tomislav Dragicevic, and Remus Teodorescu. Review of grid stability assessment based on ai and a new concept of converter-dominated power system state of stability assessment. *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, 2023.
13. Jecinta Mulongo, Marcellin Atemkeng, Theophilus Ansah-Narh, Rockefeller Rockefeller, Gabin Maxime Nguenang, and Marco Andrea Garuti. Anomaly detection in power generation plants using machine learning and neural networks. *Applied Artificial Intelligence*, 34(1):64–79, 2020.
14. Preeti Patel. Synthetic data. *Business Information Review*, page 02663821241231101, 2024.
15. Angel Esteban Labrador Rivas and Taufik Abrao. Faults in smart grid systems: Monitoring, detection and classification. *Electric Power Systems Research*, 189:106602, 2020.
16. Alireza Shahsavari, Mohammad Farajollahi, Emma M Stewart, Ed Cortez, and Hamed Mohsenian-Rad. Situational awareness in distribution grid using micro-pmu data: A machine learning approach. *IEEE Transactions on Smart Grid*, 10(6):6167–6177, 2019.
17. Kristina P Sinaga and Miin-Shen Yang. Unsupervised k-means clustering algorithm. *IEEE access*, 8:80716–80727, 2020.
18. Ferhat Ucar. A comprehensive analysis of smart grid stability prediction along with explainable artificial intelligence. *Symmetry*, 15(2):289, 2023.
19. V Voronin, F Nepsha, and M Krasilnikov. Short term forecasting peak load hours of regional power systems using machine learning methods. *CIGRE Science&Engineering*, 1(29), 2023.
20. Jian Xie and Wei Sun. A transfer and deep learning-based method for online frequency stability assessment and control. *IEEE Access*, 9:75712–75721, 2021.