cigre Turkey

TEPES
TURKISH JOURNAL OF ELECTRICAL POWER AND ENERGY SYSTEMS

TEPES, Vol. 5, Issue. 3, 168-175, 2025 **DOI:** 10.5152/tepes.2025.25038

RESEARCH ARTICLE

Machine Learning–Based Fault Diagnosis in Solar Photovoltaic Systems Using Data Balancing Techniques

Merve Demirci

Department of Electrical and Electronics Engineering, Kafkas University, Kars, Türkiye

Cite this article as: M. Demirci, "Machine learning–based fault diagnosis in solar photovoltaic systems using data balancing techniques," *Turk J Electr Power Energy Syst.*, 2025; 5(3), 168-175.

ABSTRACT

As solar energy adoption continues to rise, the demand for reliable photovoltaic (PV) systems has increased significantly. Ensuring the efficient and secure operation of PV systems requires accurate fault detection, making fault diagnosis a critical research area. This study investigates the diagnosis of short-circuit faults in PV systems by integrating machine learning algorithms with data balancing techniques. Four classifiers (Random Forest, CatBoost, Extreme Gradient Boosting, and Light Gradient Boosting Machine (LGBM)) were employed for fault classification, while Synthetic Minority Oversampling Technique (SMOTE), Random Oversampling, and Adaptive Synthetic Sampling were used to address class imbalance. Two datasets were analyzed: Dataset-1 with 11 features and Dataset-2 with 13 features. For Dataset-1, LGBM achieved the highest accuracy (79.28%) on the imbalanced data, which improved to 86.59% after applying SMOTE. By incorporating two additional features in Dataset-2, fault diagnosis accuracy increased to 98.57% on the imbalanced data and reached 100% when balanced with SMOTE. These findings demonstrate that combining LGBM with SMOTE significantly enhances short-circuit fault detection performance in PV systems.

Index Terms—Fault diagnosis, machine learning, short circuit fault, solar photovoltaic system

I. INTRODUCTION

Rapid industrial development, population growth, and the resulting rise in energy consumption have significantly increased global energy demand. This growing demand has intensified interest in renewable energy sources such as wind, tidal, geothermal, hydroelectric, biomass, and solar power. Renewable sources are widely adopted due to their environmentally friendly and reliable nature, and unlike fossil fuels, they do not contribute to greenhouse gas emissions. Among these sources, solar energy is harnessed through photovoltaic (PV) panels and converted into electricity for end users. The rapid advancements in solar technologies, supportive government policies, and the declining cost of panel installation have further accelerated the adoption of solar energy [1, 2].

According to the IRENA Renewable Energy Statistics 2025 report, global solar PV capacity continues to grow steadily, increasing from approximately 1407 GW in 2023 to around 1859 GW in 2024 [3]. The global change in PV capacity over the past nine years is illustrated in Fig. 1. Investments in solar PV also expanded significantly, rising from 35 billion USD in 2022 to 83 billion USD in 2023 [4]. As of June 2025, Turkey's total installed electricity generation capacity reached 119 632 MW, of which 19.2% is supplied by solar energy [5].

Photovoltaic systems are composed of multiple components, each of which can fail due to physical, environmental, or electrical factors [6–8]. Fast and accurate fault diagnosis is essential to prevent efficiency losses, ensuring that power generation and system safety remain unaffected. With the rapid growth of solar investments, the number of studies on PV fault detection has significantly increased and diversified [6, 7].

A review of the literature reveals that artificial intelligence (AI)-based techniques are frequently employed for fault detection and classification, utilizing historical data in combination with expert knowledge. To enhance the diagnostic performance of these methods, researchers often apply data preprocessing, augmentation, reduction, and feature engineering techniques.

For instance, Lazzeretti et al. [9] developed a monitoring system to collect real-time and historical data and proposed a recursive

Corresponding author: Merve Demirci **E-mail:** merve.demirci@kafkas.edu.tr

Received: September 6, 2025
Revision Requested: September 11, 2025
Last Revision Received: September 20, 2025
Accepted: September 22, 2025
Publication Date: October 20, 2025

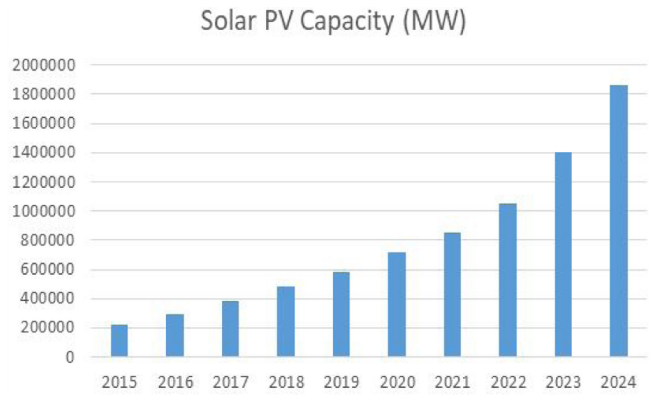

Fig. 1. Change in solar PV capacity in the world over the years [3].

method for fault detection. They employed an artificial neural network (ANN) to detect and classify short-circuit, open-circuit, distortion, and shading faults using PV panel temperature and irradiance values. Similarly, Quiles-Cucarella et al. [10] applied ensemble algorithms, ANNs, and machine learning methods to detect and classify seven different PV fault types, achieving the highest accuracy with the Bagged Trees algorithm. El-Katheri et al. [11] combined I–V and P–V curve analysis with ANNs to identify DC-side faults in PV systems, testing three scenarios with different input datasets.

Dhimish et al. [12] explored different ANN structures and fuzzy logic models to detect faults such as partial shading, defective PV arrays, faulty modules, and Maximum Power Point Tracking errors, aiming to mitigate their negative impact on system performance. Yang et al. [13] used the Random Forest (RF) algorithm to classify four distinct fault types. Since their dataset was imbalanced, they employed Modified Independent Component Analysis for oversampling and undersampling, achieving superior performance compared to other classifiers. Yi et al. [14] proposed a two-stage Support Vector Machine model to classify line-to-line faults based on current and voltage data, demonstrating reliable detection even under low-irradiance and high-impedance fault conditions.

The aim of this study is to investigate the effects of different data balancing methods used to eliminate the imbalance in datasets for the accurate classification of short circuit faults occurring in solar PV systems on the performance of classification algorithms. Faults occurring in solar PV systems were classified into four different cases. For this purpose, a dataset consisting of 700 data points from the literature was used [15]. The data numbers of fault cases in the dataset are imbalanced. Since this imbalance in the dataset is known to affect the performance of classification algorithms, different data balancing methods were used. Random Oversampling Minority Class (ROM) and Synthetic Minority Oversampling Technique (SMOTE) methods, which aim to equalize class labels by augmenting the data in the minority class according to their own rules, were used. In addition, Adaptive Synthetic Sampling (ADASYN) method was used, which aims to augment the data in the difficult-to-learn datasets without equalizing the class labels. Datasets created with these methods were used with classification algorithms such as RF, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LGBM), and CatBoost algorithm (CA) to obtain comparative fault diagnosis accuracies and performance metrics.

Main Points

- It has been shown that imbalance in the dataset negatively affects classification performance, but data balancing methods such as Synthetic Minority Oversampling Technique, Random Oversampling, and Adaptive Synthetic Sampling reduce this effect and improve accuracy rates.
- The performances of Random Forest, Extreme Gradient Boosting, Light Gradient Boosting Machine, and CatBoost algorithms were compared and significant performance improvements were obtained by applying data balancing methods.
- High accuracy rates were achieved in classifying four different fault conditions demonstrating that the methods are applicable in fault detection in photovoltaic systems

This study is organized as follows. The second section summarizes the short-circuit faults that occur in solar PV systems and are classified in this study. The machine learning methods used for classification, including RF, XGBoost, LGBM, and CA, as well as the data balancing methods ROM, SMOTE, and ADASYN, are briefly described in the third section, Materials and Methods. The fourth section presents the analyses and their results, and the fifth section presents the study's conclusions.

Solar Photovoltaic Faults

Environmental, physical, and electrical failures may occur in solar PV systems. Early diagnosis of these failures is crucial to ensure efficiency, safety, and cost-effectiveness. Accurate fault detection enables timely intervention, thereby minimizing failures and reducing their negative impact on system performance [6–8].

Environmental faults occurring in PV systems are defined as partial shading, hotspot faults, and bypass diode faults [6, 16, 17]. Physical faults include microcracks and fractures that occur during production, transportation, and assembly, and internal corrosion caused by external factors such as humidity [7, 8]. Electrical faults include faults occurring in the inverter or PV array components in PV systems [18, 19]. In this study, a classification process was performed to determine the fault type using data from String Fault, String to String Fault, and String to Ground Fault situations occurring in PV systems.

A string fault is a fault that occurs in a structure consisting of series-connected PV modules. This fault occurs due to the failure of any module within the string, or breaks or damage to electrical connections. If not detected and repaired, the string cannot produce electricity, reducing system efficiency.

A string-to-string fault is a short-circuit fault that occurs between two different strings. It occurs due to incorrect connections during the installation phase or insulation defects in the connecting cables. It causes current flow within the circuit, creating the risk of overheating and fire.

A string-to-ground fault occurs between a string and the ground. It occurs when insulation breaks down and ground contact occurs. A leakage current from the system to ground occurs, posing a risk of shock to personnel if undetected.

II. METHODS

A. Classification Algorithms

1) Random Forest:
It is one of the ensemble learning algorithms proposed by L. Breiman in 2001 [20]. It is a method that aims to reach a conclusion by combining the classification results obtained with different numbers of decision tree structures. It combines the results obtained with the decision trees with the voting method and determines the result that reaches the maximum number of votes as the final result [21].

2) Extreme Gradient Boosting:
It is an improved version of the Gradient Boosting algorithm. It applies decision trees sequentially, and the next decision tree aims to learn from the errors of the previous algorithm and perform classification. Its advantages are fast, flexible, and low overfitting [22, 23].

3) Light Gradient Boosting Machine:
It is an effective and open-source Gradient Boosting algorithm introduced by Microsoft in 2017. It is a scalable algorithm that can make fast decisions using decision trees. Its advantages include fast operation, high accuracy, and low memory usage [24].

4) CatBoost Algorithm:
Derived from the words "Categorical" and "Boosting," the Gradient Boosting algorithm is a widely used algorithm in R and Python developed by Yandex. It allows for accurate processing of class labels and variables across multiple decision trees with fewer parameters, resulting in highly accurate results. Its operations solve the overfitting problem and produce fast results [25, 26].

B. Data Balancing Methods

1) Random Oversampling Minority Class:
The main goal of this method is to label the classes by randomly increasing the number of data belonging to the minority class. Its advantages are that it is fast and easy to implement and does not cause data loss. However, it is a disadvantage that the same data must be repeated to equalize the class labels. This can lead to overfitting problems in classification algorithms [27, 28].

2) Synthetic Minority Over-sampling Technique:
It is one of the most commonly preferred oversampling methods. It generates synthetic new samples among the samples in the minority class according to the nearest neighbor rule. This process is repeated until the number of minority class data equals the number of majority class data. The generated samples differ from the existing samples. Therefore, it does not cause an overfitting problem. By generating different data, it provides data diversity, which increases model performance. However, its disadvantages include the possibility of outliers and the computational cost [29, 30].

3) Adaptive Synthetic Sampling:
In this method, a weighted distribution is used based on the learning difficulty of the dataset when augmenting the data for the minority classes, and synthetic data is generated from these difficult examples. Data that are very close to the majority class in the dataset, creating classification difficulties, are augmented. This allows the classification model to learn better at the boundary values between classes. In this method, the minority class data is augmented as needed, without expecting it to equal the majority class. By focusing on difficult examples, it increases model learning and produces diverse data, thus reducing the risk of overfitting. Its disadvantages include computational cost and, in some cases, the generation of complex data [31, 32].

C. Performance Metrics

Classification metrics are used in classification processes to evaluate classification success and present model performance. True Positive (TP) represents the number of correctly detected faults, while True Negative (TN) represents the number of non-faulty cases. False Positive (FP) represents the detection of non-faulty cases as faults, while False Negative (FN) represents undetected faults. The Recall parameter is a measure of the ability of the classification algorithm used to detect true faults, presented in (1). Precision is the proportion of true positives among detected faults, presented in (2). The F1 Score is the harmonic average of the Precision and Recal values [33].

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2}$$

IV. RESULTS

In this study, machine learning classification algorithms were employed to diagnose faults in solar PV systems. To address class imbalance in the dataset and enhance algorithm performance, data augmentation techniques including SMOTE, Random Oversampling (ROS), and ADASYN were applied. Fault diagnosis performance was assessed using both the original (imbalanced) and balanced datasets, allowing for an evaluation of the impact of data balancing on classification accuracy. The flowchart of the study is illustrated in Fig. 2.

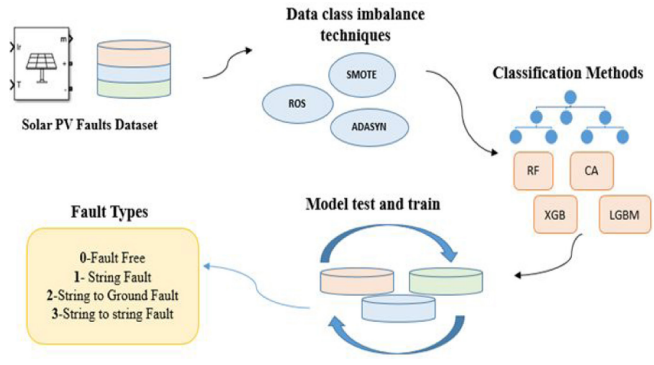

Fig. 2. Flow chart of study.

A. Dataset

The dataset used in this study consists of 700 data points and classifies the solar PV system into four different states: Fault-free, String fault, String to ground fault, and String to string fault [15]. 11 features obtained directly from the PV system in the original data set are used and detailed information for the data set is presented in Table I. As can be seen from Table II, the dataset is unbalanced because the number of data points for fault types is unequal.

In the dataset, the current values I_1, I_2, I_3, and I_4 represent measurements taken in the same direction from ammeters located at the beginning and end of each string. To enhance the dataset, the current differences between the beginning and end of the strings were also calculated and included, resulting in an extended dataset with 13 features.

B. Performance Analysis and Results

In this study, a dataset containing 11 features was initially used. Because this dataset had an imbalanced data class distribution, it was balanced using various data augmentation methods and used for fault classification. The data was augmented using the data augmentation methods ROS, SMOTE, and ADASYN to reach the appropriate number. Because the ROS and SMOTE methods aimed to have equal class labels, each class label was augmented to 223. In the ADASYN method, augmentations were made as necessary to account for data that created learning difficulties inherent in the method's application principles, resulting in a total dataset of 870 data points. Table II shows the results of the balancing process for Dataset-1.

TABLE I.
DATASET CONTENT [15]

Dataset Content		
I_1	String 1 top average current	
I_2	String 1 bottom average current	
I_3	String 2 top average current	
I_4	String 2 bottom average current	
I_5	String 3 top average current	
I_6	String 3 bottom average current	
I_{total}	Total average current	
V_{dc}	Total average DC Voltage	
P_{dc}	Total average DC Power	
T	Temperature	
IR	Radiation	
Fault Type		**Count**
0	Fault free	123
1	String fault	174
2	String to ground fault	178
3	String to string fault	223

TABLE II.
DATA BALANCING RESULTS FOR DATASET-1

	Count	
Fault Type	**ROS/SMOTE**	**ADASYN**
0	223	225
1	223	223
2	223	211
3	223	211
Total	892	870

After balancing the dataset using data augmentation methods, the training and testing of the classification algorithms were completed, and the results were obtained. The classification results of the algorithms for fault diagnosis are presented in Table III. In fault diagnosis performed with an imbalanced dataset, the highest fault diagnosis performance was achieved by the LGBM algorithm with 79.28%. On the dataset balanced with the SMOTE method, the LGBM algorithm also achieved the highest performance with 86.59%. On the dataset processed with the ROS method, the best performance belonged to the XgBoost and RF algorithms with 82.68%. On the dataset augmented with ADASYN, the highest performance was achieved by the

TABLE III.
FAULT DIAGNOSIS ACCURACY RESULTS FOR DATASET-1

	Imbalanced Dataset	**SMOTE**	**ROS**	**ADASYN**
LGBM	79.28	86.59	78.77	78.73
XGB	73.57	81.56	82.68	81.60
RF	75.71	77.65	82.68	75.86
CA	75.00	78.77	81.00	79.31

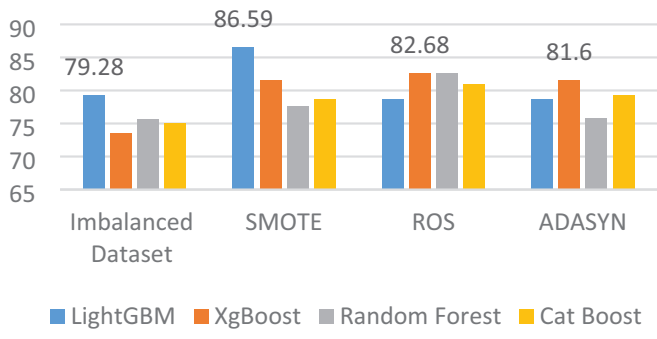

Fig. 3. Fault diagnosis accuracy for Dataset-1.

TABLE IV.
PERFORMANCE METRICS FOR RAW DATASET-1 AND SMOTE METHOD IN LGBM ALGORITHM

		LightGBM			
Imbalanced Dataset		**Precision**	**Recall**	**F1-Score**	**Support**
	0	0.68	0.60	0.64	25
	1	0.92	0.94	0.93	36
	2	0.83	0.83	0.83	35
	3	0.72	0.75	0.73	44
	Accuracy			0.79	140
SMOTE		LightGBM			
		Precision	Recall	F1-Score	Support
	0	0.87	0.89	0.88	45
	1	0.95	0.91	0.93	44
	2	0.79	0.82	0.80	45
	3	0.80	0.80	0.80	45
	Accuracy			0.86	179

XgBoost algorithm with 81.6%. The results are presented comparatively in Fig. 3.

In the fault diagnosis operations performed for Dataset-1, the LGBM algorithm showed the best performance in diagnosis with an unbalanced dataset. The highest performance in the balancing operation performed with data balancing methods was achieved with the LGBM algorithm in the SMOTE method. Performance metrics are presented in Table IV. In unbalanced dataset conditions, the results show that it diagnosed the string fault, labeled 1, with the highest accuracy, while it diagnosed the Fault Free condition with the lowest accuracy. The results show that increasing the data with the SMOTE method increases the algorithm's fault diagnosis accuracy. Performance metrics demonstrate that the Fault Free condition, which performed least well in fault diagnosis with an imbalanced dataset, was better diagnosed by increasing the data count from 25 to 45. It has been shown that increasing and equalizing the data belonging to classes in the dataset with the SMOTE algorithm also improves performance for all classes.

In the solar PV system, the difference between the current values measured at the beginning and end of the strings in strings 1 and 2 was also added to the dataset as a feature. Thus, the number of features in the dataset was increased to 13, creating Dataset-2, and fault diagnosis was performed using algorithms. Because Dataset-2 was an imbalanced dataset, the dataset content was first increased using data balancing methods. In the augmentation process performed with the ROS and SMOTE methods, the aim was to ensure equal data for each fault type, so 223 data points were augmented for each fault. However, because the augmentation process was performed in the ADASYN method according to the learning difficulty between the feature and fault class, a total of 876 data points were obtained. The number of data points obtained for each fault as a result of the data augmentation methods is presented in Table V.

TABLE V.
DATA BALANCING RESULTS FOR DATASET-2

	Count	
Fault Type	**ROS/ SMOTE**	**ADASYN**
0	223	223
1	223	222
2	223	221
3	223	210
Total	892	876

TABLE VI.
FAULT DIAGNOSIS ACCURACY RESULTS FOR DATASET-2

	Imbalanced Dataset	**SMOTE**	**ROS**	**ADASYN**
LGBM	98.57	100	98.32	98.86
XGB	97.14	97.76	96.08	95.45
RF	97.14	97.2	98.32	98.29
CA	95.71	98.32	98.32	97.72

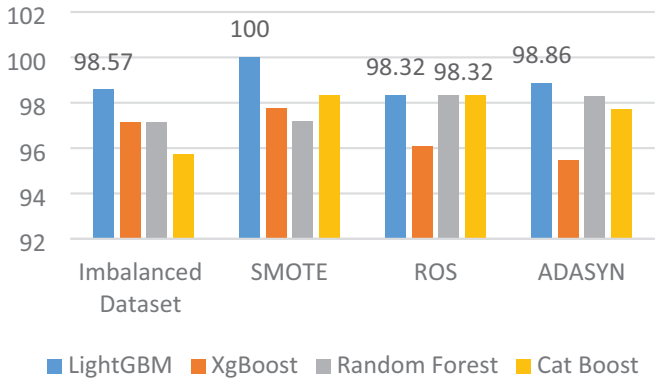

Fig. 4. Fault diagnosis accuracy for Dataset-2.

Fault diagnosis was performed using Dataset-2, which contained 13 features. When the fault diagnosis accuracies of the classification algorithms were first obtained using an imbalanced dataset, it was observed that the addition of two features to the dataset had significant effects on classification performance. The classification results of the algorithms for fault diagnosis are presented in Table VI. The highest performance achieved by the LGBM algorithm was 78.29%, while with these features, the performance increased to 98.57%. It was observed that the performances of all algorithms increased, with the lowest performance belonging to the CA at 95.71%. With the dataset balanced with the SMOTE method, the fault diagnosis accuracy of the LGBM algorithm was 100%. When the results of the other classification algorithms were examined, higher performances were achieved compared to Dataset-1. Using the dataset augmented with the ROS method, the LGBM algorithm achieved the highest accuracy of 98.32%. On the dataset processed with the ADASYN method, the highest accuracy was also achieved by the LGBM algorithm at 98.86%. Fig. 4 presents the results comparatively.

The results of the classification process performed on the unbalanced dataset and the balanced dataset using the SMOTE method for fault diagnosis on Dataset-2 are presented in Table VII. In the classification process performed on the unbalanced dataset, the highest diagnostic accuracy was achieved with LGBM. The algorithm's performance was lower in the Fault Free case and highest in the String-to-string fault case. This is because the data in the Fault Free case is a minority compared to the other cases. When fault diagnosis was performed on the balanced dataset using the SMOTE method, the LGBM algorithm performed best. An examination of the obtained results reveals that the diagnostic accuracy for states 0 and 1 increases with the increase in data.

The performance metrics for the CA, which performed best in fault diagnosis using the dataset balanced with the ROS method, are presented in Table VIII, and the results for the LGBM algorithm on the dataset augmented with the ADASYN method are presented in Table IX. When the results were examined, it was observed that data balancing methods increased the overall fault diagnosis ability of the algorithms in fault diagnosis using Dataset-2. When examined with performance metrics, it was observed that the increased data also increased the algorithms' ability to identify individual faults.

IV. DISCUSSION

This study investigates the use of machine learning methods for diagnosing short-circuit faults in solar PV systems. A dataset obtained

TABLE VII.
PERFORMANCE METRICS FOR RAW DATASET-2 AND SMOTE METHOD IN LGBM ALGORITHM

		LightGBM			
Imbalanced Dataset		***Precision***	***Recall***	***F1-Score***	***Support***
	0	0.93	1.00	0.96	25
	1	0.97	0.94	0.96	36
	2	1.00	0.97	0.99	35
	3	1.00	1.00	1.00	44
	Accuracy			**0.98**	140
SMOTE		LightGBM			
		Precision	*Recall*	*F1-Score*	*Support*
	0	0.99	0.99	0.99	45
	1	0.98	0.98	0.98	44
	2	1.00	1.00	1.00	45
	3	1.00	1.00	1.00	45
	Accuracy			**1.00**	179

TABLE VIII.
PERFORMANCE METRICS FOR ROS METHOD IN CATBOOST ALGORITHM

CatBoost				
	Precision	**Recall**	**F1-Score**	**Support**
0	1.00	0.91	0.95	45
1	0.92	1.00	0.96	44
2	1.00	1.00	1.00	45
3	1.00	1.00	1.00	45
Accuracy			0.98	179

from the literature, which categorizes PV systems into four different operating states, was employed. However, the dataset was inherently imbalanced, with unequal numbers of samples across the fault classes. To address this issue, data balancing techniques were applied to preprocess the dataset prior to classification. The impact of these balancing methods on fault diagnosis performance was then evaluated and compared.

To address class imbalance, data balancing and augmentation techniques including ROS, SMOTE, and ADASYN were applied. These balanced datasets were then used with machine learning algorithms such as RF, LGBM, XGBoost, and CA for fault diagnosis, and the corresponding results were obtained.

In the fault diagnosis experiments using Dataset-1, the highest accuracy under the imbalanced condition was obtained with the LGBM algorithm at 79.28%. When data augmentation was applied, SMOTE achieved the best performance with an accuracy of 86.59%. Using the ROS method improved the diagnostic capability of all classifiers, with the highest accuracy observed for XGBoost) and RF at 82.68%. With ADASYN, the best performance was recorded with XGBoost, achieving an accuracy of 81.6%.

In Dataset-2, two additional features were incorporated, which significantly enhanced the classification performance. Using the imbalanced dataset, the LGBM algorithm achieved an accuracy of 98.57%. Among the data augmentation methods, the highest performance was obtained with SMOTE, where LGBM reached 100% accuracy. With the ROS method, an accuracy of 98.32% was achieved by LGBM, RF, and CA. Using ADASYN, LGBM attained a fault diagnosis accuracy of 98.86%.

TABLE IX.
PERFORMANCE METRICS FOR ADASYN METHOD IN LGBM ALGORITHM

LightGBM				
	Precision	**Recall**	**F1-Score**	**Support**
0	0.98	0.98	0.98	45
1	0.98	0.98	0.98	42
2	1.00	1.00	1.00	44
3	1.00	1.00	1.00	45
Accuracy			0.99	176

V. CONCLUSION

The comparison of results clearly demonstrates that data augmentation methods have a significant impact on algorithm performance. Among them, SMOTE consistently provided the greatest improvement across all cases. In contrast, the ROS method simply replicates existing samples, while ADASYN generates new samples for harder-to-learn instances without fully balancing the dataset, leading to relatively lower performance. Overall, considering the dataset and fault types analyzed in this study, the best performance was obtained by combining the LGBM algorithm with the SMOTE method.

In future work, fault diagnosis will be extended to different datasets with varying input parameters and fault types. By increasing the number of available features within the dataset, we will provide more input features for classification algorithms, and the data will be scaled by applying different preprocessing methods. Additionally, hybrid approaches that integrate multiple diagnostic methods will be explored to further enhance classification accuracy and robustness.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – M.D.; Design – M.D.; Supervision – M.D.; Resources – M.D.; Materials – M.D.; DataProcessing – M.D.; Analysis and/or Interpretation – M.D.; Literature Search – M.D.; Writing – M.D.; Critical Review – M.D.

Declaration of Interests: The author has no conflicts of interest to declare.

Funding: The author declare that this study received no financial support.

REFERENCES

1. A. Khoshnami, and I. Sadeghkhani, "Sample entropy-based fault detection for photovoltaic arrays," *IET Renew. Power Gener.*, vol. 12, no. 16, pp. 1966–1976, 2018. [CrossRef]
2. F. Aziz, A. U. Haq, S. Ahmad, Y. Mahmoud, M. Jalal, and U. Ali, "A novel convolutional neural network-based approach for fault classification in photovoltaic arrays," *IEEE Access*, vol. 8, pp. 41889–41904, 2020. [CrossRef]
3. IRENA, "Renewable energy statistics 2025," 2025. Abu Dhabi: International Renewable Energy Agency. Available at: https://gensed.org/wp-content/uploads/2025/07/IRENA_DAT_RE_Statistics_2025.pdf. [accessed: 20 August 2025].
4. "Renewables," 2024, Global Status Report, *A Comprehensive Annual Overview of the State of Renewable Energy*. Available at: https://www.ren21.net/gsr-2024/. [accessed: 20 August 2025].
5. "Republic of Türkiye ministry of energy and natural resources." Available at: https://enerji.gov.tr/bilgi-merkezi-enerji-elektrik. [accessed: 20 August 2025].

6. E. D. Chepp, and A. Krenzinger, "A methodology for prediction and assessment of shading on PV systems," *Sol. Energy*, vol. 216, pp. 537–550, 2021. [CrossRef]
7. T. Cheng, M. Al-Soeidat, D. D. C. Lu, and V. G. Agelidis, "Experimental study of PV strings affected by cracks," *J. Eng.*, vol. 2019, no. 18, pp. 5124–5128, 2019. [CrossRef]
8. L. L. Jiang, and D. L. Maskell, "Automatic fault detection and diagnosis for photovoltaic systems using combined artificial neural network and analytical based methods," In 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, 2015. [CrossRef]
9. A. E. Lazzaretti et al., "A monitoring system for online fault detection and classification in photovoltaic plants," *Sensors (Basel)*, vol. 20, no. 17, p. 4688, 2020. [CrossRef]
10. E. Quiles-Cucarella, P. Sánchez-Roca, and I. Agustí-Mercader, "Performance optimization of machine-learning algorithms for fault detection and diagnosis in PV systems," *Electronics*, vol. 14, no. 9, p. 1709, 2025. [CrossRef]
11. A. A. Al-Katheri, E. A. Al-Ammar, M. A. Alotaibi, W. Ko, S. Park, and H. J. Choi, "Application of artificial intelligence in PV fault detection," *Sustainability*, vol. 14, no. 21, 13815, 2022. [CrossRef]
12. M. Dhimish, V. Holmes, B. Mehrdadi, and M. Dales, "Comparing Mamdani Sugeno fuzzy logic and RBF ANN network for PV fault detection," *Renew. Energy*, vol. 117, pp. 257–274, 2018. [CrossRef]
13. N. C. Yang, and H. Ismail, "Robust intelligent learning algorithm using random forest and modified-independent component analysis for PV fault detection: In case of imbalanced data," *IEEE Access*, vol. 10, pp. 41119–41130, 2022. [CrossRef]
14. Z. Yi, and A. H. Etemadi, "Line-to-line fault detection for photovoltaic arrays based on multiresolution signal decomposition and two-stage support vector machine," *IEEE Trans. Ind. Electron.*, vol. 64, no. 11, pp. 8546–8556, 2017. [CrossRef]
15. S. S. Ghoneim, A. E. Rashed, and N. I. Elkalashy, "Fault detection algorithms for achieving service continuity in photovoltaic farms," *Intell. Autom. Soft Comput.*, vol. 29, no. 3, pp. 467–479, 2021. [CrossRef]
16. M. Ma, H. Liu, Z. Zhang, P. Yun, and F. Liu, "Rapid diagnosis of hot spot failure of crystalline silicon PV module based on IV curve," *Microelectron. Reliab.*, vol. 100, 113402, 2019. [CrossRef]
17. R. G. Vieira, F. M. de Araújo, M. Dhimish, and M. I. Guerra, "A comprehensive review on bypass diode application on photovoltaic modules," *Energies*, vol. 13, no. 10, p. 2472, 2020. [CrossRef]
18. Y. Zhao, L. Yang, B. Lehman, J. J. F. de Palma, and R. Lyons, "Decision tree-based fault detection and classification in solar photovoltaic arrays," in Proceedings of the 27th Annual IEEE Applied Power Electronics Conference and Exposition, 2012. [CrossRef]
19. M. K. Alam, F. Khan, J. Johnson, and J. Flicker, "A comprehensive review of catastrophic faults in PV arrays: Types, detection, and mitigation techniques," *IEEE J. Photovolt.*, vol. 5, no. 3, pp. 982–997, 2015. [CrossRef]
20. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. [CrossRef]
21. M. Demirci, H. Gözde, and M. C. Taplamacioglu, "Improvement of power transformer fault diagnosis by using sequential Kalman filter sensor fusion," *Int. J. Electr. Power Energy Syst.*, vol. 149, 109038, 2023. [CrossRef]
22. G. Abdurrahman, and M. Sintawati, "Implementation of xgboost for classification of Parkinson's disease," In *J. Phys. Conf. S.*, vol. 1538, no. 1, p. 012024, May, 2020. [CrossRef]
23. S. Li, and X. Zhang, "Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm," *Neural Comput. Appl.*, vol. 32, no. 7, pp. 1971–1979, 2020. [CrossRef]
24. P. Tao, H. Shen, Y. Zhang, P. Ren, J. Zhao, and Y. Jia, "Status forecast and fault classification of smart meters using LightGBM algorithm improved by random forest," *Wirel. Commun. Mob. Comput.*, vol. 2022, no. 1, 3846637, 2022. [CrossRef]
25. J. T. Hancock, and T. M. Khoshgoftaar, "CatBoost for big data: An interdisciplinary review," *J. Big Data*, vol. 7, no. 1, p. 94, 2020. [CrossRef]
26. W. Chang, X. Wang, J. Yang, and T. Qin, "An improved CatBoost-based classification model for ecological suitability of blueberries," *Sensors (Basel)*, vol. 23, no. 4, 1811, 2023. [CrossRef]
27. J. M. Johnson, and T. M. Khoshgoftaar, "Deep learning and data sampling with imbalanced big data," In 20th IEEE international conference on information reuse and integration for data science (IRI). New York: IEEE, pp. 175–183, 2019. [CrossRef]
28. M. Hayaty, S. Muthmainah, and S. M. Ghufran, "Random and synthetic over-sampling approach to resolve data imbalance in classification," *Int. J. Artif. Intell. Res.*, vol. 4, no. 2, pp. 86–94, 2021. [CrossRef]
29. G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for handling imbalanced data problem: A review," In 2021 sixth international conference on informatics and computing (ICIC), pp. 1–8, 2021. [CrossRef]
30. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002. [CrossRef]
31. M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive analysis of random forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels," *Technologies*, vol. 13, no. 3, p. 88, 2025. [CrossRef]
32. K. Zhong, X. Tan, S. Liu, Z. Lu, X. Hou, and Q. Wang, "Prediction of slope failure probability based on machine learning with genetic-ADASYN algorithm," *Eng. Geol.*, vol. 346, 107885, 2025. [CrossRef]
33. H. Yakupoglu, H. Gozde and M.C. Taplamacioglu, "Online noise-adaptive Kalman filter integrated novel autoencoder for multi-fault detection and early warning of wind turbines," *Measurement*, vol. 256, 118538, doi: [CrossRef]