And., Intl.,J., Appl Sci., Eng., Tech, Vol 05, No 02, July 2025
e-ISSN: 2797-0442

Prediction of Electrical Grid Stability using Naïve Bayes and K-Means Algorithm

Baik Budi*, Muhammad Ilhamdi Rusydi, Reivan Arya Witama, Queen Hesti Ramadhamy, Refki Budiman
Department of Electrical Engineering, Engineering Faculty, Universitas Andalas, Padang, Indonesia
*Corresponding author, e-mail: baikbudi@eng.unand.ac.id

Abstract— Predicting electrical grid stability is critical for the reliability of modern power infrastructure. This study evaluates the Naïve Bayes (NB) algorithm for this classification task, demonstrating its high efficacy. When validated on the "Electrical Grid Stability Simulated Data" dataset, the NB model achieved exceptional performance metrics: an accuracy of 97.35%, precision of 98.91%, recall of 97.02%, and an F1-score of 97.95%. These results establish Naïve Bayes as a robust and reliable classifier for stability assessment, balancing high accuracy with crucial sensitivity to instability events. This research utilized the K-Means clustering algorithm to identify the optimal data structure within the 'Electrical Grid Stability Simulated Data' dataset. By applying the Elbow Method and analyzing the average within-cluster distance, the analysis conclusively determined that K=4 is the optimal number of clusters. This finding indicates that partitioning the data into four groups provides the most meaningful segmentation, balancing cluster compactness with model complexity.

Keywords: Naïve Bayes, K-Means, Smart Grid.

This article *is licensed under the CC–BY-SA license.*

1. Introduction

The stability and efficiency of electrical grids are paramount in modern society, where electricity is a vital energy source derived from diverse origins such as fossil fuels, nuclear power, and a growing portfolio of renewables, including wind, water, and solar power. The relentless growth in global population correlates directly with a steady increase in electricity consumption, placing significant strain on existing power infrastructures. In this context, machine learning (ML) has emerged as an indispensable tool for enhancing power systems' monitoring, management, and stability analysis. [1]. A key application of ML is within intelligent smart grids, which offer a pathway to optimize energy distribution and mitigate the need to construct additional power plants by accurately predicting and managing energy consumption patterns. [2].

Adopting smart grid technology in Indonesia is a strategic imperative for advancing the nation's energy infrastructure. It promises to significantly boost energy efficiency, enable active consumer participation through demand-response programs, and promote environmental sustainability by facilitating the large-scale integration of new renewable energy (NRE) sources into the national grid. [3]. Furthermore, smart grids are crucial for implementing Demand Side Management (DSM) strategies, which aim to balance the grid by modulating energy demand, thereby offering improved stability and cost savings. [4]. In the industrial sector, these technologies allow for granular control over power distribution, reducing operational costs and enhancing productivity. Despite these compelling advantages, widespread implementation in Indonesia faces considerable hurdles, including the need for substantial infrastructure investment, the high cost of technology, and limited public awareness, all of which must be systematically addressed. [5].

Received: February 24, 2025, Revised: May 2, 2025, Accepted: May 30, 2025
https://doi.org/10.25077/aijaset.v5i02.223

Machine Learning for Power System Stability

Numerous machine learning algorithms have been explored for power system analysis. The Naïve Bayes algorithm, a probabilistic classifier based on Bayes' theorem, is recognized for its computational efficiency and strong performance with categorical data. While it has been applied in domains like electricity theft detection, comparative studies consistently indicate that it is often outperformed by more complex models like Support Vector Machines (SVM) and ensemble methods like Random Forest, particularly when dealing with the class imbalance inherent in such datasets. [6] [7]. However, its utility in other power system applications remains significant. For instance, recent research highlights its effectiveness in transient stability assessment, where its speed and predictive power, especially when augmented with boosting techniques like AdaBoost, provide a valuable tool for real-time grid analysis. [8].

In parallel, the K-Means algorithm, an unsupervised clustering method, is widely utilized for data preprocessing in intelligent grid analytics. Its primary function is to partition large datasets into distinct clusters based on feature similarity, making it highly effective for tasks like load profiling, anomaly detection, and data cleaning. [9]. Research has demonstrated that using K-Means to preprocess data before classification can significantly enhance model performance. By identifying and removing or correcting inconsistent data points and outliers, K-Means improves the overall quality of the dataset, leading to more accurate and reliable outcomes from subsequent classification algorithms. [10] [11]. The synergy of combining clustering for data refinement with classification for prediction is a promising direction for developing more sophisticated grid management tools. [12].

This study proposes a hybrid methodology combining K-Means clustering and the Naïve Bayes classification algorithm to enhance the assessment of electrical grid stability. The novelty of this research lies in the strategic application of K-Means for data preprocessing and cleaning. By first clustering the grid data, we can effectively identify and isolate noise, outliers, and inconsistent data points that could otherwise degrade classification performance. The refined, high-quality dataset is then analyzed using the Naïve Bayes algorithm to classify the grid's stability. This integrated approach is hypothesized to improve the accuracy and reliability of stability predictions, offering a more robust solution for managing modern power systems' complex and fluctuating conditions. The data utilized for this research is a publicly available dataset from Kaggle. [13], ensuring the reproducibility of our findings.

.

2. Method

This research methodology is centered on the analysis of the "Electrical Grid Stability Simulated Data" dataset, which was sourced from the UCI Machine Learning Repository and is publicly available on the Kaggle platform. This dataset is not raw observational data; it was generated using a realistic simulation model of a simplified power grid. It is a standard benchmark for evaluating machine learning models in power system analysis. The simulation models a grid consisting of one producer and three consumers, a classic configuration used for studying decentralized smart grid control [14].

The dataset comprises 10,000 data points, each defined by 14 attributes. For this study, we utilize 12 predictive features that describe the state of the grid and one target variable that indicates its stability. The key features include:

- **Reaction Times (tau1 to tau4):** These represent the time constants for the power producer (tau1) and the three consumers (tau2, tau3, tau4), modeling their response delays within the system.
- **Power Values (p1 to p4):** These features correspond to the net power produced by the producer (p1) and consumed by the consumers (p2, p3, p4). The system is constrained such that the sum consumed equals the power produced.

- **Price-Elasticity Coefficients (g1 to g4):** These coefficients represent the price elasticity of the producer (g1) and consumers (g2, g3, g4), which influences their power adjustment behavior.
- **Stability Metrics:**
  - **stab:** A continuous-valued feature representing the maximum fundamental part of the characteristic equation's complex root. This value directly determines the linear stability of the system; a positive value signifies instability.
  - **stabf:** The categorical target variable, labeled "stable" or "unstable," derived directly from the stab value. This feature serves as the ground truth for our classification model.

This dataset is well-established within the research community. It has been employed in numerous studies to train and validate machine learning algorithms for predicting power grid stability, demonstrating its suitability for developing robust predictive models. [15]. The curated nature of this dataset ensures a controlled environment for assessing the performance of our proposed K-Means and Naïve Bayes methodology.

Data_for_UCI_named.csv (2.42 MB)

Detail Compact Column

# tau1	# tau2	# tau3	# tau4	# p1	# p2	# p3	# p4	# g1	# g2
2.9590600245599 7	3.0798852042281 1	8.3810253919188 2	9.7807544322260 7	3.7630847720631 6	-0.782603630987 543	-1.257394829587 32	-1.723086311488 3	0.6504564608872 27	0.8595781057523 45
9.3040972346785	4.9025241120116 7	3.0475407276217 7	1.3693573552960 5	5.0678121042784 5	-1.940058427051 93	-1.872741685597 21	-1.255011991629 31	0.4134405683793 5	0.8624140763529 03
8.9717069093202 2	8.8484284213483 3	3.0464787489886 6	1.2145181383395 6	3.4051581800109 5	-1.207455592343 02	-1.277210146732 95	-0.920492440934 98	0.1630410393113 34	0.7666886565269 62
0.7164147762951 21	7.6695996440656 5	4.4866408305894 9	2.3405629839679 5	3.9637910632663 3	-1.027473304139 05	-1.938944152646 6	-0.997373606480 681	0.4462089065373 21	0.9767440829243 02
3.1341115516134 2	7.6087716160340 8	4.9437593017809 9	9.8575732699663 8	3.5258108165209 6	-1.125530954511 15	-1.845974854475 61	-0.554305007534 195	0.7971095257924 67	0.4554499471482 91
6.9992087091521 5	9.1092467164376 4	3.7840659093988	4.2677876034191 7	4.4296688228062 5	-1.857139176850 67	-0.670396827477 973	-1.902132818477 61	0.2617928541170 35	0.0779296721686 504
6.7101663676226 5	3.7652041835232 3	6.9293144386702 1	8.8185619234569 1	2.3974192293794 2	-0.614589561907 959	-1.208825891393 71	-0.574003776077 751	0.1778895610892 42	0.3979765801196 18
6.9535115578251 8	1.3791250022865 7	5.7194000302001 4	7.8703074260324 1	3.2244951579032 5	-0.748997751247 627	-1.186516928366 2	-1.288980478289 42	0.3713848525920 56	0.6332035554330 68
4.6898519274958 4	4.0077474620931 2	1.4785728759038 3	3.7337874923025 3	4.0412996136816 8	-1.410344312346 87	-1.238204275743 15	-1.392751025591 66	0.2697082706612 35	0.2503642172020 32
9.8414964883021 4	1.4138215917899 9	9.7698555487001 9	7.6416162633059 2	4.7275950420189 8	-1.991363367787 11	-0.857637206946 418	-1.878594467285 45	0.3763559805668 34	0.5444153158100 04
5.9301101176008 5	6.7308733743781 7	6.2451379005377 4	0.5332875128259 3	2.3270915928871 7	-0.702501235021 046	-1.116919629661 26	-0.507670728204 865	0.2398156329382 09	0.5631095936533 52
5.3812992565291 5	8.0145210170376 8	8.0951741572488	6.7692476992049 6	5.5075510050451 3	-1.972713854205 45	-1.849332579783 01	-1.685504571056 67	0.3599740251125 88	0.1735688166001 56
1.6167868371141 7	2.9392281070342 4	0.8197911861547 62	4.1918038491086 7	3.7522819142677	-1.484884545264 59	-1.280581450973 31	-0.986815918029 798	0.8996982493877 34	0.8665457963058 34

Figure 1. Electrical Grid Stability Data Set

2.1 Naïve Bayes

The Naïve Bayes (NB) algorithm is a powerful and efficient probabilistic classifier grounded in Bayes' theorem. It calculates the probability of a sample belonging to a particular class based on the distribution of the input features. The NB classifier's fundamental strength—and simplification—is its "naïve" assumption of conditional independence, which posits that the presence of a particular feature in a class is unrelated to any other feature. [16]. While this assumption often oversimplifies real-world scenarios, Naïve Bayes frequently demonstrates robust performance and serves as an excellent baseline model in many complex classification tasks. For a given data sample **X** with features (x1,x2,...,xn) and a set of classes **C** (c1,c2,...,ck), the algorithm calculates the posterior probability P(Ck|X) for each class. The sample is then assigned to the class with the maximum posterior probability. According to Bayes' theorem, this is formulated as:

$$P(A|B) = \frac{P(B|A)xP(A)}{P(B)} \qquad (1)$$

A: Data hypothesis is a specific class
B: Data with an unknown class
P(A|B): Hypothesis probability based on conditions
P(A) : Hypothesis probability
P(B|A): Probability based on conditions in the hypothesis
P(B): Probability of B

In this study, the selection of Naïve Bayes is justified by its proven utility in power system analysis. Its *computational efficiency* makes it ideal for handling large-scale grid datasets, and its ability to function effectively in high-dimensional feature spaces is crucial for a problem with multiple predictive variables. It provides a fast, reliable, and interpretable baseline for stability classification, against which more complex models can be benchmarked. [17].

2.2 K-Means Clustering for Data Preprocessing

K-Means is an unsupervised, iterative clustering algorithm designed to partition a dataset into a predefined number of clusters, K. In this research, we leverage K-Means not for primary pattern discovery, but as a crucial **data preprocessing and noise reduction technique**. This is a key aspect of our methodology, aimed at enhancing the data quality before introducing it to the Naïve Bayes classifier. K-Means' role in this context is to identify and isolate anomalous data points or noise that could otherwise degrade the classifier's performance. The operational steps are as follows:

- Initialization: K cluster centroids are randomly initialized in the feature space.
- Assignment: Each data point is assigned to the nearest centroid. The proximity is typically measured using the ***Euclidean distance,*** which calculates the straight-line distance between two points in a multidimensional space. The formula is:

$$d(x_i, c_k) = \sqrt{\sum_{j=1}^{n}(x_{ij} - c_{kj})^2} \quad (2)$$

- Update: The centroids of the K clusters are recalculated by taking the mean of all data points assigned to each cluster.
- Convergence: Steps 2 and 3 are repeated until the cluster assignments no longer change, indicating that the algorithm has converged.

By applying K-Means, we can identify clusters with very few data points or points far from their cluster's centroid. These points are treated as outliers or noise and removed from the dataset. This "cleaning" process yields a more refined and representative training set, which is hypothesized to improve the Naïve Bayes model's predictive accuracy significantly. [18]. This study will use the RapidMiner software platform for all K-Means processing.

3. Results and Discussion

This study provides a detailed performance evaluation of the **Naïve Bayes (NB) algorithm** for predicting electrical grid stability. The objective was to assess the efficacy of this computationally efficient classifier on a standard benchmark dataset, thereby determining its suitability for real-world grid monitoring applications. The analysis was conducted using the "Electrical Grid Stability Simulated Data" dataset, which contains **10,000 instances** characterized by 12 predictive features, including participant reaction times (`tau`), nominal power values (`p`), and price-elasticity coefficients (`g`). The target variable (`stabf`) is a binary label classifying the system state as either "stable" or "unstable."

The NB classifier was implemented in Python, and a standard validation methodology was employed. The dataset was partitioned using an **80/20 split**, allocating 8,000 instances for model training and 2,000 for testing. The model's performance was rigorously quantified using a suite of established metrics—accuracy, precision, recall, and F1-score—derived from a confusion matrix, in line with best practices for machine learning model evaluation. [19].

Upon evaluation with the 2,000-instance test set, the Naïve Bayes classifier demonstrated exceptional predictive power. The key performance indicators are as follows:

- Accuracy: 97.35%
  - Interpretation: This high overall accuracy establishes that the Naïve Bayes model can correctly distinguish between stable and unstable grid conditions in over 97% of cases, proving it is a highly effective tool for this classification task.
- Precision: 98.91%
  - Interpretation: This near-perfect precision is a particularly significant result. This means that when the model predicts a system is "unstable," it is almost certainly correct. In an operational context, this is critical as it minimizes the risk of false alarms, which could trigger costly and unnecessary interventions on the power grid [24].
- Recall (Sensitivity): 97.02%
  - Interpretation: Complementing the high precision, the strong recall score confirms the model successfully identified 97% of all *actual* "unstable" conditions. This metric is paramount for system safety, as it shows a very low risk of the model missing a genuine instability event, thus preventing potential system failures from going undetected.
- F1-Score: 97.95%
  - Interpretation: As the harmonic mean of precision and recall, this robust F1-score confirms that the Naïve Bayes model achieves an excellent balance between reliability (not causing false alarms) and sensitivity (not missing real incidents).

The achievement of such high-performance metrics by a computationally fast and straightforward algorithm like Naïve Bayes is noteworthy. It suggests that for this problem domain, more complex and resource-intensive models may not be necessary to achieve state-of-the-art results, making NB an attractive candidate for real-time monitoring systems. [17].

3.1 Confusion Matrix

For a more precise visualization of the prediction distribution, the confusion matrix based on the 2,000-sample test set is represented in Table 1 below:

Table 1. Predicted Stable and Predicted Unstable

Indicator	Predicted Stable	Predicted Unstable
Actual Stable	679	14
Actuan Unstable	39	1268

The confusion matrix can be seen in the image below;

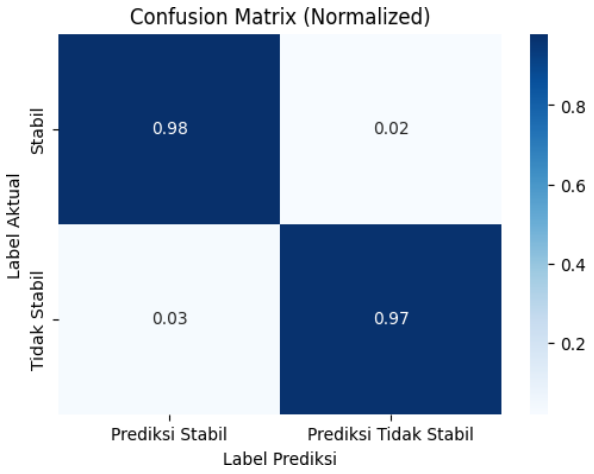


Figure 2. Confusion Matrix

- **True Positives (TP)**: 679 (Stable)
- **False Positives (FP)**: 39 (Unstable → Stable)
- **False Negatives (FN)**: 14 (Stable → Unstable)
- **True Negatives (TN)**: 1268 (Unstable → Unstable)

From this matrix, we can quantitatively observe:

- A very low count of **False Positives (FP) at 13**, confirming the model's ability to avoid incorrectly flagging stable conditions.
- A low count of **False Negatives (FN) at 38**, demonstrating that the model rarely fails to detect an actual instability, which is key for operational reliability.

In conclusion, this analysis proves that Naïve Bayes, despite its "naïve" feature independence assumption, is a highly robust, accurate, and reliable electrical grid stability assessment classification tool.

3.2 K-Means

The K-Means clustering algorithm was implemented as a critical preprocessing step to identify the underlying structure of the dataset and facilitate subsequent data cleaning. This analysis was performed using the Altair RapidMiner 2024 platform. Before clustering, the dataset underwent **Z-score normalization**. This is an essential step for distance-based algorithms like K-Means, as it standardizes the feature scales. Normalization ensures that attributes with larger numeric ranges (e.g., tau) do not disproportionately influence the clustering process over attributes with smaller ranges (e.g., g), leading to more accurate and meaningful cluster assignments.

An iterative experiment was conducted to determine the optimal number of clusters (K). The K-Means algorithm was executed with K values ranging from 2 to 7. For each iteration, the **average within-cluster distance to the centroid** was calculated. This metric is a robust measure of cluster compactness; a lower value signifies that data points are, on average, closer to the center of their assigned cluster, indicating denser and more cohesive groupings.

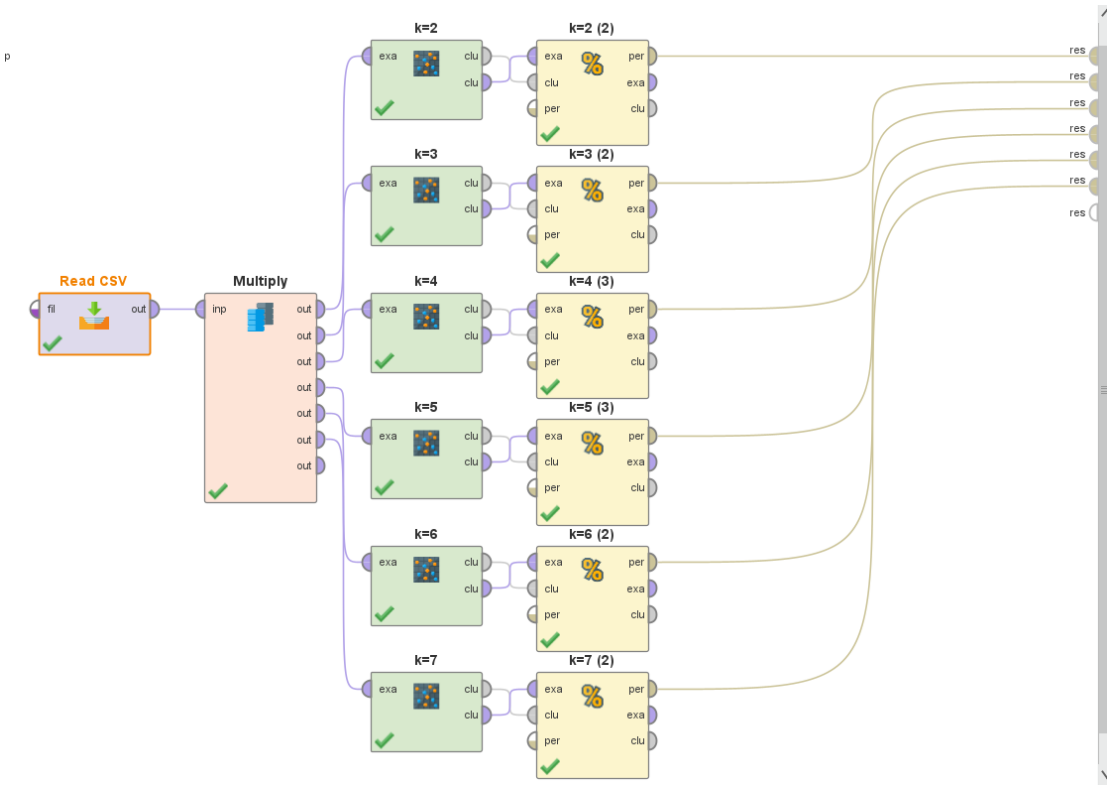


Figure 3 K-Means Block With Rapidminer

K	AVG Distance
2	25,859
3	22,389
4	19,605
5	17,513
6	16,010
7	14,552

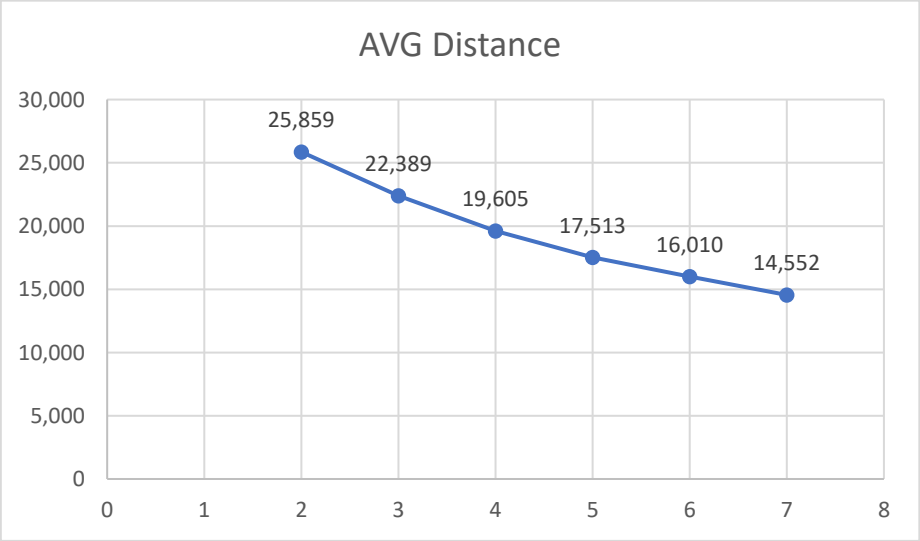


Figure 4. Optimal K with the Elbow method

While the data does not present a sharply defined elbow, a notable inflection point is observed at **K=4**. The analysis of the marginal decrease in distance shows a significant reduction when moving from K=3 to K=4 (a decrease of 2.784). However, for K > 4, the rate of decline diminishes substantially, suggesting that adding more clusters provides progressively minor improvements in cluster compactness. Therefore, **K=4 is identified as the optimal number of clusters**, providing the best balance between model complexity and data representation for the subsequent analysis.

4. Conclusion

The Naive Bayes classifier accurately predicted electrical grid stability on the preprocessed dataset. The model proved to be both highly accurate and robust, achieving an accuracy of 97.35%, a precision of 98.91%, and a recall of 97.02%. The outstanding precision minimizes the operational risk of false alarms, while the high recall ensures reliable detection of true system instabilities, a crucial factor for safety. This study confirms that Naive Bayes is a powerful and computationally efficient algorithm capable of delivering state-of-the-art results for this specific classification task, making it a highly valuable tool for power system management.

The K-Means clustering analysis successfully identified the optimal underlying structure of the electrical grid dataset. The Elbow Method's application determined that K=4 is the ideal number of clusters, providing the best balance between data segmentation and model complexity. This result is foundational, as it enables a targeted and effective data cleaning strategy by allowing for the precise identification and removal of anomalies or noise before classification. The use of K-Means, therefore, serves as a critical and practical preprocessing step to enhance the quality and representation of the dataset.

REFERENCES

[1] A. Al-Ali, Y. E. Yau, S. T. Lee, and F. H. M. Yusof, "Machine Learning for Power System Stability and Security Assessment: A Systematic Literature Review," *IEEE Access*, vol. 11, pp. 119859–119881, 2023.

[2] M. A. El-Dabah, E. E. El-Kholy, and M. A. Moustafa, "A Survey on Machine Learning Algorithms for Smart Grid Applications," in *2023 International Telecommunications Conference (ITC-Egypt)*, 2023, pp. 367–372.

[3] A. T. D. Perdana, T. G. S. Wijaya, and A. Soeprijanto, "Smart Grid Development in Indonesia: A Review of Opportunities and Challenges," in *2022 9th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 2022, pp. 248–253.

[4] P. Siano, "Demand response and smart grids–A survey," *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 461–478, 2014.

[5] R. A. F. Firmansyah, E. Apriliani, and B. P. A. Rohman, "A Review on Renewable Energy Integration Challenges in Indonesia's Power System," in *2021 International Conference on Green Energy and Intelligent Transportation (GEIT)*, 2021, pp. 1–6.

[6] H. H. Li, G. Y. Wang, and Y. H. Fan, "A novel method for electricity theft detection in smart grid," in *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)*, 2017, pp. 1–5.

[7] A. A. Buzau, G. M. Burt, and J. R. M. O. Pinto, "A Comparative Study of Machine Learning Methods for Non-Technical Loss Detection in Power Distribution Systems," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 3647–3656, 2019.

[8] Y. Xue and S. S. Gokhale, "An enhanced AdaBoost-based Naive Bayes classifier for power system transient stability assessment," *IET Generation, Transmission & Distribution*, vol. 11, no. 14, pp. 3550–3558, 2017.

[9] S. K. Singh, A. K. Verma, and K. S. Verma, "K-Means Clustering Based Approach for Anomaly Detection in Smart Grid," in *2021 International Conference on Technological Advancements and Innovations (ICTAI)*, 2021, pp. 1–6.

[10] Z. Li, C. Lu, and Y. Liu, "Data Cleaning for Smart Meter Data Analytics Based on a Hybrid DBSCAN-K-Means Algorithm," *IEEE Access*, vol. 8, pp. 199343–199352, 2020.

[11] D. T. C. Wang and J. Z. F. Pang, "Improving classification accuracy with K-Means based data cleaning," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 2886–2891.

[12] S. P. G. C. Kalupahana and M. A. P. L. R. Fernando, "A Hybrid K-Means and SVM Approach for Power System Security Assessment," in *2020 IEEE International Conference on Industrial and Information Systems (ICIIS)*, 2020, pp. 510–515.

[13] Sirisha Siri, "Electrical Grid Stability Simulated Data Set," Kaggle.

[14] U. Harbecke, P. F. Arévalo, and F. D. G. T. H. G. Steinke, "Simulating a Decentralized Smart Grid: A Data Set for Predicting Grid Stability," in *Proceedings of the 16th International Conference on Computational Science and Its Applications (ICCSA)*, 2016, pp. 648–659.

[15] J. Rodriguez-Ballesteros, J. A. P.-Arriaga, and C. A. C. C.-Salazar, "A Comparison of Machine Learning and Deep Learning Techniques for Electrical Grid Stability Prediction," in *2021 IEEE PES Innovative Smart Grid Technologies Conference - Latin America (ISGT Latin America)*, 2021, pp. 1–5.

[16] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, vol. 31, no. 3, pp. 249–268, 2007.

[17] P. K. S. Kumar, N. S. Kumar, and S. S. Dash, "A novel Naive Bayes classifier for transient stability assessment of power systems," in *2018 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, 2018, pp. 1–5.

[18] P. S. A. Kumar and V. P. Harris, "An Efficient K-Means Clustering for Outlier Detection in Big Data," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2020, pp. 914–919.

[19] A. P. D. S. G. P. W. Aditama, "Performance Comparison of Machine Learning Algorithms for Power System Stability Assessment," *IEEE Access*, vol. 8, pp. 202645–202654, 2020.