International Journal of Electrical, Energy and Power System Engineering
Vol. 07, No. 02, June 2024
e-ISSN: 2654-4644

Detection and Classification of Physical and Electrical Fault in PV Array System by Random Forest-Based Approach

Syed Sikandar Shah[1*], Bin Li[2], Anqi Zheng[3]
[1,2,3]College of Electrical, Energy and Power Engineering, Yangzhou University, China
[1]Sikandershah2020@gmail.com, [2]libin@yzu.edu.cn, [3]1192479263@qq.com

[*]Corresponding author: libin@yzu.edu.cn

Abstract— The importance of solar photovoltaic (PV) systems has increased over the past ten years due to the solar PV industry's explosive growth. To ensure the reliable, safe, and efficient operation of residential PV systems, fault detection is crucial. Early classification of faults can improve PV system performance and reduce damage and energy loss. Many recent studies have focused on classifying and detecting PV faults but most of them are limited to specific reasons like Real-world data can be restricted, unbalanced, or include noise, all of which may decrease the effectiveness of ML models. This paper proposes a method for identifying and classifying both physical and electrical faults in the PV array system applying a machine learning (Random Forest) model to that is trained using a synthetic photovoltaic training database. Make use of a synthetic PV database opens the door to a more precise, effective, and scalable PV system by addressing the limitations of real-world data. MATLAB is used to create a synthetic database while scikit-learn tool in Jupyter Notebook is used to train an ML model are the two main steps in this paper. The performance of the proposed model is compared with the existing ML model and achieves the most effective algorithm offering higher accuracy in detection of 98.6% and classification accuracy is 94.2% for both physical and electrical faults after being successfully tested on real-world datasets and trained on historical data from the PV array system (PV Database).

Keywords: AC Protection, Electrical Fault, Machine Learning, Physical Fault, PV Database, Solar Photovoltaic.

INTRODUCTION

Fault identification is crucial for power grid operation. Utilities are working to reduce outages caused by natural events, which can lead to long interruptions and economic impact on customers [1] [2]. The cost of a one-hour outage ranges from USD 3 to USD 82,000 depending on the customer type and time of year [3]. Predicting faults and their duration can reduce unplanned outages and enable utilities to deploy maintenance crews and sequence operations efficiently [4]. Balanced or unbalanced faults are the two general categories for physical problems in power systems [5]. Unbalanced or asymmetrical faults, which might be series or shunt-type faults, are often encountered faults. Regarding fault detection methods, many machine-learning (ML) approaches are currently developed and published in the field of research [7]. The detection and classification of common faults, (like open circuits, short circuits, partial shading, soiling, and deterioration, as well as complex problems, such as multiple faults, have shown good competency in this trained model. Much research on

Received: February 12, 2024, Revised: March 12, 2024, Accepted: March 23, 2024
https://doi.org/10.31258/ijeepse.7.2.67-84

machine-learning (ML) methods for fault detection and classification has been developed, and they have shown good performance in finding both common and complicated faults. However, these methods frequently rely on simulated or measured data, which could not accurately describe the intricate features of PV systems in the real world [8]. Utilizing a real-time implementation of the described techniques, only a small number of works have been experimentally validated. ML-based models are often trained and tested on measured or simulated data, which can be collected using MATLAB/Simulink. A fault detection system often consists of several tasks, including the localization, identification, classification, and detection of faults [9]. To address this problem, techniques for locating PV system faults have been developed that make use of PV yield databases, mostly based on output power. The use of machine learning (ML) to improve these techniques is yet largely unexplored. Open databases are not suited for accurately identifying faults since they are highly susceptible to data entry mistakes, and noise-full data such as PVoutput.org. An alternate method provides control and precision for training machine learning models by modeling PV systems and adding faults to create synthetic data. Compared to open databases containing historical data, synthetic databases offer advantages [10]. First of all, it removes the need for PV system owners to precisely enter the parameters of their system design, guaranteeing accurate and error-free system parameter input. Second, it is possible to balance a synthetic training database more successfully in terms of both Healthy and faulty systems. This study, on the other hand, had 783418 data points, which is more than 32 times as numerous. Other features such as operational temperature, array current, array voltage, fill factor, and others were provided in addition to irradiance, system power, and ambient temperature. When it comes to uncommon fault types that might have a major influence on PV system performance and safety, historical data frequently tends to be biased toward operating systems in a Healthy state. Machine learning models may be trained to identify and classify faults more accurately, by manually modifying the balance in the training database, particularly the less common ones [11]. A Random Forest machine learning model built on a synthetic photovoltaic (PV) training database is used in our work. AUC curve, F1-score, specificity, recall, accuracy, and precision are some of the performance matrices that are used to compare the Random Forest model's performance to that of other models. The model's output is also graphically represented through the use of ROC curves and confusion matrices. Hyperparameter tuning is necessary to improve the model's performance and its predictive power. Scikit-learn Python library is an extensible tool for machine learning applications it creates a predictive model that is both effective and accurate for fault identification and classification. The sequence of this paper is as stated below: Sec 2 explain common residential solar system faults and synthetic database generation. Sec 3 explains in more detail of proposed ML model construction. Sec 4 covered the performance of the ML models using the synthetic database. Finally, to summarize, the important conclusion is discussed in sec 5.

FAULT IDENTIFICATION AND PARAMETER EXTRACTION

A. Faults in a PV Array

Faults are a common source of operational difficulties for photovoltaic (PV) systems; the most common types are Line to Line (L-L) and Ground Faults, which can result in short circuits and low power output. Although not a typical fault, inverter failures, such as inverter clipping, are still regular, impacting about 22% of residential PV systems and limiting maximum power [12]. broken string faults are less frequent but significant since they can cause open circuits and decreased power output. These faults can be caused by problems such as faulty solder connections or broken cables [13].

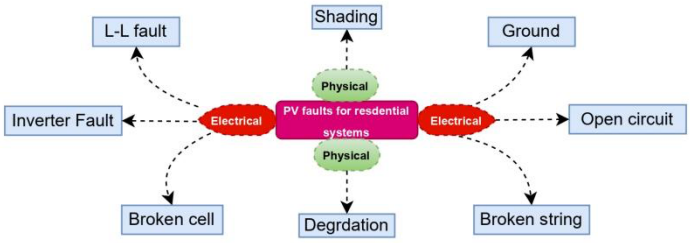

Figure 1. Common residential PV faults

The electrical components of PV systems are the main source of these issues. Physical faults that impact system performance include broken cells or connections brought on by shade or poor manufacture. For instance, shading's effect on system efficiency varies. This study focuses on identifying and classifying undetected faults such as short circuits and routine failures to improve system performance, even though it admits the difficulty of diagnosing these problems, particularly when inverter faults and short circuits may be confounded [14].

B. Dataset Parameter for PV Modeling and Analysis

To enhance the efficiency of fault detection and classification in photovoltaic (PV) systems, a new synthetic training database using MATLAB PVMD Toolbox was proposed. This database resembles both optimal and unfavorable system ages and types of meteorological data. The database contains hourly weather data for the Jiangsu station in which a random simulation created over 48-hour profiles of days picked from there. On occasion, the consequences of low irradiance were also mitigated. By combining a variety of healthy and faulty system states, this method increased the efficacy of machine learning. The main components of this database simulations are "CELL," which stands for a single solar cell and the number of single solar cells that must be simulated; "MODULE," which represents mounted panels and describes the unit part of the structural part of a PV system; "WEATHER," which holds environmental conditions such as temperature and humidity in the weather, crucial for realism in simulations; "ELECTRICAL," which contains the electrical characteristics of the system that must be simulated, aiding in performance analysis; and "CONVERSION," which refers to the conversion of data into a format that is more useful for the PVMD Toolbox. These are fixed, except for the weather data, which is added to increase the simulation performance.

The first factor was the addition of a dynamic factor to improve its applicability in machine learning database applications under a variety of weather conditions, especially low light conditions. This factor includes hourly irradiance and input from the previous day. As shown in Table 1 below crucial cell and module parameters, as well as the weather and temperature, are listed. These factors are essential to the synthetic database that we utilized to train our machine-learning algorithm to identify and classify faults in PV systems.

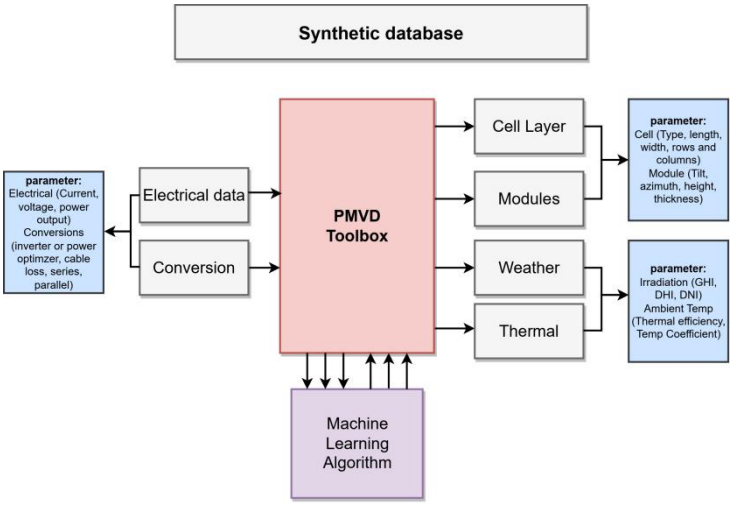

Figure 2. Schematic diagram of PMVD toolbox

Table 1. System Parameters for Database

PV System Parameters			
Cell and module parameters		**Weather and thermal parameters**	
Parameters	**Values**	**Parameters**	**Values**
Cell type	MonoPERC	Longitude	51.955278
Module tilt	35*	Latitude	4.347778
Module azimuth	15*	Efficiency of PV module	0.22
Module height above ground	75 cm	Temperature Coefficient	0.00321*C
Number of cell rows	10	Glass thickness	0.35cm
Number of cell columns	8	**Electrical and Conversion Parameters**	
Module thickness	0.4 cm	**Parameters**	**Values**
Module cell spacing	0.2 cm	Shading loss due to metallization	2%
Module edge spacing	0.5 cm	Metal grid resistance	0.0055
Cell length	16 cm	Number of bypass diodes	2
Cell width	16 cm	Inverter type	hybrid
Module albedo	0.15	Inverter capacity	4500W
		Number of modules in	3
		Number of modules in series	6
		Cable losses	0.80%

Additionally, a new feature called the power ratio has been included to provide insights into the health of the system and greatly increase the fault detection model prediction accuracy. Figure 3, which shows the Chi-square test results for seven important features that have a significant impact on the system's performance analysis and fault diagnostics, will be presented to graphically describe the impact of each variable.

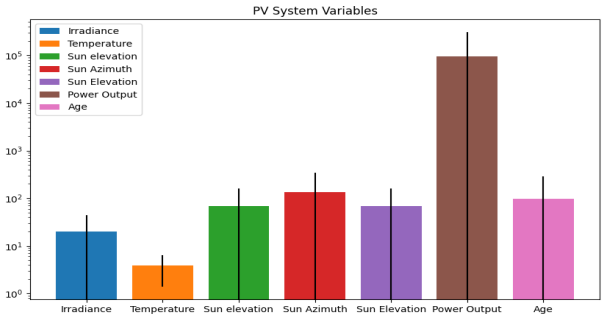

Figure 3. Chi-square test values for 7 feathers

C. Database Result

To simulate various faults, several system scenarios were simulated using the PVMD Toolbox [15]. A comparison of several fault classes for the same system under the same meteorological conditions is shown in Figure 4. For training machine learning models, the training database, which included 25,000 scenarios, collected important variables such as system age, system power output, and weather conditions. A 48-hour weather attribute, a matching power output profile, and the system status are all included in each database scenario, along with hourly values for the connected variables. In the database, there are 40% faulty systems and 60% Healthy systems. In further cases, 25% are caused by broken cells, broken string, Ground faults, and Line-to-Line faults combined. Zero efficiency faults were initially listed but were later deleted since their 100% detection rate made them uninteresting to study.

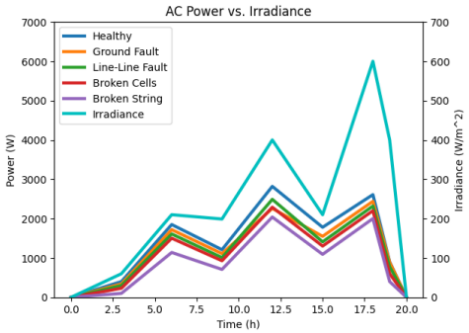

Figure 4. Photovoltaic Fault Analysis with AC Power Output

PROPOSED MODEL

A. Feather Selection and Extraction

In order to train an ensemble learning model more specifically, a Random Forest (RF) algorithm to detect faults in PV systems, this study uses a synthetic database. Nineteen characteristics that represent the various circumstances of PV arrays during normal operation and faults serve as the basis for the model. In order to improve the model capacity to accurately classify faults, an initial feature selection technique is used to simplify the dataset by eliminating less useful information. The primary approach used for fault detection is the RF model. Its effectiveness in classifying faults in PV systems is then verified by comparing its

performance against that of the other algorithms in the ensemble, which include Decision Trees (DT), Support Vector Machines (SVM), and Neural Networks (NN).

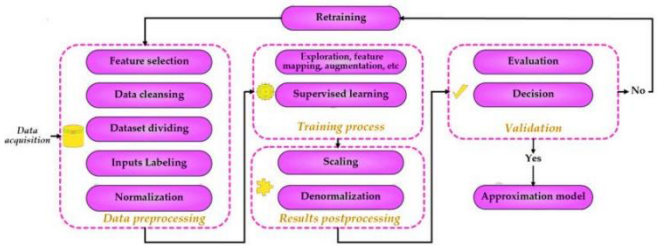

Figure 5. Workflow Diagram for Predictive Machine Learning Model Modified From [16]

Voltage, temperature, and performance measurements are included in the dataset, also known as matrix X. Each row in the matrix denotes an observation, and each column a characteristic. Labeled data is used to train the model, enabling it to distinguish between PV states that are functional and those that are not. The predictive accuracy of the model is verified by contrasting its fault predictions with current conditions using a separate, unlabeled test dataset. The predictive model's workflow is depicted in Figure 5, which highlights the critical roles that feature selection, data preparation, and repeated retraining serve in the model evaluation process.

B. Optimized Model

It is important to assess the kind of data that is available and select the appropriate machine-learning method. The study employed discrete, labeled data for PV fault classification and detection. As a result, a supervised learning algorithm that is capable of classifying data should be applied. The Decision tree (DT), Random Forest (RF), support vector machine (SVM), and neural network (NN) are the most widely used machine learning methods in the field of PV fault identification and classification. This research uses a synthetic PV training database from sec 2 to train four algorithms. The final model is developed through the Python Jupyter Notebook. Metrics like recall, accuracy, specificity, F1-score, precision and AUC are used to evaluate performance; confusion matrices are used to display results. To improve model performance, hyperparameter tuning requires to modifying parameters particular to certain algorithms.

Random Forest

Since its introduction in 2001 by Leo Breiman, the Random Forest (RF) algorithm has grown to be one of the most popular machine learning methods [17]. The purpose of this work is to identify faulty operations and classify certain fault types using the Random Forest (RF) algorithm in residential PV systems. Both binary and multiclass classification are considered necessary to accomplish this aim. At first, several fault types are combined into one 'faulty' class, whereas the 'Healthy' class is constructed up of standard operating systems. The study uses multiclass classification to evaluate the recommended method's capacity to correctly classify various kinds of faults.

$$\mathrm{p} = (1 - \tfrac{1}{\mathrm{n}})^{\mathrm{n}} \tag{1}$$

$$\lim_{n\to\infty} p = \lim_{n\to\infty} \left(1 + \tfrac{1}{n}\right)^{n} = e^{-1} \approx 36.8\% \tag{2}$$

Analysis Of RF Tree Creation

Multiple Decision Trees (DTs) are built using the Random Forest (RF) algorithm, an ensemble learning technique, using resampled datasets from the actual training data. The Classification and Regression Tree (CART) algorithm creates each DT, and to improve accuracy and decrease correlation between DTs, a Random feature selection technique is used. When creating an RF tree, a given total feature count is taken into account. The best possible split-point approach is then used to choose a Random subset of features for node selection. After RF tree creation, Decision trees are used to classify the sensed and transmitted data. This process begins with a root node and repeatedly splits nodes X{1} and X[2] based on feature best values lower or higher than 845.457 until a stopping requirement, such as a certain tree depth, is satisfied. The best splitting is calculated depending on the Gini level to gradually reduce the weighted average of 0.49 for purer nodes. A minimum number of samples required to split a node or create a leaf node might be one of the stopping criteria.

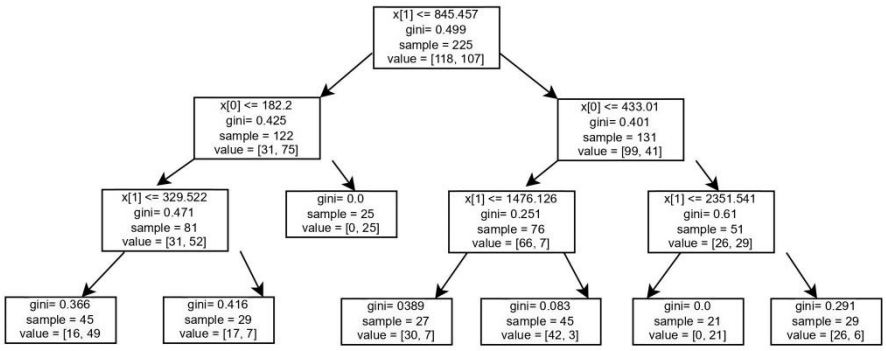

Figure 6. Tree Structure of The RF Model with Three Depths

A group of Randomized Decision trees, each trained on a subset of features and a bootstrapped sample of the training set, is called a Random Forest. Each tree contributes to the prediction of the most likely class for each sample, resulting in a strong classification from the total of the data gathered from individual trees.

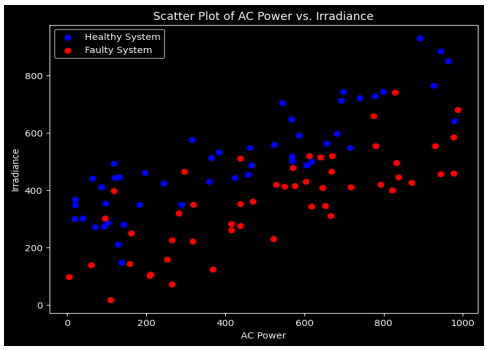

Figure 7. Predicted System Status for RF Model

This method improves overall accuracy by reducing the over-reliance on certain feature subsets that may be found in individual Decision trees. By contrast, even if a single Decision tree performs quite well on training data, it frequently fails to generalize to test data.

Other ML Model

Three machine learning approaches are compared with our proposed method: DT, SVM, and ANN (more specifically, a multilayer perceptron feedforward artificial neural network). Random forest (RF) successfully minimizes overfitting and often achieves greater accuracy than other machine learning models as shown in accuracy analysis Table 1.

Table 2. Accuracy Analysis Between ML Models

Model	Data type	Accuracy
RF	Supervised	Detection = 98.6% Classification = 94.2%
DT	Supervised	Detection = 93.1% Classification = 89.3%
SVM	Supervised	Detection = 97.1% Classification = 93.4%
NN	Supervised	Detection = 97.1% Classification = 94.0%

By employing multiple Decision Trees, each with a distinct feature subset, Random Forest (RF) reduces over-fitting that is sometimes present in single Decision Trees and enhances generalization. In contrast to Support Vector Machines (SVM) and neural networks, RF improves performance on both linear and non-linear data by adding variety through its ensemble technique.

C. Performance Metrics Evaluation

A confusion matrix is used to critically evaluate a machine learning model that predicts system problems by classifying predictions into True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Whereas TNs are correctly detected healthy systems, TPs are correctly identified problems. On the other hand, flaws overlooked by the model are known as FNs, while good systems are known as FPs when mislabeled [18].

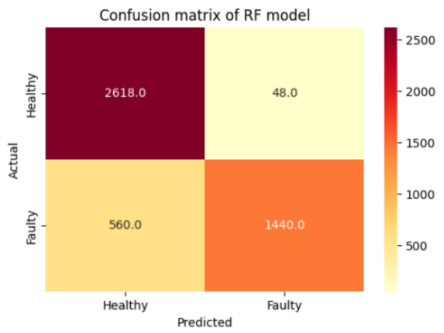

Figure 8. RF Model Confusion Matrix for Faulty System

To improve the model's F1 score, accuracy, recall, precision, and Area Under the Curve (AUC), hyperparameter tweaks are guided by these measures, which guarantee a balanced performance across different fault prediction characteristics. These metrics are essential to model optimization procedures because a larger ratio of TP and TN to FP and FN denotes improved model performance [19].

Accuracy: $\frac{(TP+TN)}{(TP+TN+FP+FN)}$ (3)

Recall (Sensitivity, True Positive Rate): $\frac{TP}{(TP+FN)}$ (4)

Precision (Positive Predictive Value): $\frac{TP}{(TP+FP)}$ (5)

Specificity (True Negative Rate): $\frac{TN}{(TN+FP)}$ (6)

F1 Score: $\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$ *(7)*

The model's total discriminatory power is further evaluated by considering the Area Under ROC curve (AUC). A careful balance between model complexity and generalization to new data is necessary, as evidenced by the small variation in ROC curves between training and validation data that was observed. This study ensures that the model works well in practical applications by laying the foundation for future research and ongoing enhancements.

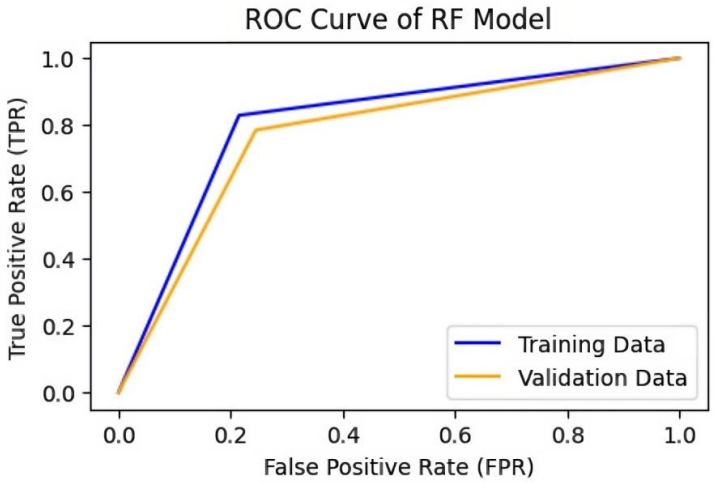

Figure 9. ROC Curve of RF Model for Faulty Sets

ML MODEL PERFORMANCE

The Random Forest (RF) model designed for fault identification using data from the PMVD toolbox is shown in the flowchart and the specifics mentioned in Chapter 3. After giving careful consideration to each component's requirement, a flowchart is created as shown in Figure 4.1. This begins with gathering data using the PMVD toolbox, which is a collection of tools for predictive maintenance, and then moves on to pre-processing the data so that it can be analyzed. When a fault is strike this is decision point, the system determines whether fault occur or not if not, the process will end because there is no fault occur and move to data collection box again. If yes then it split for a testing set of labeled data is generated to assess the model detection efficiency. This results in improved data, which serves as the test set to evaluate how well the RF model performs in terms of generalization and accuracy. For classification the system gathers similar events, labels them in the similar duration, and creates a complete report that includes information on the fault causes and effects. The prepared data are fed into a Random Forest (RF) model to classify possible faults, a machine-learning technique recognized for its high classification performance. Hyperparameter adjustment is a crucial step toward better predictive performance if the dataset is further refined

and applied to the RF model. The result is an improved RF model, especially for precise problem detection in the system it monitors.

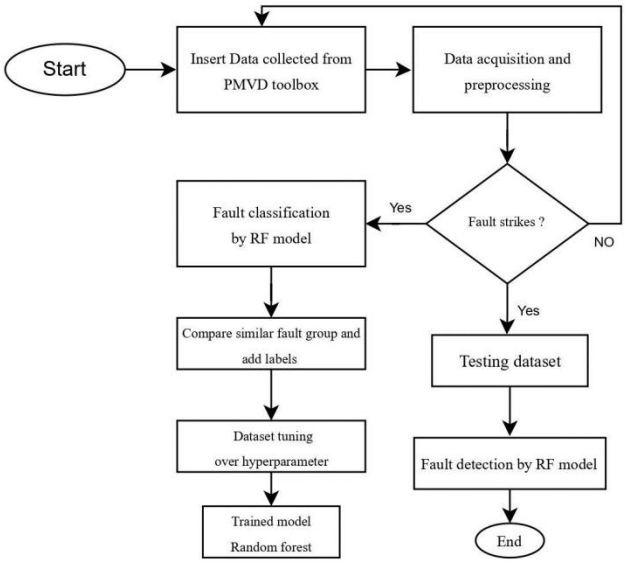

Figure 10. Flowchart of RF Model

The data that follows is divided into test, validation, and training sets to address any bias in hyperparameter optimization. A validation set avoids hyperparameter biasing, which results in a more precise representation of model performance during final testing, even if training and test sets boost model performance. Variance is ensured by the widely used division of 60% training, 20% validation, and 20% test, even if class balance may vary somewhat within sets.

Table 3. Original Feature of The Mean and Standard Deviation

Dataset	Training	Validation	Test
Irradiance W/m^2	[251, 201]	[254, 205]	[251, 201]
Temperature °C	[14.6, 8.2]	[14.6, 8.2]	[14.6, 8.2]
Azimuth°	[0.2, 51.3]	[-2.2, 51.4]	[2.1, 51.3]
Elevation°	[26.1, 16.9]	[26.3, 17.1]	[26.3, 16.7]
Power W	[1503, 1410]	[1506, 1401]	[1505, 1411]
Duration	[2.30, 5.45]	[2.42, 5.56]	[2.29, 5.43]

Faults were simulated at different periods between 09:00 and 15:00, taking into account the effect of variable irradiance on power production, to improve the model fault prediction accuracy. To control complexity, models were initially trained on certain hourly data. The objective of the random selection was to include possible low irradiance times, which are known to impede fault identification [20]. Additional data, such as power output, irradiance from certain hours (09:00, 13:00, and 15:00) the day before, and the power ratio (output to irradiance) were added to handle irradiance fluctuation. This improved fault identification under various irradiance scenarios by expanding the feature set to 19.

Table 4. Shows The Calculated Feather for The Train Model

Case	Number of Cases	Features
1A	6	Irradiance, Ambient Temperature, Sun (azimuth + elevation), System (Power + Age)
1B	13	1A features + Irradiance Day before, Irradiance at 2 pm, System power day before, System power at 2 pm, Power ratio, Power ratio day before, Power ratio 2 pm
1C	16	1A and 1B features + Irradiance at 9 am, Irradiance at 3 pm, System power at 9 am, System power at 3 pm, Power ratio at 9 am, Power ratio at 3 pm

A. Fault Detection Results

In this study, binary classification is used for fault detection where the Random Forest (RF) model beats SVM, DT, and ANN, demonstrating efficacy even in the presence of nineteen intricate characteristics. Table 5 and Figure 4.2 illustrate RF's accuracy, which shows that even with a longer training time, it is still superior at differentiating between healthy and faulty systems. The typical problem of real-world datasets bias towards Healthy systems was solved by using a synthetic database to guarantee class balance. We varied the ratio of the faulty-to-total system from 1% to 70% while keeping a consistent dataset size of 15,000 data points, taken from an initial pool of 25,000, to thoroughly assess model performance across different dataset balances. Because there were few faulty systems, a dataset with 14,286 data points was employed to achieve the greatest failure ratio of 70%.

Table 5. RF Model Results for Different Feature Values on The Test Set

Metric	F1 Score	Recall	Specificity	Precision	Accuracy	AUC
6 features	0.674	0.651	0.821	0.692	0.75	0.731
13 features	0.752	0.73	0.863	0.784	0.801	0.792
19 features	0.756	0.733	0.879	0.786	0.819	0.791

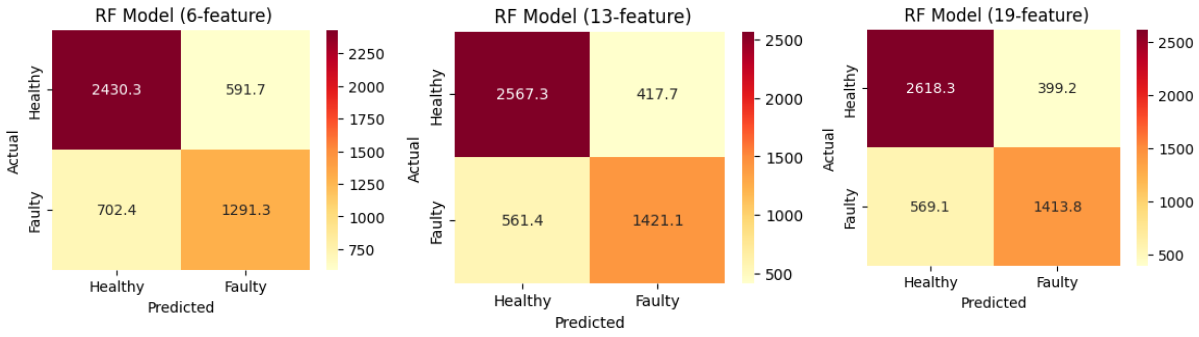

Figure 11. RF model confusion matrix on test set including 6, 13, and 19 features

To determine which model performed best at various fault ratios, our investigation examined four models with varied dataset balances. In datasets with 30% and 50-70% fault rates in particular, the Random Forest (RF) model performed exceptionally well, demonstrating its durability and dependability in binary classification tasks. The SVM model works better at first

in datasets with a fault ratio of 0–20%, but it becomes less successful in datasets that are more balanced. The Neural Network is the best performer with a 30% fault ratio. All dataset balances and fault ratios, however, consistently demonstrate that the Random Forest model performs better. These variant highlights how flexible the Random Forest is, and Figure 11 shows how dominant it is using F1-scores and AUC comparisons.

Table 6. RF Model Result on Test Set for Changing Dataset of Faulty System Ratio

Metrics	Accuracy	Recall	Precision	Specificity	F1	AUC
1%	0.996	0.292	1.00	1.00	0.452	0.646
10%	0.929	0.349	0.967	0.999	0.512	0.674
30%	0.801	0.457	0.847	0.963	0.594	0.71
50%	0.817	0.836	0.799	0.786	0.819	0.808
70%	0.851	0.929	0.844	0.621	0.893	0.771

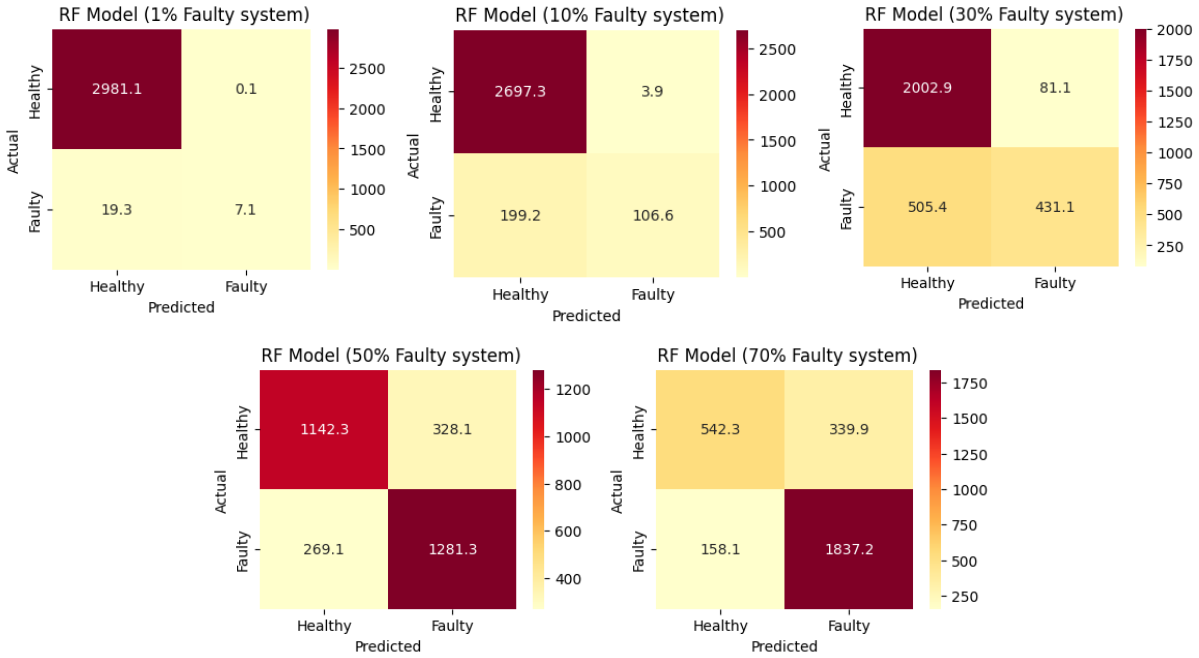

Figure 12. RF model confusion matrix on test set for 1% to 70% of the dataset faulty systems

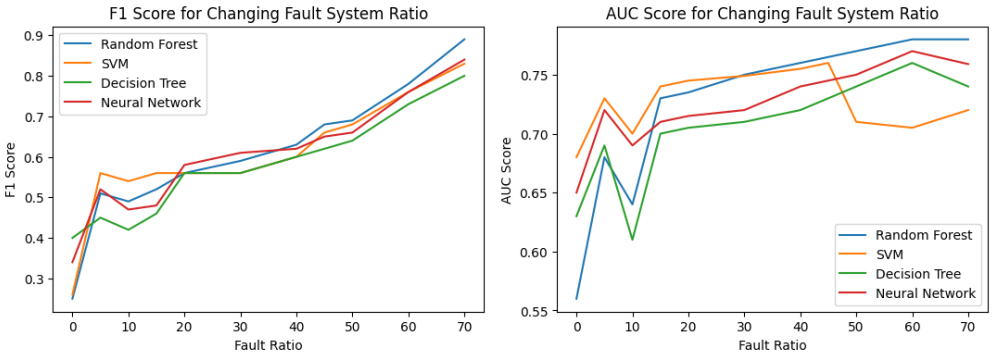

Figure 13. AUC and F1 score with changing fault ratio

B. Fault Classification Results

We used OVO and OVR approaches in our work on multi-classification for residential PV system failures, which allowed us to divide the faults into four subcases. The Random Forest (RF) model demonstrated higher recall, F1, and accuracy scores across five fault classes Table VII, outperforming SVM and neural networks in both OVO and OVR techniques. Notably, the RF model consistently demonstrated high scores in identifying "Broken String" faults.

Table 7. RF Model Performance Using Multi-Classification on Test Set (OVR classifier)

Class	Recall Score	F1 Score	Precision Score
Healthy	0.526	0.605	0.712
Line-to-Line	0.496	0.504	0.512
Ground	0.654	0.526	0.44
Broken Cell	0.586	0.645	0.716
Broken String	0.952	0.936	0.92
Average	0.643	0.643	0.66

Table 8. RF Model Performance Using Multi-Classification on Test Set (OVO classifier)

Class	Recall Score	F1 Score	Precision Score
Healthy	0.55	0.62	0.71
Line-to-Line	0.508	0.512	0.519
Ground	0.639	0.519	0.441
Broken Cell	0.605	0.663	0.731
Broken String	0.952	0.94	0.929
Average	0.62	0.649	0.671

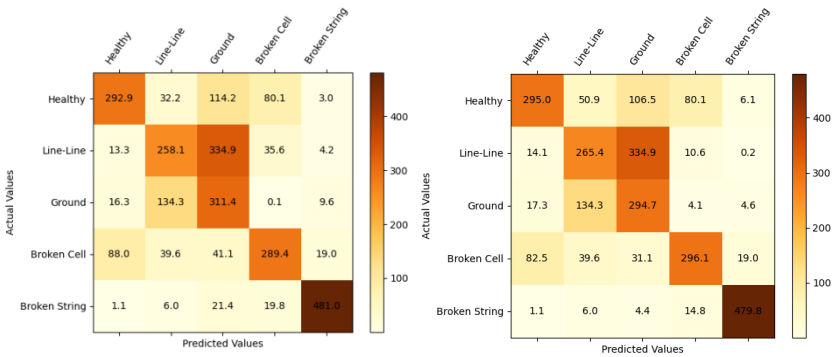

Figure 14. RF model confusion matrix of (a) OVO (b) OVR

In order to improve overall model performance, "Line-to-Line" and "Ground" faults were integrated into a single "Short-Circuit" category to overcome the difficulty of differentiating between them. Even though there is still some ongoing challenge in correctly recognizing Healthy systems, this strategic tweak together with the use of a balanced dataset of 12,500 samples enhanced fault classification, especially for "Healthy" and "Broken Cell" classes Figure 14.

Table 9. RF Model Performance Using Multi-Classification on a Test Set with Collective Short Circuit Faults Dataset (OVO classifier)

Class	Recall score	F1 score	Precision score
Healthy	0.624	0.564	0.751
Short-Circuit	0.772	0.929	0.682
Broken Cell	0.669	0.582	0.789
Broken String	0.948	0.989	0.936
Average	0.749	0.782	0.792

Table 10. RF Model Performance Using Multi-Classification on a Test Set with Collective Short Circuit Faults Dataset (OVR classifier)

Class	Recall score	F1 score	Precision score
Healthy	0.635	0.565	0.723
Short-Circuit	0.767	0.928	0.653
Broken Cell	0.658	0.572	0.774
Broken String	0.95	0.978	0.925
Average	0.752	0.761	0.769

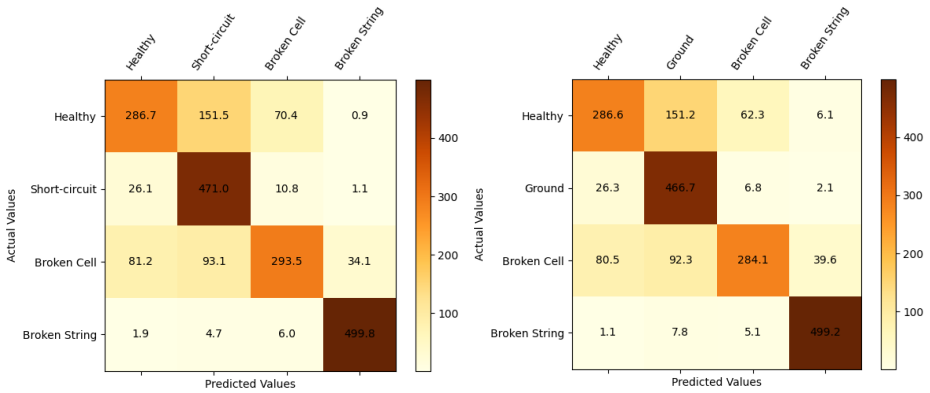

Figure 15. RF Model Confusion Matrix Using a Test Set of Datasets with Short Circuit Fault Class of OVO and OVR

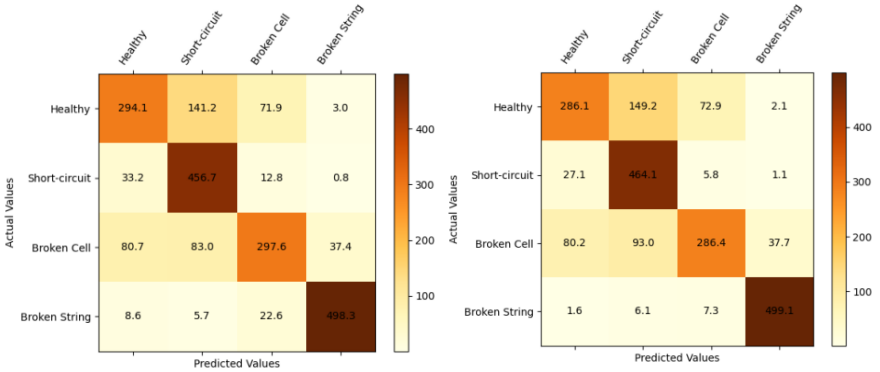

Figure 16. Rf Model Confusion Matrix of Only (OVR) Classifier of $R_{fault.max}$= (a)100Ω (b)1000Ω

We reduced the maximum fault resistance from 1000Ω to 100Ω to address the misclassification of healthy systems as short-circuit faults. This improved model performance and fault classification, particularly for the healthy and short-circuit classes. Figure 4.7 shows how clearly fault types could be distinguished in the Confusion Matrices for both resistance levels, indicating that the Random Forest model OVR algorithm worked best. This tactical modification highlights the advantage of the Random Forest model in fault classification, which is confirmed by comparable performance metrics in Table 11.

Table 11. RF Model Performance Using Multi-Classification on a Test Set with Collective Short Circuit Faults Dataset of $R_{Fault,\ Max=100\Omega}$

Class	Recall Score	F1 Score	Precision Score
Healthy	0.639	0.564	0.731
Short-Circuit	0.782	0.936	0.659
Broken Cell	0.662	0.579	0.781
Broken String	0.956	0.982	0.932
Average	0.760	0.765	0.775

Table 12. RF Model Performance Using Multi-Classification on a Test Set with Collective Short Circuit Faults Dataset of $R_{Fault,\ Max=1000\Omega}$

Class	Recall Score	F1 Score	Precision Score
Healthy	0.656	0.590	0.731
Short-Circuit	0.785	0.940	0.682
Broken Cell	0.671	0.602	0.870
Broken String	0.963	0.981	0.936
Average	0.768	0.778	0.804

CONCLUSION

In the work, model that relies on machine learning exhibited encouraging results in detecting and classifying faults in residential solar power systems, even in difficult situations that haven't been by earlier research before. The Random Forest model performed well in binary fault detection, with 91.1% accuracy at a 50% faulty system balance and 94.6% accuracy at a 70% faulty system ratio. For multiclass fault classification, the model with the one-vs-rest classifier obtained an average (Recall: 0.778, F1-score: 0.768, Precision: 0.804). Particularly, the short-circuit fault class had an appropriate F1 score of 0.785, whereas broken string faults came in with an outstanding F1 score of 0.963. The study focuses on how the model's simplicity and the advantages of synthetic databases in getting a defined balance dataset make it a feasible option for residential PV systems. However, because different studies have different dataset sizes, features, and fault types, it can be difficult to compare accurately with the research that already exists.

Table 13. Summary of Existing Literature on PV Fault Detection

Ref.	Study	Approach	Data Source	Advantages	Limitations
[21] [22]	Machine Learnin g	NN + Levenberg–Marquardt	Real-time data system	Detection inaccuracy is under 3%. Real time fault location, High power accuracy and determine fault type with distant source of end tolerance	The training procedure has a long completion time. Fault timing is not examined

[23]	NN-based	Real-time data system	when the network topology changes, the findings for predicting the fault distance from the substations are optimal. Noise tolerance is high.	Fails to precisely detect faults in live data streams coming from the Power grid.
[24]	CNN-based	Real-time data system	Diverse fault types, fast re-closures, and complicated transient states after a fault event make real-time fault location	
[25]	RF + DT	Open PV Database	The accuracy of identifying the location of the fault is 91% when there are few no of buses (5-7%).	Research focus on fault location, not duration
[26]	RF	PV system data	fault detection algorithm has an accuracy of 90.69%. The fault location algorithm has a performance of 0.30% on average. carried out on a large-scale distribution feeder	
[27]	KNN	PV system data	The accuracy of fault location identification is 98.70%. Absolute error ranges between 0.61 to 6.5%	A model trained and tasted only on a PV system
Propose d method	Random forest model	Synthetic database	Error-free and controlled data. Balanced dataset with a healthy and faulty system. Scalable for training and experimentation	

REFERENCES

[1] H. Haes Alhelou, M. E. Hamedani-Golshan, T. C. Njenda and P. Siano, "A Survey on Power System Blackout and Cascading Events: Research Motivations and Challenges," Energies, vol. 12, 2019.

[2] Y. Zhang, Y. Xu and Z. Y. Dong, "Robust Ensemble Data Analytics for Incomplete PMU Measurements-Based Power System Stability Assessment," IEEE Transactions on Power Systems, vol. 33, pp. 1124-1126, 2018.

[3] L. Lawton, M. Sullivan, K. Van Liere, A. Katz and J. Eto, "A framework and review of customer outage costs: Integration and analysis of electric utility outage cost surveys," 2003.

[4] A. Jaech, B. Zhang, M. Ostendorf and D. S. Kirschen, "Real-Time Prediction of the Duration of Distribution System Outages," IEEE Transactions on Power Systems, vol. 34, pp. 773-781, 2019.

[5] S. S. Gururajapathy, H. Mokhlis and H. A. Illias, "Fault location and detection techniques in power distribution systems with distributed generation: A review," Renewable and sustainable energy reviews, vol. 74, p. 949–958, 2017.

[6] A. Et-taleby, Y. Chaibi, M. Benslimane and M. Boussetta, "Applications of machine learning algorithms for photovoltaic fault detection: a review," Statistics, Optimization & Information Computing, vol. 11, p. 168–177, 2023.

[7] R. Nijman, "Automatically and real-time identifying malfunctioning PV systems using massive on-line PV yield data," 2018.

[8] O. Tsafarakis, P. Moraitis, B. B. Kausika, H. Van Der Velde, S. ’t Hart, A. de Vries, P. de Rijk, M. M. De Jong, H.-P. van Leeuwen and W. Van Sark, "Three years experience in a

Dutch public awareness campaign on photovoltaic system performance," IET Renewable Power Generation, vol. 11, p. 1229–1233, 2017.

[9] N. A. Engerer and J. Hansard, "Real-time simulations of 15,000+ distributed PV arrays at sub-grid level using the regional PV simulation system (RPSS)," in Proceesings of the Solar World Congress, 2015.

[10] Y. Zhao, L. Yang, B. Lehman, J.-F. de Palma, J. Mosesian and R. Lyons, "Decision tree-based fault detection and classification in solar photovoltaic arrays," in 2012 Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition (APEC), 2012.

[11] A. Charki, P.-O. Logerais, D. Bigaud, C. M. F. Kébé and A. Ndiaye, "Lifetime assessment of a photovoltaic system using stochastic Petri nets," International Journal of Modelling and Simulation, vol. 37, p. 149–155, 2017.

[12] M. Bressan, Y. El-Basri and C. Alonso, "A new method for fault detection and identification of shadows based on electrical signature of faults," in 2015 17th European Conference on Power Electronics and Applications (EPE'15 ECCE-Europe), 2015.

[13] A. Sayed, M. El-Shimy, M. El-Metwally and M. Elshahed, "Reliability, availability and maintainability analysis for grid-connected solar photovoltaic systems," Energies, vol. 12, p. 1213, 2019.

[14] T. Berghout, L.-H. Mouss, T. Bentrcia, E. Elbouchikhi and M. Benbouzid, "A deep supervised learning approach for condition-based maintenance of naval propulsion systems," Ocean Engineering, vol. 221, p. 108525, 2021.

[15] L. Breiman, "Random forests," Machine learning, vol. 45, p. 5–32, 2001.

[16] G. T. Klise, O. Lavrova and R. L. Gooding, "PV System Component Fault and Failure Compilation and Analysis.," 2018.

[17] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen and S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," Scientific reports, vol. 12, p. 5979, 2022.

[18] C. Del Cañizo, A. B. Cristóbal, L. Barbosa, G. Revuelta, S. Haas, M. Victoria and M. Brocklehurst, "Promoting citizen science in the energy sector: Generation Solar, an open database of small-scale solar photovoltaic installations," Open Research Europe, vol. 1, 2021.

[19] S. A. M. Javadian, A. M. Nasrabadi, M.-R. Haghifam and J. Rezvantalab, "Determining fault's type and accurate location in distribution systems with DG using MLP neural networks," in 2009 International conference on clean electrical power, 2009.

[20] Y. Aslan, "An alternative approach to fault location on power distribution feeders with embedded remote-end power generation using artificial neural networks," Electrical Engineering, vol. 94, p. 125–134, 2012.

[21] F. Dehghani and H. Nezami, "A new fault location technique on radial distribution systems using artificial neural network," 2013.

[22] W. Li, D. Deka, M. Chertkov and M. Wang, "Real-time faulted line localization and PMU placement in power systems through convolutional neural networks," IEEE Transactions on Power Systems, vol. 34, p. 4640–4651, 2019.

[23] A. Zainab, S. S. Refaat, D. Syed, A. Ghrayeb and H. Abu-Rub, "Faulted line identification and localization in power system using machine learning techniques," in 2019 IEEE International Conference on Big Data (Big Data), 2019.

[24] H. Okumus and F. M. Nuroglu, "A random forest-based approach for fault location detection in distribution systems," Electrical Engineering, vol. 103, p. 257–264, 2021.

[25] S. R. Madeti and S. N. Singh, "Modeling of PV system based on experimental data for fault detection using kNN method," Solar Energy, vol. 173, p. 139–151, 2018.

[26] N. Dahal, O. Abuomar, R. King and V. Madani, "Event stream processing for improved situational awareness in the smart grid," Expert Systems with Applications, vol. 42, p. 6853–6863, 2015.

[27] U. Adhikari, T. H. Morris and S. Pan, "Applying hoeffding adaptive trees for real-time cyber-power event and intrusion classification," IEEE Transactions on Smart Grid, vol. 9, p. 4049–4060, 2017.

BIOGRAPHIES OF AUTHORS

SYED SIKANDAR SHAH was born in kpk, Pakistan on 5 July 1996. he completed his Bachelor in Electrical (Telecommunication) Engineering at the University of Science and Technology Bannu, Pakistan in 2018. Right now, he pursued his MS in Power and Automation at College of Electrical, Energy and Power Engineering Yangzhou University of China.

LI BIN was born in July 1966 in China, and is a Full-time professor at the School of Water Conservancy and Energy Power Engineering of Yangzhou University. He is currently engaged in research on automation and management of water conservancy and hydropower projects, and pump station automation and informatization.

ANQI ZHENG was born in Suzhou, China in 1998. She is doing her MS in in Power and Automation at College of Electrical, Energy and Power Engineering Yangzhou University of China.