International Journal of Reconfigurable and Embedded Systems (IJRES)
Vol. 14, No. 2, July 2025, pp. 507~517
ISSN: 2089-4864, DOI: 10.11591/ijres.v14.i2.pp507-517

Enhanced fault detection in photovoltaic systems using an ensemble machine learning approach

Mohammed Salah Ibrahim[1], Hussein K. Almulla[2], Anas D. Sallibi[3], Ahmed Adil Nafea[1], Aythem Khairi Kareem[4], Khattab M. Ali Alheeti[5]

[1]Department of Artificial Intelligence, College of Computer Science and Information Technology, University of Anbar, Ramadi, Iraq
[2]Computer Center, University of Anbar, Ramadi, Iraq
[3]Department of Computer Sciences, College of Sciences, University of Al Maarif, Ramadi, Iraq
[4]Department of Heet Education, General Directorate of Education in Anbar, Ministry of Education, Heet, Iraq
[5]Department of Computer Networking System, College of Computer Science and Information Technology, University of Anbar, Ramadi, Iraq

Article Info

Article history:

Received Sep 7, 2024
Revised Apr 20, 2025
Accepted Jun 10, 2025

Keywords:

Artificial intelligence
Ensemble models
Fault detection
Machine learning
Photovoltaic systems
Solar energy

ABSTRACT

Malfunctioning of photovoltaic (PV) systems is a main issue affecting solar panels and other related components. Detecting such issues early leads to efficient energy production with low maintenance costs and high system performance consistency. This paper proposed an ensemble model (EM) for fault detection (FD) in PV systems. The proposed model utilized advanced machine learning algorithms containing random forest (RF), k-nearest neighbors (KNN), and gradient boosting (GB). Traditional approaches often do not handle the several situations that PV systems can have. Our EM leveraged the power of GB’s algorithm in handling complex data patterns through iterative boosting, KNN’s capability in capturing local data structures, and RF’s strength in handling overfitting and noise through its tree structure randomness. Combining these models enhanced fault detection capabilities, providing excellent accuracy compared to individual models. To evaluate the performance of our EM, different experiments were conducted. The results demonstrated substantial improvements in detection fault, achieving an accuracy rate of 95%. This accuracy rate considered high underscores the model’s capability to handle fault detection of PV systems, posing a consistent solution for instant fault detection and maintenance scheduling.

Corresponding Author:

Ahmed Adil Nafea
Department of Artificial Intelligence, College of Computer Science and Information Technology
University of Anbar
Ramadi, Iraq
Email: ahmed.a.n@uoanbar.edu.iq

1. INTRODUCTION

The industry of photovoltaic (PV) systems got a high attention in recent years due to the environmental and economic advantages of solar energy [1]. Although it is free availability and additional sensual aspects, the PV industry encounters challenges such as reduced output power, reliability, exposure to faults, high cost, and its dependence on environmental changes [2]. Because PV system's functionality is affected by environment conditions, they are affected by various faults like line-to-ground (LG), line-to-line (LL), open-circuit (OC), short-circuit (SC), hot spot (HS), partial shading (PS), and dust. Such issues can lead to low energy production and lower system efficiency. In addition to that, it has been reported that these conditions are the primary cause of PV lifetime reduction [3].

PV faults parts like panels, arrays, modules, connection lines, inverters, and converters divided into three primary types corresponding to their time characteristics: abrupt, incipient, or intermittent [4]. External aspects like PS (dead birds, sand dust, and fall of leaves) or internal aspects (OS and SC, elements parameters switching, and elements aging) conduct the happening of faults in the PV or in its immediate circuit, conduct to the general failure system [5].

Various fault detection (FD) methods have been placed in the last decade for PV. Manual and traditional methods of detecting faults in PV are inaccurate and take a long time. The conveyed fault diagnosis techniques regarding the sensing principles include electroluminescence imaging, thermal imaging, time domain reflectometry, earth capacitance measurement, and electrical elements monitoring [6]. Machine learning (ML) techniques have been proposed due to the high performance provided by these techniques [7], [8]. ML techniques have been proposed due to the high performance provided by these techniques. ML is a collection of algorithms that permit software applications without complicated programming to predict products more accurately. This is done by modeling predictive models established on statistical mechanisms, which, set on providing input data, will predict outcomes and rework the availability of renewed input data. It utilizes different algorithms and techniques to reach expected results, such as regression, decision tree (DT), random forest (RF), k-nearest neighbors (KNN), logistic regression (LR), and support vector machine (SVM).

The current study will evaluate three fault types: OC, PS, and SC. Thus, this study aims to introduce an empirical mode (EM) created FD structure for PV designs multifarious. The EM contains three algorithms gradient boosting (GB), KNN and RF techniques have been employed for model investigations in PV designs.

2. RELATED WORKS

Chaibi *et al.* [9] suggested that PV systems detect four types of FD. These types are inverter disconnection (ID), PS, SC, and OC faults are the most frequent failures. This study proposes a straightforward, effective technique for detecting them. The suggested approach introduces three indicators: current, voltage, and power indicator, with the primary goal of identifying normal and abnormal operating circumstances. The best-yet artificial bees colony (ABC) optimization technique is used with the single-diode model to produce a trusted PV model and extract the unknown model parameters. The maximum power point (MPP) coordinates, which simulate the actual operational PV system, are then estimated. Current, voltage, and power indicators can be calculated using measured and expected MPP coordinates. Trial and error have been used to determine upper and lower criteria for each indication. The value of every indicator whether it is within, up, or lower than the threshold will signal the situation of the PV system and whether it works correctly or not. Different experiments were conducted using the studied data from a 3.2 kW grid-connected PV system mounted on Algeria’s renewable energy development centre (CDER).

Wang *et al.* [10] proposed SVM to detect faults in PV systems. The SVM algorithm utilized the OC voltage, SC current, maximum power voltage, and maximum power current as the main setting parameters. This selection was based on an analysis of faults and the I-V featured curves of PV arrays. The fault dataset was enhanced using data preparation techniques, providing high-quality data for the SVM algorithm’s efficiency. These parameters were optimized using grid search and k-fold cross-validation approaches. To test the proposed system, 400 data points were used, and the results of the evaluation showed a test accuracy of 97%. The results of the experiments demonstrated that the SVM-based fault diagnosis algorithm surpassed many algorithms in terms of accuracy.

While this paper Kalogerakis *et al.* [11] proposed a global maximum power point tracking (GMPPT) method aimed at locating the GMPPs focal point rapidly. Unlike traditional GMPPT methods, the approach proposed in this research uses a ML algorithm and does not require previous knowledge of the PV modules' operating properties or their configuration. This method is particularly effective in PV systems in that shading patterns can vary rapidly, such as in wearable PV systems or building-combined PV systems, because of its characteristic of learning ability that allows quicker detection of the GMPP with less search space solutions. Numerical results in the paper demonstrate that the designed Q-learning-based GMPPT algorithm reduces the execution time needed to detect the global MPP by 80.5% to 98.3% compared to a GMPPT method built based on the particle swarm optimization (PSO) algorithm when using unknown PS patterns.

Eskandari *et al.* [12] suggest a brand-new, intelligent fault monitoring mechanism. The critical characteristics of current-voltage (I-V) curves with various fault situations and typical circumstances are obtained for this purpose. Using the hierarchical classification (HC) framework, the faults are categorized. Later, ML techniques are used to identify and categorize the LL and LG problems. Compared to existing fault diagnostic approaches, the suggested method seeks to decrease the dataset needed for the training procedure and achieve advanced accuracy in recognizing and categorizing the occurrences of fault at low mismatch levels and high fault impedance. The experimental results show that, with an accuracy of 96.66% and 91.66%, the presented approach accurately classifies and categorizes LL and LG defects on PV systems over various situations and degrees of difficulty.

Kapucu and Cubukcu [13] proposed ensemble learning (EL) and a ML technique. By mixing the predictions of various algorithms, EL approaches strive to achieve more generalizability and prediction accuracy than a single ML algorithm. In this situation, grid-search with cross-validation is applied first to choose the most pertinent features [10]. Then, the parameter optimization of each learning method and the EL model that would integrate them has been enhanced. Results reveal that the suggested method has an excellent generalization capacity for PV system defect diagnosis and improves the classification performance with the correct data and optimum settings for each algorithm and the EL model.

Padmavathi *et al*. [14] developed a regression controller-based maximum power point tracking (MPPT) to attain maximum peak voltage under partial shade. Based on recorded datasets of PV system output voltage and load, the regression algorithm forecasts the cycle for the converter during partial shade effect or instant separation for that specific geographic location. In MATLAB R2018a Simulink, the regression-based duty cycle prediction system is designed. Regression system is also used in the test bench for PV systems. During partial shade conditions in PV, the simulation and hardware results of regression controller-based MPPT outperform PSO, flower pollination algorithm (FPA), and perturb and observe (P&O) algorithms by roughly 20%, 16.96%, and 15% in efficiency, respectively.

3. RESEARCH METHOD

The proposed creates an EM using the voting classifier class. Mainly, the proposed model is composed of three essential stages: the first is the preprocessing stage, and the EM combines the predictions as a second stage by implementing three individual classifiers: RF, GB, and KNN. Finally, three algorithms that have been suggested to predict the voting card need to be evaluated. Figure 1 explains the proposed model, where it shows the main stages and their details represented by the preprocessing (Labelencoder) at this stage, the dataset is divided into the training part, which has the largest percent (80%) and the lowest percent (20%) for the testing part. Then three algorithms are applied to both dataset parts to obtain the final prediction, which depends on the large voting result, and finally the evaluation. In the following sections, the main stages have been detailed.

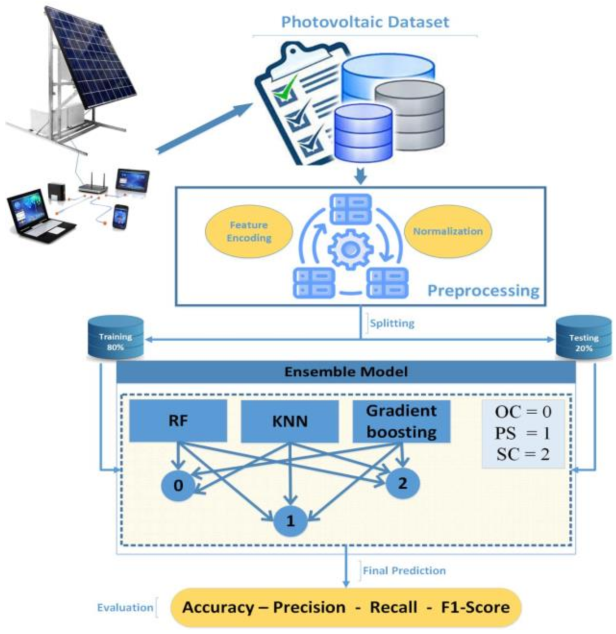


Figure 1. Research methodology

3.1. Dataset description

In this section PV fault dataset that has been used to validate the proposed model. Mendeley has this dataset available [15]. This dataset provides a thorough simulation using Python and simulation program with integrated circuit emphasis (SPICE) implements, including LTSpice. The different PS, OS, and SC factors were simulated using the Python-induced SPICE netlist. Subsequently, other datasets for additional configurations at different operating temperatures are generated by the LTSpice simulation.

That data is preserved in a comma-separated values (CSV) file containing 6965234 instances. Every sample contains ten attributes. Power measured as watts (W), shade voltage, full voltage, current measured as ampere (A), temperature, voltage measured as volt (V), parallel and series cells, and number of cells are some of these qualities. Table 1 describes an instance of the PV features that are used as illustrative features and as input for the ML methods.

Table 1. Input features for PV dataset [15]

	Voltage (V)	Current (A)	Power (W)	Full voltage	Shade voltage	Temperature	# of cells	Series cells	Parallel cells
0	3.73	2.09	7.80	1000	700	40	10	5	2
1	0.18	4.48	0.81	1000	600	30	10	2	5
2	0.34	5.98	2.03	1000	500	40	8	2	4
3	8.42	0.67	-5.60	1000	600	36	9	9	1
…	…	…	…	…	…	…	…	…	…
6965234	6.95	1.05	7.30	800	700	40	10	10	1
6965235	3.34	1.79	5.99	1000	600	38	10	5	2
6965236×10 columns									

The dataset delivers 6965236 instances; each instance has ten features that represent the PV panel. These features are current, voltage, power, full voltage, shade voltage, temperature, number of cells, series cells, and parallel cells. This study investigates three kinds of faults in PV (OC, PS, and SC).

3.2. Preprocessing

In real-world scenarios, datasets often come with issues like missing values, noise, and formats that are unsuitable for EM, necessitating preprocessing to make them usable for deep learning models. Preprocessing helps to enhance the performance and efficiency of EM methods by addressing these issues. Two key preprocessing operations in this context are data normalization and feature encoding.

3.2.1. Data normalization

This step adjusts the scale of features to ensure that all attributes participate evenly in the model’s learning process, and methods such as standardization, Z-score normalization, and Min-Max normalization are commonly used. In this research, a Python standard scaler function was used. This function standardizes data to fit within a particular range rather than keeping the relationships between data points. It transforms the data into a range among 0 to 1 [16].

3.2.2. Feature encoding

This is another preprocessing step for the EM that processes only numerical inputs. A feature encoding is applied to convert categorial values into numerical values. There are a lot of methods like OneHotEncoder and LabelEncoder are usually utilized for this purpose. For this proposed, the LabelEncoder was chosen. This method transfers unique integers to categorical values [17], [18]. For example, many fault types in the dataset were encoded as 0, 1, and 2, instead of ('OC', 'PS,' and 'SC'), enabling the model to read and utilize such values in its computations.

3.3. Data splitting

The final step in preparing the dataset involves data splitting, which is crucial for evaluating model performance. In this study, the dataset was separated into 80% for training and 20% for testing to evaluate the model’s performance. This splitting guarantees that the model is trained on a good portion of the data whilst preserving a subset to assess how well it deals with unseen examples keeping a robust evaluation of its effectiveness.

3.4. Proposed ensemble model

This study proposed an EM that contains of RF, GB, and KNN algorithms combined for classification tasks. RF is well known for its capability to deal with complex datasets and reduce overfitting, while GB creates weak models and combines them to make a strong predictive model. On the other hand, KNN categorizes new cases based on their KNN. The EM controls the power of every algorithm to boost the accuracy of the system. However, the efficiency of the model depends on factors such as model quality, diversity, and ensemble technique.

3.4.1. Random forest

RF is a commonly utilized in EL algorithm for both classification and regression tasks [19], [20]. It enhances accuracy via aggregating the results of many decision trees and combining them to produce a final prediction, making it effective for complex datasets. The core concept is to estimate a classification function, where a random vector X is represented as the input, and the output is the square-integral random response Y. This Y is computed using an estimating function n(x)=E[Y |X=x]. An RF is a predictor consisting of N-randomized classification trees, with the ith tree's predicted value at the query point denoted by nm (x;Θi,Tm). The parameter Θ is utilized to resample the training set before developing individual trees and selecting successive directions for splitting.

$$n_m(x;\Theta_i\,.T_m\,) = \sum_{j\in T_m(\Theta_i)} \frac{X_j\in A_m(x\,;\,\Theta_i\,.T_m)Y_j}{M_m(x\,;\,\Theta_i\,.T_m)} \tag{1}$$

Where $T_m(\Theta_i)$ is the set of data points chosen before tree creation, $A_m(x\,;\,\Theta_i\,.T_m)$ is the cell including x, and $M_m(x\,;\,\Theta_i\,.T_m)$ is the set of the points that fall into $A_m(x\,;\,\Theta_i\,.T_m)$. At this step, a set of trees are conquered to construct the forest estimate [15]:

$$n_{N.m}(x;\ \Theta_1.\cdots\cdots.\Theta_N.T_m) = \frac{1}{N}\sum_{i=1}^{N} M_m(x\,;\,\Theta_i\,.T_m) \tag{2}$$

In the T package for RF, the number of trees (N) can be arbitrarily large, constrained only by available computing resources. From a modeling perspective, it is beneficial to allow N to approach infinity, thereby considering the estimate of an (infinite) forest instead of a finite number of trees:

$$n_{\infty.m}(x;\ T_m) = E_\Theta[n_m(x;\ \Theta.T_m\ ]$$

The E_Θ refer to the expectation including detail to random parameter Θ, restricted on Sm. The operation "$N\ \to \infty$" is explained via the theory of big numbers, which declares that this expectation converges, restricted to Sm:

$$\lim_{N\to\infty} m_{N.m}(x;\ \Theta_{1.}\ldots\ldots.\Theta_N.T_m = n_{\infty.m}(x;\ T_m) \tag{3}$$

3.4.2. Gradient boosting

GB is a popular EL approach that has gained recognition for its efficiency in solving various ML tasks [21]. By merging different weak predictive models, GB makes a predictive model capable of handling complicated datasets and capturing non-linear relationships. Its ability to provide high predictive accuracy has made it widely used in domains such as finance, healthcare, and natural language processing.

However, a point to account to is that GB can be expensive and sensitive to hyperparameter tuning. Careful optimization of the algorithm's parameters is necessary to achieve optimal performance. This process involves finding the right balance between model complexity and generalization ability.

$$f(x) = \sum_{m=1}^{M}\beta_{j_m} b(x;T_{j_m}) \tag{4}$$

Where all basis function is $b(x;T) \in$ R it is a naive function of the feature vector indexed via a parameter T and $\beta_j$is the weak learner which is a constant of the jth. β_{j_m} with T_{j_m} are chosen in an adaptive manner to improve data consistency. Let $l(y.f(x))$ a rate of data fidelity at the observation (y, x) for the loss function l, expected to be distinct in the second manage. One of the main goals of ML is to develop a function f that minimizes the accepted loss E_p(EP ($l(y.f(x))$)), where the probability is taken over the unknown distribution of (y, x) (denoted by P). One way to achieve this goal stays to study the practical loss and nearly minimize it utilizing algorithms that finds a high value of f via nearly minimizing the realistic loss:

$$\min_f \sum_{j=1}^{m} l\left(y_j.f(x_j)\right) \tag{5}$$

where the $l\left(y_j.f\left(x_j\right)\right)$ used to process data conformity for the ith sample $(y_j.x_j)$. The previous version of GBM. The GBM algorithm can be presented:

initialization: initialize with:

$$f^0(x) = 0$$

for m=0,….. M-1 do:Compute pseudo-residual:

$$r^m = -\left[\frac{\sigma l\left(y_j.f^m(x_j)\right)}{\sigma f^m(x_j)}\right]_{j=1.\cdots\cdots.m} \quad (6)$$

discover the best weak learner:

$$j_m = arg\ min_j\ min_\sigma \sum_{j=1}^{m} \left(r_i^m - \sigma b\left(x_{i;}\, T_j\right)\right)^2 \quad (7)$$

select the step-size p_m by line search:

$$p_m = arg\ min_p \sum_{j=1}^{m} l\left(y_j.f^m\left(x_j\right)\right) + pb\left(x_j;\ T_{j_m}\right) \quad (8)$$

update the model

$$f^{m+1}(x) = f^m(x) + p_m b\left(x;\ T_{j_m}\right) \quad (9)$$

output:

$$f^m(x) \quad (10)$$

3.4.3. K-nearest neighbors

KNN is one of ML algorithms which is usually applied for both regression and classification tasks. The widespread of this algorithm is due to its simplicity and effectiveness [22]. It is considered a non-parametric algorithm because it does not make any assumptions about the underlying data distribution. Instead, KNN makes predictions based on distance measurements between input data points. By combining the predictions from multiple classifiers, the EM aims to enhance prediction accuracy and generalization. Each classifier possesses its own strengths and weaknesses, and by leveraging their collective knowledge, the EM can generate more robust predictions.

In ML issues, a method to characterize data points is crucial. A feature vector, with length M, is a mathematical presentation of data with M unique attributes. KNN classifies new objects based on attributes and training data. To class new items, KNN performs specific phases:

- Estimate the space between the item to be classified with every point in the dataset training.
- Choose the nearest data points with the K lowest space.
- Performing a “majority vote” between those data points.

The KNN mathematical model uses the entire training dataset for prediction, exploring multiple equivalent samples. The Euclidean distance formula is used to determine which K samples are more equivalent to the new object [16]:

$$d(p.q) = \sum_{j=1}^{m} (p_k - q_k)^2 \quad (11)$$

where $d(p.q)$ is the space among testing and training object.

3.5. Evaluation

Precision, recall, and F-score are widely used performance metrics used to assess how a ML model is efficient. Such metrics provide important facts on several situations related to the model performance. The first metric, which is precision, measures the model ability to appropriately detect right outcomes out of the total predicted positives outcomes, showing how low is false positives outcomes. Recall, on the other hand, evaluates the model's ability to identify all positive cases out of the total actual positives, highlighting a low false negative rate [23].

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives)}} \tag{12}$$

Recall: recall is the ratio of correctly predicted positive instances to the total actual positive instances. It is a measure of the model's ability to find all positive instances. High recall means that the model has a low false negative rate [24].

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \tag{13}$$

The other metric is F-score measurement. It represents the mean of precision and recall. The result of such metric is a score that shows the model's performance when optimizing both precision and recall is desired. A high score of F-score value means high scores of precision and recall [25].

$$\text{F} - \text{measure} = \frac{2\times(\text{Precision} * \text{Recall})}{\text{Precision+Recall}} \tag{14}$$

In the accuracy measurement in (15), the score is calculated based on true positives, false positives, true negatives, and false negatives. Such values collected from the model’s outcomes and ground truth dataset [26].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

4. RESULT AND DISCUSSION

The results of this work highlight the efficiency of the proposed EM which combines RF, KNN, and GB for fault detection in PV systems. The model achieved an accuracy of 95% in classifying three major PV fault types (open circuit, shaded, and short circuit) utilizing a dataset of 6,965,236 samples collected in 2023. This remarkable accuracy demonstrates the models ability to effectively handle large-scale datasets and reliably diagnose fault types, making it a valuable tool for real-world PV system monitoring and maintenance.

The training and validation accuracy curves, shows in Figure 2, provide visual evidence of the model's learning progression. Both metrics increased steadily with successive epochs, indicating efficient learning from the training data and strong generalization to unseen data. Accompanying this trend, the consistent decrease in training and validation losses confirms the model’s ability to minimize errors and converge toward an optimal solution. These patterns collectively validate the robustness and reliability of the model's training process.

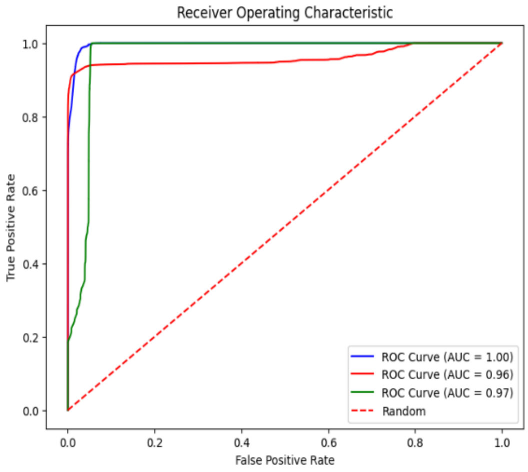


Figure 2. Training and validation accuracy

The models confusion matrix, shown in Figure 3, further supports its classification accuracy. The diagonal elements, representing correct predictions, dominate the matrix, while off-diagonal elements, indicating false positives or negatives, and remain minimal. This outcome confirms the model's high precision in distinguishing between fault types with negligible misclassifications.

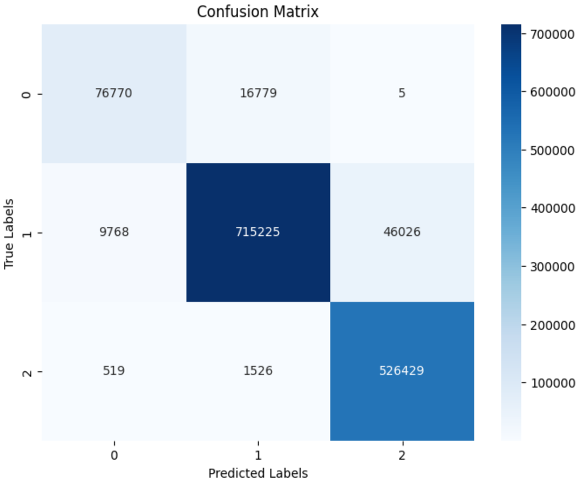


Figure 3. A confusion matrix of the proposed model

Table 2 shows the evaluation of the proposed model, summarizing its performance and comparison of our model with other models. Where accuracy, precision, recall, and F-measure were all equal to 95%. The results presented shows the EM efficiency in detecting and classifying faults in PV systems. This analysis proves the ability of the model to accurately diagnose faults and provides a foundation for future enhancements and applications in solar energy fault detection systems.

Table 2. Results of the proposed model

	Precision	Recall	F-measure	Accuracy
Proposed model	0.95	0.95	0.95	0.95

Performance metrics such as precision, recall, F-measure, and accuracy reached 95%, as detailed in Table 3. This consistent performance across multiple evaluation criteria underscores the model's reliability, not only in detecting faults but also in minimizing both false positives and false negatives. The proposed EM metrics surpass those of existing approaches, such as the study by Ramaprabha and Gokularaman [27], which achieved an accuracy of 91.91% using a decision tree-based approach. Table 3 compares the performance of various models, highlighting the superiority of the ensemble approach in handling complex fault scenarios. Compared to traditional models like LR (67%), naive Bayes (NB) (62%), and SVM (79%), the EM demonstrates a clear advantage by leveraging the strengths of multiple algorithms. By combining RF, KNN, and GB, the model benefits from diverse learning mechanisms, improving 0its ability to adapt to complex and diverse fault patterns. This integration results in enhanced diagnostic accuracy and robustness, essential for applications in solar energy maintenance.

Table 3. Accuracy of different models for fault detection

Study	Model	Accuracy (%)
Ramaprabha and Gokularaman [27]	NB	62
	LR	67
	SVM	79
	RF	83
	DT	91.91
Proposed model	EM combined (RF, GB, and KNN)	95

The study’s findings have profound implications for the solar energy sector. The EM ability to detect faults with high accuracy supports early identification, which is critical for minimizing system downtime and ensuring uninterrupted energy production. This capability also reduces maintenance costs by enabling precise interventions rather than broad and costly system inspections.

The models scalability and generalizability make it suitable for large-scale deployment across various PV systems and environmental conditions. Its success sets the stage for adapting it to diverse operational scenarios, enhancing its practical utility in global solar energy initiatives.

The EM strengths include its high performance across evaluation metrics, scalability to handle extensive datasets, and adaptability to various PV fault scenarios. However, the study’s limitations must be acknowledged. The dataset, while comprehensive, may not fully capture all fault types or environmental conditions encountered in diverse geographical regions. Furthermore, the model currently addresses only three fault categories. Expanding its scope to include additional faults would increase its applicability in real-world conditions. In the future incorporating deep learning techniques or hybrid models that combine deep learning with traditional ML methods could further improve detection accuracy, particularly under diverse and dynamic operational conditions. Additionally, the use of more diverse datasets would allow for better adaptation to various PV system configurations and environmental influences, paving the way for more robust and adaptable fault detection systems in the future.

5. CONCLUSION

The findings of this paper have significant implications for the research field of PV fault detection and for the broader solar energy community. By demonstrating that an EM combining RF, KNN, and GB achieves 95% accuracy in classifying common PV faults-open circuit, shaded, and short circuit-this research offers a robust and scalable solution for early fault detection in PV systems. This is important for the solar energy industry, as it can help reduce downtime, minimize maintenance costs, and ultimately improve the efficiency and longevity of solar power systems. The ability to detect faults early in the operational lifecycle enables more proactive maintenance, reducing the need for costly emergency repairs and maximizing energy production. In the research field, these results contribute to the growing body of work in ML-based fault detection and provide a strong case for the efficiency of EL approaches in real-world applications. For the solar energy community, the model’s high performance demonstrates the potential of automated fault detection systems to optimize the maintenance of PV installations, which is critical as solar energy adoption continues to expand. This research paves the way for future advancements, including the integration of deep learning techniques or hybrid models that could enhance detection capabilities and adaptability to diverse fault types and environmental conditions, further improving the resilience and sustainability of solar power systems globally.

REFERENCES

[1] J. C. R. Kumar and M. A. Majid, “Floating solar photovoltaic plants in India – A rapid transition to a green energy market and sustainable future,” *Energy and Environment*, vol. 34, no. 2, pp. 304–358, 2023, doi: 10.1177/0958305X211057185.

[2] G. M. El-Banby, N. M. Moawad, B. A. Abouzalm, W. F. Abouzaid, and E. A. Ramadan, “Photovoltaic system fault detection techniques: a review,” *Neural Computing and Applications*, vol. 35, no. 35, pp. 24829–24842, 2023, doi: 10.1007/s00521-023-09041-7.

[3] F. Aziz, A. Ul 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, doi: 10.1109/ACCESS.2020.2977116.

[4] A. Malik, A. Haque, V. S. B. Kurukuru, M. A. Khan, and F. Blaabjerg, “Overview of fault detection approaches for grid connected photovoltaic inverters,” *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 2, p. 100035, 2022, doi: 10.1016/j.prime.2022.100035.

[5] R. Fazai *et al.*, “Machine learning-based statistical testing hypothesis for fault detection in photovoltaic systems,” *Solar Energy*, vol. 190, pp. 405–413, 2019, doi: 10.1016/j.solener.2019.08.032.

[6] U. Hijjawi, S. Lakshminarayana, T. Xu, G. P. M. Fierro, and M. Rahman, “A review of automated solar photovoltaic defect detection systems: Approaches, challenges, and future orientations,” *Solar Energy*, vol. 266, p. 112186, 2023, doi: 10.1016/j.solener.2023.112186.

[7] S. Singh, S. K. Patro, and S. K. Parhi, “Evolutionary optimization of machine learning algorithm hyperparameters for strength prediction of high-performance concrete,” *Asian Journal of Civil Engineering*, vol. 24, no. 8, pp. 3121–3143, 2023, doi: 10.1007/s42107-023-00698-y.

[8] M. Al-Mahdawi, A. K. Kareem, A. A. Nafea, A. M. Shaban, S. A. S. Aliesawi, and M. M. Al-Ani, “An Effective Deep Learning Approach for the Estimation of Proton Energy by Using Artificial Neural Network,” in *2024 21st International Multi-Conference on Systems, Signals and Devices, SSD 2024*, 2024, pp. 257–262, doi: 10.1109/SSD61670.2024.10549706.

[9] Y. Chaibi, M. Malvoni, A. Chouder, M. Boussetta, and M. Salhi, “Simple and efficient approach to detect and diagnose electrical faults and partial shading in photovoltaic systems,” *Energy Conversion and Management*, vol. 196, pp. 330–343, 2019, doi: 10.1016/j.enconman.2019.05.086.

[10] J. Wang, D. Gao, S. Zhu, S. Wang, and H. Liu, “Fault diagnosis method of photovoltaic array based on support vector machine,” *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 45, no. 2, pp. 5380–5395, Jun. 2023, doi: 10.1080/15567036.2019.1671557.

[11] C. Kalogerakis, E. Koutroulis, and M. G. Lagoudakis, “Global MPPT based on machine-learning for PV arrays operating under partial shading conditions,” *Applied Sciences (Switzerland)*, vol. 10, no. 2, p. 700, 2020, doi: 10.3390/app10020700.

[12] A. Eskandari, J. Milimonfared, and M. Aghaei, “Fault Detection and Classification for Photovoltaic Systems Based on Hierarchical Classification and Machine Learning Technique,” *IEEE Transactions on Industrial Electronics*, vol. 68, no. 12, pp. 12750–12759, Dec. 2021, doi: 10.1109/TIE.2020.3047066.

[13] C. Kapucu and M. Cubukcu, “A supervised ensemble learning method for fault diagnosis in photovoltaic strings,” *Energy*, vol. 227, p. 120463, 2021, doi: 10.1016/j.energy.2021.120463.

[14] N. Padmavathi, A. Chilambuchelvan, and N. R. Shanker, “Maximum Power Point Tracking During Partial Shading Effect in PV System Using Machine Learning Regression Controller,” *Journal of Electrical Engineering and Technology*, vol. 16, no. 2, pp. 737–748, 2021, doi: 10.1007/s42835-020-00621-4.

[15] K. Sood, N. Ruppert, and R. Mahto, “Partial Shading and Fault Simulation Dataset of Photovoltaics Module,” *IEEE Dataport*, 2022, doi: 10.21227/fjbq-0321.

[16] B. Deepa and K. Ramesh, “Epileptic seizure detection using deep learning through min max scaler normalization,” *International Journal of Health Sciences*, vol. 6, pp. 10981–10996, 2022, doi: 10.53730/ijhs.v6ns1.7801.

[17] J. Bai, S. Kong, and C. Gomes, “Gaussian Mixture Variational Autoencoder with Contrastive Learning for Multi-Label Classification,” in *Proceedings of Machine Learning Research*, 2022, vol. 162, pp. 1383–1398.

[18] N. Bhattacharya, A. Subudhi, S. Mishra, V. Sharma, A. P. Aderemi, and C. Iwendi, “A Novel Ensemble based Model for Intrusion Detection System,” in *Proceedings - International Conference on Computing, Power, and Communication Technologies, IC2PCT 2024*, 2024, vol. 5, pp. 620–624, doi: 10.1109/IC2PCT60090.2024.10486584.

[19] O. J. Kadhim, A. A. Nafea, S. A. S. Aliesawi, and M. M. Al-Ani, “Ensemble Model for Prostate Cancer Detection Using MRI Images,” in *Proceedings - International Conference on Developments in eSystems Engineering, DeSE*, 2023, pp. 492–497, doi: 10.1109/DeSE60595.2023.10468866.

[20] H. Tao *et al.*, “Development of integrative data intelligence models for thermo-economic performances prediction of hybrid organic rankine plants,” *Energy*, vol. 292, p. 130503, 2024, doi: 10.1016/j.energy.2024.130503.

[21] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, 2021, doi: 10.1007/s10462-020-09896-5.

[22] M. Bansal, A. Goyal, and A. Choudhary, “A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning,” *Decision Analytics Journal*, vol. 3, p. 100071, 2022, doi: 10.1016/j.dajour.2022.100071.

[23] O. A. Egaji, G. Evans, M. G. Griffiths, and G. Islas, “Real-time machine learning-based approach for pothole detection,” *Expert Systems with Applications*, vol. 184, p. 115562, 2021, doi: 10.1016/j.eswa.2021.115562.

[24] M. S. I. Alsumaidaie, K. M. A. Alheeti, and A. K. Al-Aloosy, “Intelligent Detection System for a Distributed Denial-of - Service (DDoS) Attack Based on Time Series,” in *Proceedings - International Conference on Developments in eSystems Engineering, DeSE*, 2023, vol. 2023, pp. 445–450, doi: 10.1109/DeSE58274.2023.10100180.

[25] A. A. Nafea, M. AL-Mahdawi, K. M. A. Alheeti, M. S. I. Alsumaidaie, and M. M. AL-Ani, “A Hybrid Method of 1D-CNN and Machine Learning Algorithms for Breast Cancer Detection,” *Baghdad Science Journal*, 2024, doi: 10.21123/bsj.2024.9443.

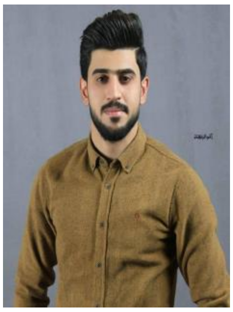
[26] M. Mahmood, F. M. Jasem, A. A. Mukhlif, and B. Al-Khateeb, “Classifying cuneiform symbols using machine learning algorithms with unigram features on a balanced dataset,” *Journal of Intelligent Systems*, vol. 32, no. 1, p. 20230087, 2023, doi: 10.1515/jisys-2023-0087.

[27] R. Ramaprabha and S. R. Gokularaman, “Analysis and Modification of Fault Detection Methods in Photovoltaic Array,” in *2nd International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE 2024*, 2024, pp. 1–6, doi: 10.1109/ic-ETITE58242.2024.10493547.

BIOGRAPHIES OF AUTHORS

Mohammed Salah Ibrahim is a computer science teacher who finished his Ph.D. from University of Arkansas-Fayetteville from United States of America in 2021. He got his M.Sc. from the Department of Computer Science, University of Anbar in 2011. He has published many papers in different fields and his major interest in information retrieval and text processing area. He can be contacted at email: moh.salah@uoanbar.edu.iq.

Hussein K. Almulla is a tech developer at the Center of Computation at University of Anbar. He finished his Ph.D. from University of South Carolina from United States of America in 2021. He got his M.Sc. from the Department of Computer Science University of Anbar in 2011. He has published many papers in different fields and his major interest in image processing and machine learning, and deep learning. He can be contacted at email: hussein.k.almulla@uoanbar.edu.iq.



Anas D. Sallibi hold Master’s degree from the University of Anbar, Iraq in 2024. He hold a Bachelor's degree from the University of Anbar, Iraq. He currently working as an Assistant lecturer at the Department of Computer Sciences, College of Sciences, University of Al Maarif, Al Anbar, 31001, Iraq. His research interest is in the areas of data science, deep learning, machine learning, artificial intelligence, big data, and data. He can be contacted at email: anas.diab@uoa.edu.iq.

Ahmed Adil Nafea hold Master’s degree from Universiti Kebangsaan Malaysia, Malaysia in 2020. He holds a Bachelor's degree from the University of Anbar, Iraq. He was born in Al-Anbar, Iraq on October 7, 1995. He currently working as an Assistant Lecturer at the College of Computer Science and Information Technology in the University of Anbar, Ramadi, Iraq. His research interest is in the areas of data science, deep learning, machine learning, natural language processing, artificial intelligence, computer vision, reinforcement learning, data analysis, statistical techniques, data preprocessing, big data, data visualization, and linear algebra. He can be contacted at email: co.khattab.alheeti@uoanbar.edu.iq.

Aythem Khairi Kareem hold Master's degree from the University of Anbar, Iraq in 2021. He hold a Bachelor's degree from the University of Anbar, Iraq. He currently working as an Assistant lecturer at the Department of Heet Education, General Directorate of Education in Anbar, Ministry of Education, Heet, Anbar, Iraq. His research interest is in the areas of data science, deep learning, machine learning, natural language processing, artificial intelligence, computer vision, reinforcement learning, data analysis, statistical techniques, data preprocessing, big data, data visualization, and linear algebra. He can be contacted at email: ayt19c1004@uoanbar.edu.iq.

Khattab M. Ali Alheeti a computer scientist working at the security for autonomous systems, self-driving vehicles and drones. obtained his M.Sc. in Information Technology from Al Al-Bayt University, Mafreq, Jordan, in 2008, and his Ph.D. in Computer Science from the University of Essex, Colchester, United Kingdom, in 2017. He carried out a doctoral degree in Computer Science at the University of Essex. Since December 2017, he has been a lecturer in the Department of Computer Science, College of Computer, University of Anbar, where he has taught in the B.Sc. (Computer Science) program. His personal research interests include security, image processing, bioinformatics and computational technology, Petri nets modeling and simulation, data mining, and fuzzy logic applications to predictive modeling. He has authored over 65 publications, including more than 25 journal papers and 35 refereed international conference papers. He is a member of the IEEE Computer Society and was the winner of the 2017 Best Researcher Award from the Ministry of Higher Education and Scientific Research. He can be contacted at email: co.khattab.alheeti@uoanbar.edu.iq.