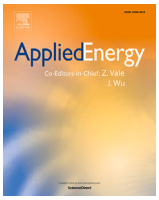
Applied Energy 393 (2025) 126108

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apen



A multi-stage review framework for AI-driven predictive maintenance and fault diagnosis in photovoltaic systems

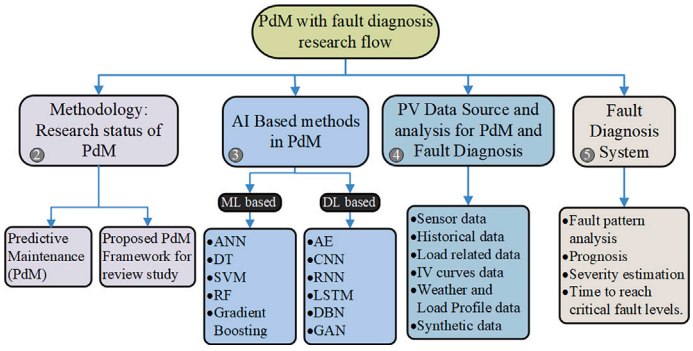
Ali Hamza [a], Zunaib Ali [a,*], Sandra Dudley [a], Komal Saleem [b], Muhammad Uneeb [c], Nicholas Christofides [d]

[a] School of Engineering and Design, London South Bank University (LSBU), London, UK
[b] Department of Engineering and Construction, University of East London (UEL), London, UK
[c] School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan
[d] Department of Electrical Engineering, Computer Engineering and Informatics, Frederick University, Nicosia, Cyprus

HIGHLIGHTS

- Discussion on PdM and future forecasting of the faults.
- AI-based PdM and fault diagnosis for PV System.
- PdM potential framework for PV System.
- PV data sources and analysis.
- Experimental validation of AI algorithms for PdM and fault diagnosis.
- Discussion about description of data and data types applied for PdM and fault diagnosis.
- Challenges of data standardization and its effects on AI performance for PdM and fault diagnosis.
- Exploration of PdM/anomaly detection concerning specific classes and future forecasting of faults.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
AI algorithms
Data analysis
Fault diagnosis
Predictive maintenance

ABSTRACT

The photovoltaic (PV) sector encounters challenges such as high initial costs, reliance on weather, susceptibility to faults, irregularities in the grid, and degradation of components. Predictive maintenance (PdM) aims to proactively identify issues, thereby enhancing reliability and efficiency but may lack specific fault details without additional diagnostic efforts. This research presents an advanced PdM and fault diagnosis framework that integrates fault pattern analysis, severity assessments, and critical fault predictions. It aims to improve the functionality of PV systems, minimize downtime, and enhance reliability by identifying and analyzing specific fault patterns. Consequently, our article provides a critical review of current Artificial Intelligence (AI) methodologies for PdM and fault diagnosis in PV systems. Moreover, this study highlights the significance of data standardization and offers recommendations on how PdM, when combined with fault diagnosis, can utilize various data sources to anticipate faults in advance, assess their severity, and optimize system performance and maintenance activities. To the best of the authors' knowledge, no such review study exists.

* Corresponding author.
Email address: aliz29@lsbu.ac.uk (Z. Ali).

https://doi.org/10.1016/j.apenergy.2025.126108
Received 5 December 2024; Received in revised form 15 March 2025; Accepted 10 May 2025
Available online 21 May 2025

Nomenclature

1SVM One-class Support Vector Machine
ABC Artificial Bee Colony
AE Autoencoder
AI Artificial Intelligence
ANFIS Adaptive neurofuzzy inference system
ANN Artificial Neural Network
ARIMA Autoregressive Integrated Moving Average
CBM Condition-based Maintenance
CM Condition Monitoring
CNN Convolution Neural Network
CT Current Transformer
DAQ Data Acquisition
DKASC Desert Knowledge Australia Solar Center
DL Deep Learning
DSS Decision Support System
DT Decision Tree
EFB Exclusive feature bundling
EL Electroluminescence
FCNN Fully Connected Neural Network
FES Fuzzy Expert System
FPGA Field programmable gate array
GAN Generative adversarial network
GCPS Grid-connected photovoltaic System
GHI Global Horizontal Irradiance
GOSS Gradient-based one-side sampling
GTI Global tilted irradiance
ID one-dimensional
IMS Intelligent Monitoring System
IoT Internet of Things
IR Infrared Radiation
KNN K-Nearest Neighbors
KPI Key Performance Indicator
kVA Apparent power
kVAR Reactive power
KW/MW Active Power
KWh Power generation
LightGBM Light Gradient Boosting Machine
LL Line-to-Line
LSTM Long short-term memory
MAE Mean absolute error
MAPE Mean absolute percentage error
ML Machine learning
MLP Multi-layer perceptron
MPPT Maximum Power Point Tracking
MSE Mean square error
MSRs Maintenance Strategies Routines
NPL Natural processing languages
O&M Operations and Maintenance
OC Open Circuit
ODM Operational Design Model
PCA Principal Components Analysis
PCS Power conversion system
PdM Predictive Maintenance
PID Pelvic inflammatory disease
PM Preventive Maintenance
PMU Phasor measurement unit
PS Partial Shading
PSIM Physical security information management
PV Photovoltaic
PVGIS PV Geographical Information System
R^2 coefficient of determination
RF Random Forest
RLU Rectified Linear Unit
RM Reactive Maintenance
RNN Recurrent Neural Network
RUL Remaining Useful Life
SADA Solar Array Drive Assembly
SAPVS Stand-alone PV system
SC Short Circuit
SVM Support Vector Machine
UAV Unmanned aerial vehicle
VER Vector Auto-regression
VT Voltage Transformer

1. Introduction

Photovoltaic (PV) systems are a prominent renewable energy technology known for their modular design and flexibility in meeting diverse electrical needs. By 2030, the globally installed PV capacity is projected to reach 630 GW [1]. However, PV systems face several challenges, such as natural wear and tear, component malfunctions (including faults and issues like cracked modules, hotspots, inverter problems, and tracker misalignment), adverse weather impacts (such as snow, dirt buildup, and wind), along with other concerns (like tightening loose cable connections, changing fuses, fixing SCADA errors, and addressing tracker failures). Protective measures like over-current and ground fault protection are essential, but nonlinear PV characteristics and complications from shading limit fault detection capabilities for PV systems. It also poses inherent complexity [2], often featuring nonlinearity [3], uncertainty [4], time-dependent correlations [5], multimodality [6], multi-periodicity [7], largeness in scale [8], or intermittent attributes [9], leading to problems and challenges in managing the acquired data for PdM and fault diagnosis. Apart from that, outliers [10] and measurement noise in PV system data significantly impact the reliability and accuracy of data analysis, modeling, and decision-making processes [11]. Consequently, these factors can lead to undetected faults and continuous energy losses, highlighting thus the necessity for efficient monitoring and fault diagnosis strategies for PV systems [12].

Research on PV system monitoring and fault diagnosis is growing due to technological advancements and improved access to data. We focus on techniques for timely detection and classification of PV component faults to maintain system functionality and cost-effectiveness. Consequently, three major maintenance strategies are identified as Preventive Maintenance (PM), Reactive Maintenance (RM), and Predictive Maintenance (PdM). PdM stands out in assessing plant health through condition monitoring, minimizing thus, operational costs and enhancing equipment longevity through AI algorithms that pre-emptively address faults [13]. It is considered the best approach for industrial component maintenance [14]. Various PdM methods have been suggested for evaluating plant health utilizing specific matrices and analyzing the progression of monitoring data. These techniques can be divided into physical model-based, knowledge-based, and data-driven approaches [15]. Among these, data-driven models, in particular, are pertinent to our research, as they solely depend on the monitoring data collected and the history of maintenance events. These models utilize Machine Learning (ML) algorithms designed to detect ideal plant operations and pinpoint failures. By examining both historical and real-time

data, they can forecast possible equipment malfunctions, allowing for proactive maintenance to avoid damage. Nevertheless, effective PdM encounters obstacles like data quality issues, missing or redundant attributes [9], outliers and measurement noise [10,11] and the necessity for extensive interconnected historical data.

While PdM offers early alerts for faults, it may not provide detailed information on specific fault types without additional diagnostic techniques. PV systems exhibit various fault types, ranging from major to minor, with differing initiation speeds and severities that often vary over time. Combining PdM with fault diagnosis offers a holistic approach to addressing these issues, enabling precise fault prediction and identification. This approach is further guided by fault severity and patterns, encompassing faults such as module cracking, hotspots, inverter failures, tracker misalignment, weather-induced issues (e.g., snow, soiling, wind), and operational challenges like loose cable connections, fuse replacements, SCADA faults, and tracker repairs [16]). Predicting and identifying these faults ahead of time ensures timely maintenance and enhanced system reliability. Likewise, analyzing fault patterns and estimating the timeline before reaching a critical stage enables accurate fault diagnosis and improved maintenance planning. Moreover, it is crucial to tackle challenges related to sensor degradation, sampling frequency, and variations in environmental conditions. Long-term deterioration or drift in sensors, often caused by extended exposure to sunlight, can lead to inaccuracies in essential data needed for fault identification and prediction.

To overcome existing challenges and enhance PdM and fault diagnosis in PV systems, this research proposes an innovative multi-stage framework that integrates fault pattern analysis, prognosis, severity estimation, and predictions regarding the time to reach critical fault levels. Therefore, in this paper, authors attempt to review state-of-the-art work carried out by researchers within each stage of the proposed review framework (Fig. 2) leading to guidelines, helping engineers and researchers develop techniques and methods for advancing PdM. Consequently, addressing the entire PdM from the fault diagnosis perspective, emphasizing the comprehensive data collection process, nature of data, and diverse AI algorithms in designing a PdM model for PV systems. Various review studies exist in the literature around PdM and fault diagnosis but with certain limitations. Table 1 provides an overview of the review papers that focus on the application of AI to PV system conditioning monitoring and fault diagnosis analysis (grouped into two categories, i.e., Fault Detection and Diagnosis, and PdM). Several important research gaps and limitations have been identified. Furthermore, Table 1's last row compares our review study to previous review papers in the field, highlighting our notable contributions to this area.

The contributions of this study are given below:

- This research emphasizes the significance of AI-driven PdM and fault diagnosis algorithms in establishing a robust framework to maintain the optimal functionality of the PV system.
- This work Investigates anomaly detection, Condition Monitoring (CM), and PdM in the context of the classification of future fault forecasting simultaneously. This can serve as a valuable foundation for future studies into the advancement of the PdM.
- This study signifies various sources for PV system-related data and PV measurement data to support fault diagnosis and PdM efforts. This study establishes several data categories such as sensor, historical, weather, IV curves, synthetic, and load-power demand data. It also focuses on the data standardization challenges and how they impact AI performance for PdM and fault diagnosis.
- This study explores the experimental validation of data collection methods for PV systems, coupled with hardware analysis of AI systems for PdM and fault diagnosis, ensuring the robustness and effectiveness of the proposed methodologies in real-world applications.
- This work reviews existing PdM and fault diagnosis techniques comprehensively based on data-driven algorithms, providing insights into state-of-the-art methodologies and their applicability to PV systems.
- Various AI-based algorithms for PdM and fault diagnosis in the context of PV system components, input and output settings, data types, advantages and drawbacks have been critically evaluated in this study.
- This work presents a comparison of AI-based approaches for fault diagnosis, highlighting aspects such as the severity of faults, fault patterns, and the estimated time until a critical condition is reached.
- This study analyzes fault diagnosis methodologies in terms of their potential for future prediction, alignment with PdM objectives, real-world hardware implementation, and verification processes, as well as delineation of the inputs and outputs of data-driven models, facilitating informed decision-making and proactive maintenance strategies.

The proposed work's graphical overview and article structure are depicted in Fig. 1. Section 2 presents the methodology and research status of PdM. Section 3 discusses the traditional AI approaches applied in PdM and fault diagnosis. Section 4 provides a detailed overview of PV data sources and analysis. Section 4.6 describes experimental setup and verification of data collection and ML/DL hardware implementation. Fault diagnosis of PV systems is discussed in Section 5. Section 6 focuses on the discussion and future recommendations. Finally, the conclusion is presented in Section 7.

2. Methodology: research status of PdM

This section examines the present state of ML-based PdM algorithms, focusing on pertinent questions that shape our understanding and investigation:

- Which ML techniques are used for PdM?
- What data is required to enable PdM?
- What is the source of data (synthetic or realistic)?
- How are the ML techniques used for PdM of PV systems?
- Do the ML algorithms for PdM primarily offer fault diagnosis or anomaly detection capabilities?
- Is the feasibility of ML-based PdM algorithms in real-time applications supported by experimental verification?

To find appropriate articles for review, a search was conducted on Google Scholar using a range of keywords like 'predictive maintenance', 'fault diagnosis', 'condition monitoring', 'anomaly detection', 'defect', 'failure', 'PV System', 'AI', 'ML', 'DL', 'classification', 'prediction', 'various types of faults', 'various types of AI and ML/DL', 'data nature'. Based on the search results, over 150 articles were identified between 2018 and 2024 and included in the present review.

2.1. Predictive maintenance (PdM)

PdM, sometimes called condition-based maintenance (CBM) [34], focuses on predicting equipment failures and determining the necessary maintenance actions to balance service frequency and costs effectively, (as shown in Fig. 2). PdM uses the system's real operating conditions and components for Operation and Maintenance (O&M) optimization [35]. Data from measurement devices installed on the PV system including temperature, irradiance, current, voltage, power output, etc. are used for predictive analysis.

To avoid unexpected outages and costs from downtime a well-implemented, sensitive, and efficient maintenance strategy is required. Numerous systems these days continue to depend on spreadsheets or even handwritten records to monitor equipment, effectively taking a reactive process towards maintenance. Consequently, intermittent downtime is anticipated and frequently encountered. However, many

Table 1
Research status of PdM and fault diagnosis.

Category	Reference	Methods	Features	Limitation(s)
Fault detection and diagnosis	[17], 2021	DL	**(A)**. Reviewed most frequent Deep Learning (DL) methods for fault detection and diagnosis. **(B)**. Enhancement prospects in the DL algorithms for fault detection and diagnosis.	**(A)**. Does not cover a relation between DL algorithms and data nature. **(B)**. It does not explore the impact of data standardization, which could enhance model interoperability and scalability. **(C)**. The economic perspective of fault diagnosis for both small- and large-scale PV facilities has not been discussed, which is crucial for assessing the cost-effectiveness and financial feasibility of implementing advanced diagnostic techniques.
	[18], 2021	AI & ML	Focus on various methods for utility PV plant fault identification such as electrical parameters, AI, and thermal aging.	**(A)**. A very limited number of AI techniques are considered. **(B)**. Does not propose a framework for data extraction phases. **(C)**. There is a complete lack of information regarding fault pattern analysis, severity assessment, and predictions for critical fault levels.
	[19], 2018	AI	**(A)**. Analysis of PV fault types and causes with emphasis on PV array. **(B)**. Examined electrical methods for PV array, string fault detection, and diagnosis.	**(A)**. Limited analysis of AI algorithms for fault detection and diagnosis. **(B)**. Mostly offline data for fault detection and diagnosis, neglecting insights into online monitoring setups or data.
	[20], 2022	ML	**(A)**. ML algorithms review for performance prediction and fault detection. **(B)**. Compared conventional techniques to analyze PV systems from thermal and electrical perspectives.	**(A)**. System information, data types, and future forecasting of the faults have not been discussed.
	[21], 2021	ML	**(A)**. Reviewed AI algorithms and Internet of Things (IoT) applications for PV Systems. **(B)**. Comparison of AI (ML and DL) algorithms based on cost implementation, accuracy, and real-time hardware feasibility.	**(A)**. Mostly focus on fault classification and detection. Fault prediction algorithms have not been discussed. **(B)**. Lacks in discussing the development of online detection methods for multi-defect/faults.
	[22], 2021	AI	**(A)**. A systematic study on the AI hybridization models for PV fault diagnosis. **(B)**. AI model comparisons are based on the data nature, model structure, and fault diagnosis performance.	**(A)**. This review paper does not consider experimental investigation of the applied AI algorithm for fault diagnosis. **(B)**. The specific equipment of the PV system for the AI algorithm's fault detection and diagnosis is not discussed.
	[23], 2023	AI & ML	**(A)**. A review of diverse fault diagnosis techniques. **(B)**. Fault awareness for solar PV systems	**(A)**. A Limited number of AI and ML techniques are considered. **(B)**. Only focus on fault classification—details on the fault pattern analysis, severity assessment, and predictions for critical fault levels are not addressed. **(C)**. Lack of discussion on PV plant maintenance and fault detection.
	[24], 2023	ML	**(A)**. Reviewed ML algorithms used in PV fault detection. **(B)**. A brief overview of ML and its concepts along with various widely used ML algorithms. **(C)**. The main focus on fault detection accuracy and efficiency.	**(A)**. Fault prediction and monitoring of PV plants are not considered. **(B)**. The data extraction process and data nature have not been addressed. **(C)**. Limited application of ML algorithms for fault detection and classification has been introduced.
	[25], 2022	AI	**(A)**. Emphasized on AI-based research for PV plant fault diagnosis. **(B)**. Highlighted fault types, features, and diagnostic performance of the ANN models.	**(A)**. Limited AI and ML techniques are considered. **(B)**. Discussion on PV plant maintenance and future prediction of the faults is not considered.
	[26], 2024	ML	**(A)**. Reviewed supervised learning-based ML for fault diagnosis in PV systems. **(B)**. The study aims to explore multiple PV system faults and their types	**(A)**. Focused on ML algorithms as part of the fault diagnosis only. However, analysis of data acquisition and online monitoring of PV systems is not discussed.
	[27], 2022	Visual, thermal and electrical	**(A)**. Consideration of most PV potential faults on AC and DC sides. **(B)**. Enumeration of specific PV fault detection and classification. **(C)**. PV fault categorization based on visual, thermal, and electrical methods.	**(A)**. Infusion of monitoring architecture doesn't consider performing PV fault detection and classification.
	[28], 2023	AI and thermography	**(A)**. Various classifications of PV faults and fault detection techniques are presented. **(B)**. Fault localization and classification by thermography methods. **(C)**. Review of different AI tools for fault detection and classification.	**(A)**. Limited AI methods are considered. **(B)**. No information on the improvement of fault characterization and identification. **(C)**. PV fault prediction using small datasets. **(D)**. Does not propose the monitory system.

(*continued on next page*)

Table 1 *(continued)*

Category	Reference	Methods	Features	Limitation(s)
PdM	[29], 2022	ML and physical techniques	**(A)**. Reviewed forecasting techniques for ambient and cell temperature, solar irradiance, and their connection to PdM. **(B)**. Discussed the limited utilization of weather stations in PdM and forecasting of climate parameters. **(C)**. Highlighted the correlation between PV PdM and forecasting.	**(A)**. The paper concentrates on forecasting weather parameters in the PdM process but needs to include fault classification for a better understanding of system behavior and targeted maintenance interventions.
	[30], 2021	DL & AI	**(A)**. Explored various methods such as deep learning, ensemble learning, and transfer learning.	**(A)**. A limited number of AI techniques are considered. **(B)**. System information, data types, and fault diagnosis analysis have not been discussed.
	[31], 2020	Maintenance types	**(A)**. Focused on maintenance strategies to prevent efficiency drops due to faults. **(B)**. Reviewed techniques developed under maintenance strategies. **(C)**. Emphasized fault prediction in PV systems to enable future expansion	**(A)**. Omitting fault indicators and failure mode analysis significantly weakens the paper's insights into system health and its optimization potential.
	[32], 2022	DL	**(A)**. Remote sensing, problem detection, and diagnosis of PV systems **(B)**. Applications of DL and IoT for PV systems.	**(A)**. The focus is solely on the type of maintenance. **(B)**. There is no description of the data type or any future forecasts regarding faults.
	[33], 2022	Maintenance approaches	**(A)**. Highlights the present scenario of maintenance approaches with possible causes of degradation. **(B)**. Current approaches and opportunities for PV PdM are provided.	**(A)**. Discusses PdM from environmental issues and perspectives only. **(B)**. Operational perspectives (technical performance optimization, panel efficiency, inverter performance, and overall system health), and technological perspectives such as IoT sensors, data analytics, and ML are essential to be considered.
This Work	–	AI & ML	**(A)**. Discussion on PdM and future forecasting of the faults. **(B)**. ML-based PdM and fault diagnosis for PV Systems. **(C)**. PdM potential framework for PV Systems. **(D)**. PV data sources and analysis. **(E)**. Experimental validation of ML algorithms for PdM and fault diagnosis. **(F)**. Discussion about description of data and data types applied for PdM and fault diagnosis. **(G)**. Exploration of PdM/ anomaly detection concerning specific classes and future forecasting of faults. **(H)**. Challenges of data standardization and its effects on AI performance for PdM and fault diagnosis.	-

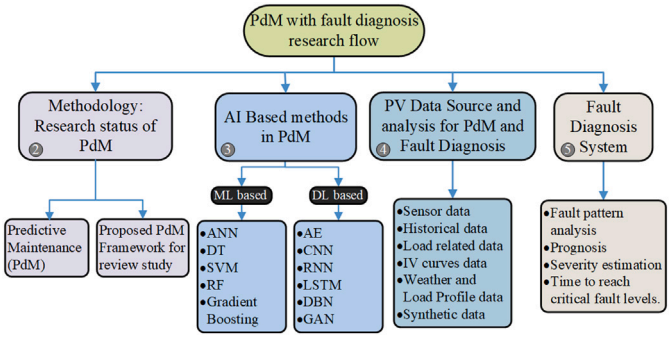


Fig. 1. Flow diagram of this review work.

of these disruptions can be prevented or mitigated through appropriate maintenance strategies. The evaluation of PdM in relation to different maintenance strategies, focusing on costs and methods, is illustrated in Fig. 2, along with a financial comparison in Fig. 3. Table 2 outlines the advantages and disadvantages of each maintenance method as it pertains to the PV system.

2.2. Proposed PdM framework for PV system

An ML-based PdM framework for PV systems has been proposed in this study, illustrated in Fig. 4, which provides a structural approach to the deployment of AI-based PdM integrated with fault diagnosis (baseline for our review study).

2.2.1. Data acquisition

The process begins with data acquisition, a critical step for implementing effective PdM and fault diagnosis in a PV system. This stage involves the systematic collection of comprehensive data. Real-time electrical parameters such as voltage, current, and power are captured using voltage sensors (including potential transformers and dividers) and current sensors (like current transformers and Hall effect sensors), delivering immediate and accurate assessments of system performance. IR cameras effectively capture thermal images, pinpointing potential hotspots that indicate faults based on module temperature distribution. Environmental conditions are rigorously monitored through a variety of advanced instruments, including humidity sensors (capacitive and resistive), rain gauges (tipping bucket and weighing models), soiling sensors (optical and ultrasonic), sky imagers/cloud cover sensors, and pyranometers/reference cells that measure solar irradiance. Moreover, the tilt and orientation of the PV module and array are precisely tracked using inclinometers/tilt sensors alongside GPS and compass systems.

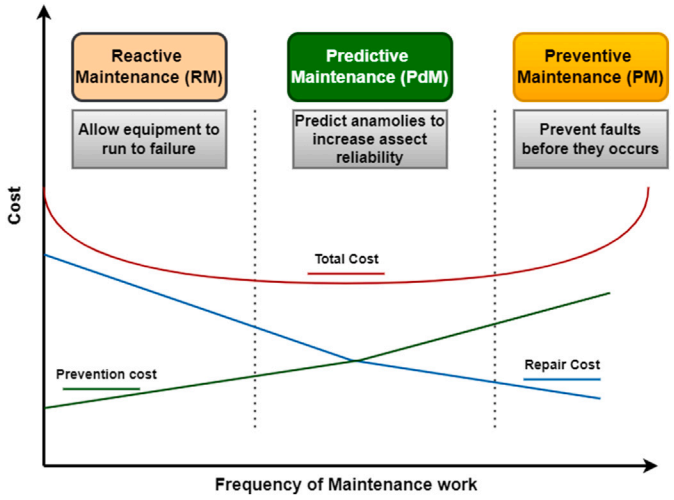


Fig. 2. Comparison of maintenance strategies based on costs and maintenance activities.

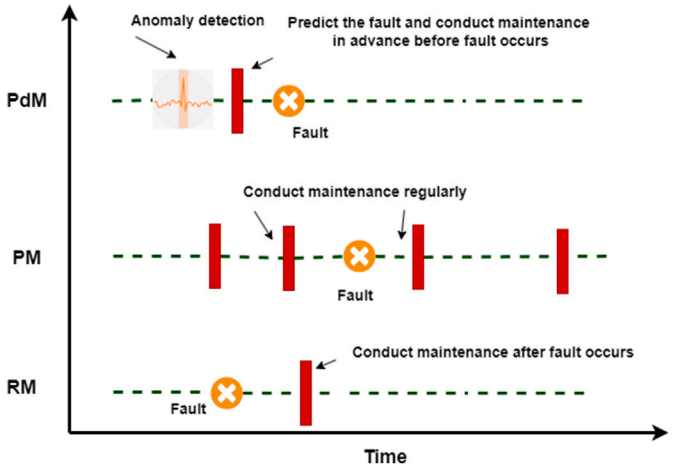


Fig. 3. Maintenance difference between PdM, RM and PM.

Additionally, simulated performance data generated from PV system simulation software, as well as calculated reference and total yields and power flow information sourced from these sensors, form integral components of the process. This multifaceted strategy, which harnesses a diverse array of sensor technologies and cutting-edge data analysis methods, guarantees a comprehensive understanding of the PV system's health. This approach enables precise fault detection and empowers the development of robust predictive maintenance strategies.

The complexity of data acquisition in PV systems is significantly increased by the diverse outputs and collection frequencies of sensors. Each sensor type, whether high-frequency current and voltage sensors or slower irradiance and temperature sensors, produces data at varying rates and exhibits distinct signal characteristics. This demands a robust Data Acquisition System (DAS) that can effectively manage these heterogeneous sensor inputs. Modern DAS solutions must employ modular architectures to ensure seamless integration of different sensor types and provide tailored signal conditioning to meet specific sensor requirements [36]. Moreover, efficient data transmission is crucial. Communication protocols such as Modbus, TCP/IP, and various wireless technologies (including Wi-Fi and cellular) must be utilized to facilitate reliable data transfer to central monitoring stations or cloud platforms. The substantial volume and variability of data introduce significant challenges regarding standardization. It is imperative to ensure data consistency and interoperability among different sensor types and DAS components for effective data analysis and fault diagnosis. Implementing standardized data formats and communication protocols is vital to overcoming these challenges, enabling seamless integration and efficient processing of the diverse data streams generated by the PV system.

2.2.2. Data processing

After acquiring data, the next vital step is data processing, which involves three main phases: data cleaning, data integration, and data transformation. Due to the diverse nature of PV system data, each phase should be customized according to the specific attributes of the data types. Firstly data cleaning involves cleaning raw sensor data to eliminate inaccuracies, noise, and incomplete readings, ensuring precision in analysis. To manage missing values in electrical parameters due to brief interruptions, techniques like linear interpolation or Kalman filtering can be employed. In the case of thermal infrared images, using noise reduction methods like median filtering can effectively reduce sensor noise. Weather data, which is often impacted by sensor drift or failures, requires outlier detection and imputation methods to maintain accuracy. Secondly, data integration involves merging information from various sources. This process may include combining real-time sensor data with data from weather APIs, retrieving historical performance data from databases, and converting simulation results into a cohesive format. For example, irradiance data obtained from pyranometers can be integrated with cloud cover data from sky imagers and forecasted irradiance from weather APIs. This phase typically demands careful synchronization of timestamps and geospatial data. Third, data transformation is crucial for preparing data for analysis and modeling. Techniques such as PCA can be employed to reduce the dimensionality of high-dimensional

Table 2
Maintenance techniques within a PV system framework.

Technique	Features	Limitation	Summary
PM	This maintenance helps in the minimization of sudden failures and enhances the component lifespan.	**(A).** Cost is ineffective due to scheduled downtime. **(B).** It can cause unnecessary maintenance whereas the machinery is still in use.	**(A).** This technique is applicable where fault impact is acute and is suitable for machinery prone to wear and tear. **(B).** Incompetent, as components rarely fail before their expected life cycle. **(C).** This technique can be suitable for components such as inverters, and PV arrays.
RM	**(A).** Lower upfront cost and minimal operational interruption until any malfunction occurs.	**(A).** More unexpected failure risks. Expensive downtown through failures.	**(A).** It is suitable when the impact is not severe and on those components that have low failure chances.
PdM	**(A).** Enhances component overall safety and reliability. It also decreases the overall maintenance costs, maximizes component lifespan, and minimizes the downtown.	**(A).** Investment in the monitoring system and experts is required for the data analysis.	**(A).** Maintenance decisions are made solely on the data analysis by ML. **(B).** For a system where the early detection of potential failures is crucial, the impact is significant.

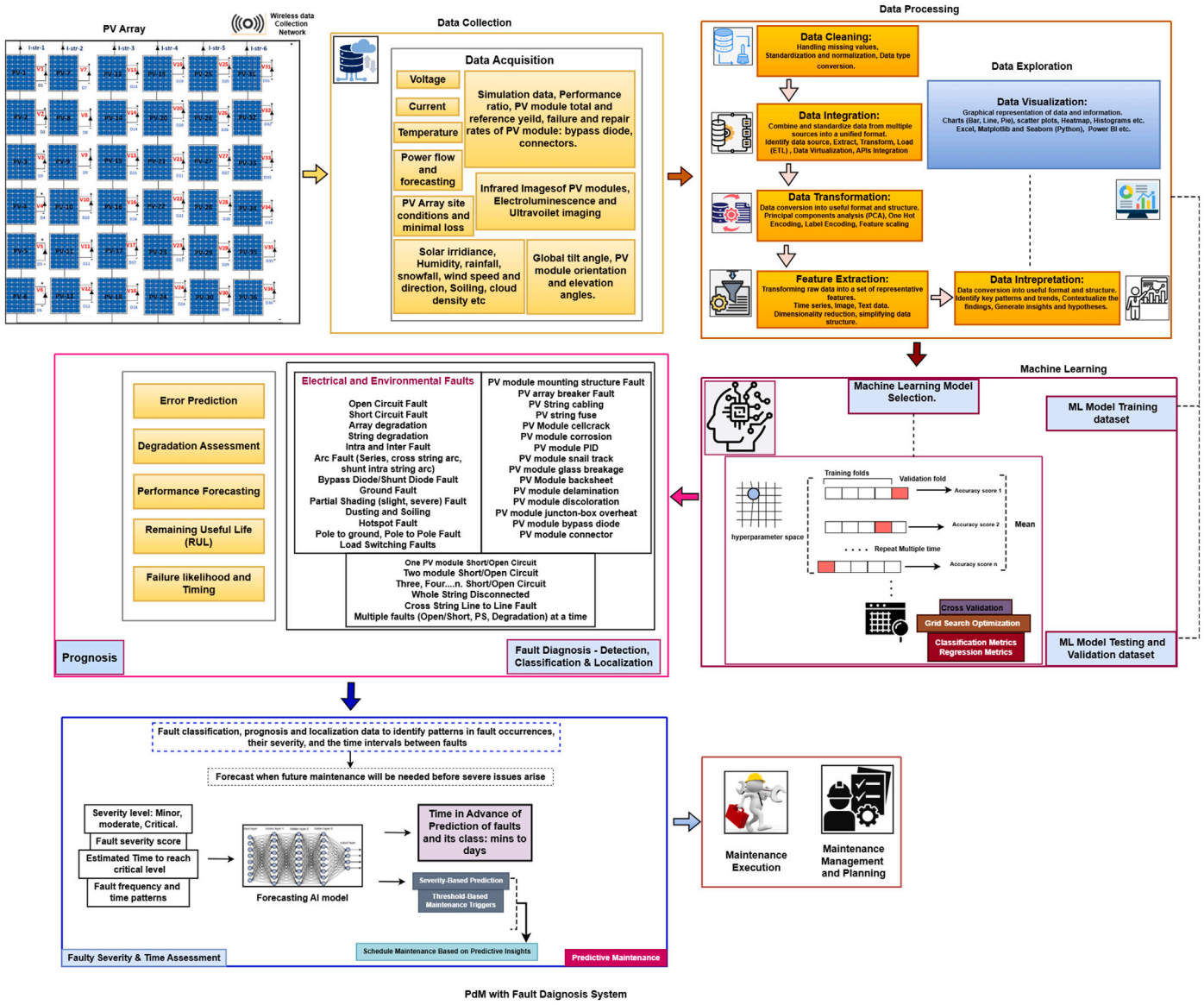


Fig. 4. Proposed ML integrated PdM structure with fault diagnosis.

datasets, like thermal IR images or weather data, while still preserving key information. Likewise, label encoding can categorize different types of faults identified from thermal anomalies or electrical fluctuations. Additionally, feature scaling methods, including standardization or robust scaling, help standardize features to a similar scale, preventing any single feature from overshadowing others in the analysis. For time-series data, such as voltage and current readings, techniques like Fast Fourier Transform (FFT) or wavelet transforms can be applied to uncover frequency-domain characteristics that signal faults. This comprehensive data processing pipeline effectively transforms raw data into a structured and informative dataset, which is prepared for fault diagnosis and PdM modeling.

After completion of the thorough data processing phase, feature engineering is performed to derive significant and distinguishing information from the cleaned data, focusing on the specific goal of PdM and fault diagnosis. Due to the variety of dataset formats, feature engineering approaches are adjusted to fit these differences. For example, features derived from I–V curve analysis are essential for a precise assessment of PV module performance degradation. Key metrics including MPP voltage and current, fill factor, and series/shunt resistance, are calculated directly from measured I–V curves. Any deviations from expected I–V curve characteristics serve as strong indications of module faults, such as shading, soiling, or cell degradation. Image-based feature extraction using thermal IR imagery involves techniques such as texture analysis and histogram analysis to detect spatial thermal anomalies that signal faults. These techniques help identify hotspots, temperature gradients, and other thermal patterns, offering vital insights into the thermal distribution of the module. Additionally, time-series feature extraction from voltage, current, and power data employs statistical moments and frequency-domain transforms (such as FFT and wavelet) to uncover temporal patterns and frequency components linked to both normal and abnormal operations. After feature extraction, the framework includes a data visualization and interpretation phase. This phase aims to convert the extracted features into easily understandable formats, such as graphs, charts, and heatmaps, which help in identifying crucial patterns and insights regarding the health of the PV system. These visual aids are vital for experts to understand the system's status and to make informed decisions regarding maintenance and troubleshooting issues.

Table 3
PdM framework design requirements for PV system components based on ML (mapping data types to ML algorithms).

For	Data	Suitable ML models	Reflection on mapping and application significance
PV arrays	**(A) Sensor Data like**: **(I)**. PV DC Current, DC-DC converter current, PV DC voltage, PV DC power. **(II)**. Solar irradiance, humidity, wind speed/direction, total radiation received by array surface, ambient light, and temperature. **(III)**. Dust and soil accumulation. **(IV)**. Azimuth and tilt angle. **(B). Generation/weather profiles** **(I)**. Historic data **(II)**. Synthetic data.	**(A) Classification algorithms:** **(I)**. Hybrid ML models with autoencoders. **(II)**. Long short-term memory (LSTM) models. **(III)**. Time series analysis **(IV)** Regression.	**(A)**. Developing a correlation model between weather parameters and PV power generation for regenerating PV system dynamical behavior, i.e., array voltage, current, temperature and solar irradiation to reduce the installation and maintenance cost in terms of sensors and real-time insights into PV system components such as PV arrays. **(B)**. Temporal dependencies and patterns can be captured by LSTM. **(C)**. To identify and control the deviation and unusual response of PV system components, classification models can be used. **(D)**. Power generation patterns of the PV system can be identified by time series analysis such as the Autoregressive Integrated Moving Average (ARIMA) model and can predict the future generation trends over time. **(E)**. The potential energy generated from PV systems can be predicted using regression models incorporating environmental and historical data. **(F)**. Fault severity progression over time can be modeled using LSTM, which captures degradation and aging effectively. **(G)**. A multi-task learning framework can be explored to improve predictions by enabling LSTM to forecast severity progression while allowing boosting models to effectively classify or score fault severity levels.
Inverter	**(A)**. Power flow measurements. **(B)**. Real-time operational data such as energy yield, and total operating time. **(C)**. Historical data such as aggregated energy data on day, week, month, and year.	**(A)**. Classification algorithms **(B)**. Regression **(C)**. Hybrid ML models with autoencoders **(D)**. LSTM **(E)**. Time series analysis	**(A)**. Inverter conditions can be tracked through both real-time operational data and historical information, while classification models assist in detecting anomalies and irregular behavior. **(B)**. Real-time insights into inverter health can be modeled and monitored.
AC grid and transmission lines	**(A)**. Instrument transformers such as Current Transformers (CT) and Voltage transformers (VT) for AC current and voltage, respectively. **(B.** Power meter for measuring active power (kW), reactive power (kVAR), and apparent power (kVA). **(C)**. Frequency meter for monitoring the frequency of AC grid. **(D)**. Load profiles **(E)**. Maintenance records	**(A)**. Classification algorithms **(B)**. Regression **(C)**. Hybrid ML models **(D)**. LSTM **(E)**. Time series analysis	**(A)**. Regression models predict remaining life or estimate potential faults in transmission lines using historical and real-time data from CTs and VTs. **(B)**. Time-stamped data from CTs, VTs, and power meters can be analyzed to detect changes and forecast line behavior using time series analysis and LSTM for anomaly detection by identifying deviations from learned patterns.

2.2.3. Implementation of AI

The next stage in the PdM framework covers the selection, implementation, training and validation of AI algorithms. The selection of AI techniques relies on the nature of the data and the application under consideration. Table 4 demonstrates the overview of designing a PdM framework for PV systems based on AI models and maps the potential AI applications to data types (providing guidelines to researchers). The AI algorithm is generally trained to recognize specific patterns in data that indicate the system's deviation from the normal condition. AI algorithms identify outliers by recognizing the features important for differentiation. To improve the accuracy and reliability, a validation step is performed, where the AI model is tested on the new set of data that has not been used for model training. In this way, the model will be generalized to new data and avoid over-fitting. After deploying the AI model and identifying anomalies, predictive alerts can be implemented.

2.2.4. PdM with fault diagnosis model

Following the AI training, the process advances to prognosis and fault classification. Prognosis entails predicting errors, assessing system degradation, forecasting performance, estimating the remaining useful life, and determining the likelihood and timing of failures. Subsequently, fault classification organizes identified problems into categories, which may include electrical or environmental faults, faults related to PV modules, or issues with PV module connections. This is followed by fault localization, which identifies the precise location of faults to enable focused maintenance efforts. In the realm of PdM, the integration of fault classification and localization helps to discern patterns in the occurrence of faults, their severity, and the intervals between them.

The proposed framework enhances traditional PdM methods by adopting a more holistic approach to evaluating fault severity, conducting temporal fault analysis, and classifying predictive faults. A key feature is the ability to quantify fault severity, which allows for a deeper understanding of system degradation beyond simple fault detection. For example, in PV systems, PS is examined based on its origin and effects. The analysis differentiates between shading caused by factors such as tree foliage, dust build-up, bird droppings, or shadows from utility poles, each of which presents unique shading patterns and severity levels. By developing metrics that account for the shaded area, the intensity of shading (derived by irradiance measurements), and the corresponding power loss, a detailed assessment of shading severity can be achieved, ranging from minor impacts on energy output to significant performance decline. Similar to other faults such as OC and SC, their fault levels are defined based on the number of affected PV modules within strings or arrays, delivering a clear assessment of the fault's severity. Incorporating fault duration into the analysis is crucial for accurate degradation modeling. By combining fault severity with aging data, it is possible to calculate degradation levels over time. This temporal analysis allows for predicting the RUL of a system and optimizing

maintenance schedules. By monitoring the progression of fault severity over time, dynamic thresholds can be established to prompt maintenance interventions before the critical breakdown of the PV system occurs.

Furthermore, the framework focuses on recognizing fault patterns. The analysis of how fault severity progresses from initial detection (level 1) to critical failure highlights distinct patterns that precede catastrophic breakdowns. These patterns are utilized to create predictive fault classification models. Specifically, after detecting an anomaly, the framework observes the anomaly's pattern to determine its classification. By keeping track of when faults occur and how quickly their severity increases, it estimates the time to critical breakdown. Dynamic thresholds, which are adjusted according to the temporal changes in severity, provide real-time insights into the health of the system. This approach allows for the forecasting of faults and their classifications days to weeks in advance, from the initial anomaly detection to the expected critical breakdown. This capability supports proactive maintenance, reducing downtime and enhancing system reliability by allowing maintenance teams to concentrate on specific fault cases and facilitating prompt responses. The forecasting AI model incorporates various inputs, including the severity score, fault frequency, temporal patterns, estimated time to critical failure, severity level, and aging data. This model anticipates upcoming maintenance tasks and produces optimized maintenance schedules that are informed by severity-based predictions and dynamically modified thresholds.

The proposed framework presents an innovative fusion of AI-driven fault diagnosis and prognosis with PdM scheduling, addressing a gap in the current PV fault diagnosis and maintenance system. Furthermore, utilizing AI algorithms empowers the system to learn from both historical and real-time data, enhancing its accuracy in predicting future maintenance needs. Overall, the suggested approach will significantly improve the effectiveness of PdM and fault diagnosis systems in PV installations, providing a robust answer to issues related to data inaccuracies, environmental changes, and communication lags.

This proposed framework acts as the basis for the following review of related literature. Table 4 in Section 3 centers on PdM, anomaly detection, and CM. It also investigates the use of various datasets for these tasks and underscores the capability to forecast faults in advance, often days or weeks prior to their occurrence. Table 6 in the fault diagnosis, Section 6, evaluates papers based on fault severity, fault occurrence time, and the ability to predict future faults, further progressing the conversation on fault diagnosis and prognosis with PdM.

3. AI based methods in PdM

In this article, we reviewed various AI algorithms used in PV predictive maintenance and fault diagnosis. Table 4 reports state-of-the-art research employing ML/DL algorithms for PV PdM, condition monitoring, and anomaly detection. The reviewed papers are compared based on the category and type of algorithm used, equipment of the PV system, description of the data applied for PdM, real-time implementation/verification of ML/DL algorithm, data type, PdM/anomaly detection/condition monitoring as well as specific fault types considered. The important columns of Table 4 are briefly described as:

- **Category**: The first column categorizes the reviewed papers based on the primary methodological approaches employed for PdM, including DL, ANN, Supervised and Unsupervised ML and other relevant techniques. It combines and organizes the research papers reviewed into groups based on their similarities in methodological approach.
- **ML methods**: The ML algorithms utilized for PdM and anomaly detection are specified in this column. These may range from supervised learning techniques like classification and regression to unsupervised learning approaches such as clustering. The choice of ML algorithms profoundly influences the accuracy, efficiency, and interpretability of the PdM framework. These algorithms are later explained briefly.
- **Equipment**: The specific PV equipment or systems used are described in this column. This includes details such as the types of solar panels, inverters, monitoring devices, or any other hardware components relevant to the study.
- **Description of the data used for PdM**: This column outlines the nature and characteristics of the data used in the existing studies. This encompasses parameters such as sensor readings, environmental conditions, performance metrics, or any other relevant data sources collected from the PV equipment. A clear description of the data helps readers understand the input variables and the information available for training and evaluation of the machine learning models.
- **Real-time implementation/validation**: In this column, we examine whether the ML-based predictive maintenance framework has been implemented for real-time operation on hardware. This indicates the practical applicability and scalability of the proposed solution in monitoring and managing PV systems in real-world settings. Real-time implementation enables timely detection and mitigation of faults or anomalies, thereby enhancing the reliability and performance of solar energy systems.
- **Data type/source**: This column describes the data types used for PdM and anomaly detection in the photovoltaic system. The focus is on the source and nature of the data employed in the study. Data types may vary widely, ranging from simulated data generated by software tools like MATLAB or LabVIEW to real-world data collected from operational PV plants. Additionally, the column includes information on the format and structure of the data, such as numerical values, time-series data, images, or text, which influence the selection of appropriate data preprocessing techniques and AI algorithms.
- **PdM, anomaly detection and condition monitoring**: This column assesses whether studies under review apply PdM, anomaly detection, or CM, and pinpoints the particular factors used, including system variables, fault indicators, or environmental conditions, for conducting these tasks. Anomaly detection involves identifying deviations from normal behavior or expected patterns in the data, which could indicate potential faults or abnormalities in the PV system. CM entails continuous observation of various parameters and performance metrics to evaluate the health and operational status of PV equipment. Integrating anomaly detection and condition monitoring towards the PdM framework provides a comprehensive approach to identifying and addressing issues affecting the reliability and performance of the PV system.
- **Specific class and future fault forecasting**: Most papers focus solely on detecting anomalies, without categorizing specific fault types or providing detailed explanations for the underlying causes of anomalies in offline analysis. Therefore, *"Yes"* indicates characterization of anomalies into specific faults, and *"No"* means generalization without categorization of fault types. Furthermore, future fault forecasting capability (as *"Yes"* or *"No"*) is investigated, which refers to the ability to predict faults in advance (days or weeks before they occur). This allows for necessary maintenance actions that can prevent downtime and extend the equipment's life.

Furthermore, Fig. 5 summarizes major AI (ML and DL) algorithms and their applications, and the most used AI methods are briefly explained in the subsequent sections. These include supervised learning models with the most common Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), K-nearest neighbors (KNN), ensemble techniques, gradient boosting algorithm, Fuzzy Expert System (FES), and unsupervised learning with Principal Components Analysis (PCA), Hierarchical clustering, K-Means, Fuzzy C-Means and semi-supervised learning with graphic and generative based models. Likewise, in deep learning, Convolution Neural Network (CNN), Long Short-Term Memory (LSTM), Bi-LSTM, Recurrent Neural Network (RNN), and Generative Adversarial Network (GAN) are discussed.

Table 4
A summary of the most recent published work on PV system PdM.

Category	Ref	ML methods	Equipment	Description of data used for PdM	Real-time implementation/validation	Data type/source	PdM, Anomaly detection, or Condition Monitoring	Specific fault class & future forecasting
DL	[37]	CNN	PV panels	Daily electrical power signal	No	**Real Data**: Measured values of targeting panels and two neighboring panels	PdM: predicting the regular pattern of shadows	No
	[38]	LSTM	PV arrays	Timestamp, rainfall, solar module output power, temperature, diffuse radiation	No	**Real Data**: Desert Knowledge Australia Solar Center (DKASC), Australia.	PdM: Prediction of solar module output power	No
	[39]	K-Means clustering & LSTM—anomaly detection	String modules of large-scale photovoltaic plant	Electrical current of string modules, PV module temperature, module plane global irradiance	No	**Real Data**: Obtained from inverters including output current of string modules for seven combiner boxes, total of 420 modules electrical current and 8400 PV modules.	Anomaly detection: Predicted string modules output current	No
	[40]	LSTM networks and auto-encoder	PV Arrays	Data include battery array output current, solar panel temperature, drive shaft temperature, and satellite telemetry parameters.	No	**Real Data**: captured from small satellite Solar Array Drive Assembly (SADA).	Fault prediction: Early fault signals identification	No & Yes: Anticipating faults occurrence ranging days in advance.
	[41]	CNN—isolated learning and transfer learning	PV modules	Infrared images of normal and mechanical loads and shading (artificial) for artificial defect induction are taken and processed for various PV modules	No	**Real Data**: Indoor and outdoor thermography setup with IR camera, PV panel and variable power supply	Anomaly detection: Normal and faulty behaviors	No
	[42]	Intelligent Monitoring System (IMS) with LSTM to predict PV output power, NB, KNN, and SVM to detect and classify fault events.	PV modules, PV arrays as a stand-alone PV system	Power forecasting data: irradiance, temperature, humidity, power, voltage, and current. Feature extraction data: Temperature, irradiance, and power.	No	**Real data**: from PV experimental setup: 3 features, 20 time steps, and 7300 samples.	Monitoring and fault classification system	Yes: Array and string degradation, open circuit faults & No.
	[43]	IoT platform architecture: isolation forest for anomaly detection, BiLSTM-multi dense, convolutional LSTM. bidirectional LSTM, LSTM, and RNN for PV power prediction and monitoring with an edge server	Power Plants of Seoul	Temperature, Humidity, Wind speed in x and y direction, irradiation intensity Cloud density, Sunshine duration, Generated PV power, Model structure, hyperparameter	No	**Real data**: PV power generation data from 2021 to 2023, hourly weather observation data with a 24-hour data length.	Early anomaly detection: normal and abnormal data	No
	[44]	CNN	Large-scale photovoltaic systems.	Real-time images: thermography, Electroluminescence (EL) and ultraviolet fluorescence imaging, image binarization.	No	**Real Data**	PdM: Generic	No
	[45]	ResNet-34 CNN with a supervised contrastive loss and KNN classifier	105,546 PV modules from six PV plants	4.16 million IR images acquired under clear sky conditions and solar irradiance above 700 W/m^2	No	**Real Data**: Images from IR videos of PV module captured by drone-mounted DJI Zenmuse XT2 camera. 39.4 images providing multiple augmented views.	Anomaly detection: Normal and anomalous	No

	[46]	AutoEncoder LSTM, facebook-prophet, and isolation forest	Two solar power plants.	Weather measurements: irradiation, ambient, and module temperatures. Generate rate: AC and DC powers, daily and total yield	No	**Real Data**: 34 days of data from plants in India with 15 min intervals.	Anomaly detection: Normal healthy condition and abnormal behavior	No
	[47]	A^2-LSTM	PV Array -Simulink model and Greece Power plant	Data include weather parameters: ambient and module temperature, wind speed, solar irradiance. Inverter DC power, PV arrays current, voltage and power	No	**Simulated Data**: dataset spans 480 days with 10 min intervals. **Real Data**: 365 days, with 20 min intervals.	Anomaly detection: Normal healthy condition and abnormal glitch in PV production	No
	[48]	CNN-LSTM	PV System	Irradiance on PV surface Ambient temperature, PV output current, voltage, and power generation	Yes	Real Data	Generic—Online monitoring	No
ANN	[49]	ANN	PV system	Predicted and actual/measured values of AC power production	No	**Real and simulated Data**: historical store data of solar irradiance and temperature. ANN for model power production	PdM: Predictive maintenance alerts for avoiding possible incoming faults	No & Yes
	[50]	Fully connected neural network (FCNN)	PV module	PV modules string current and voltage, irradiance and air temperature	No	**Simulated Data**: Single PV module connected directly to an adjustable resistive load and a current and voltage data acquisition system	Fault detection	No
	[51]	Residual NN with infrared radiation (IR) cameras.	PV modules	IR images of different solar plants captured by UAV. Images classified as classes: No anomaly, anomaly classes: cracking, shadowing, hotspot, and hotspot multi, cell and cell-multi, diode and diode-multi, soiling, vegetation, and offline module.	No	**Real Data**: A real solar plant with 20,000 IR images, and 12 different solar defects.	Anomlay detection: Prediction and classification of anomaly solar modules	Yes: Classify 12 anomaly types & No.

(continued on next page)

Table 4 *(continued)*

Category	Ref	ML methods	Equipment	Description of data used for PdM	Real-time implementation/validation	Data type/source	PdM, Anomaly detection, or Condition Monitoring	Specific fault class & future forecasting
	[52]	Pattern recognition NN and unsupervised clustering techniques	PV plant	Plant nominal power (MW) and active power (KW)	No	**Real Data**: Six PV plants 10 MW each, more than 100 inverters of three different technology brands.	Generic fault/status prediction up to 7 days	Yes—specific class prediction. & Yes—Predicting fault types and severity in the future (ranging from a few hours to seven days).
	[53]	Regression NN	Large-scale and remote PV farms	Clean and soiled modules data, solar irradiance, module temperature, and hourly maximum power measurements.	No	**Real Data**: with 3308 observations.	Real-time solar monitoring, power prediction, and anomaly detection.	No.
Supervised ML	[54]	Regression and SVM	small-scale PV generation system	Solar irradiance, cell and ambient temperature, DC current and voltage, power.	No	**Real Data**: acquired from the power conversion system	Abnormal condition detection	No
	[55]	XGBoost regression model	Test-bench PV system	Meteorological data: Wind speed and direction, ambient temperature, and in-plane irradiance. PV module data: Back surface module temperature, DC current, voltage and power, module orientation and elevation angles, array energy, PV system total and reference yield, performance ratio, AC output power.	No	**Actual (Real)**: Historical and real-time performance data and **synthetic (simulated)** data using weather parameters, historical labeled-data with field emulated failure.	Health-state architecture for PV system monitoring	No
	[56]	Ensemble ML Empirical Analysis: RF, XGBoost, CatBoost, and Light Gradient Boosting Machine (LightGBM)	PV system	DC voltage and current, instantaneous power generation, power factor, frequency, timestamp, line-line voltages, phase currents, PV output power, maximum and daily power generation.	No	**Real data**: from 99.9 kW PV system	PdM: Anticipating the maintenance schedules.	No.
	[57]	RF regression, Vector Auto-regression (VER) for power generation forecasting	1 MW solar power plant.	Five PV modules including poly-crystalline, thin-film amorphous silicon, and concentrate PV. Dataset: Timestamp, power generation (KWh), aggregate meter reading (kWh), seed data (kWh), insolation, PR (%), etc.	No	**Real Data**: Dataset generated (2012–2020) with 12 structured and 1 instructed column for maintenance notes, daily weather, and problem observations.	Solar power forecasting using maintenance activities.	No.

	[58]	Statistical and ML models	Six sub-plants with 22 inverters each	131 measurement variables sampled every minute by each inverter. Measurement variables: current, voltage, and temperature. Six weather stations per sub-plant.	No	**Real Data**: Sensor and weather data to form a multivariate time series with 92 variables. Each inverter is generated (June 2019–2022) in one row per 30 min	PdM: Optimizing maintenance decisions	No.
	[59]	NB, one vs. rest and MLP NN algorithm	Grid-connected PV system	V_{dc}, I_{pv}, V_{pv}, I_b, V_{abc}	No	**Simulated data**: IEEE dataport	Anomaly detection: normal and abnormal behavior	N/A
	[60]	One-class Support Vector Machine (1SVM)	Grid-connected PV system.	Meteorological variables: $T_{amb}, G_{i,c}, G_{i,p}, G_{h,p}$ and electrical variables: $V_{DC}, I_{DC}, V_{AC}, I_{AC}$.	No	**Real Data:** Sampling time is 1 min with 1440 samples per day. LabView software is used to perform the monitoring process. **Simulated data:** Simulated data using co-simulation between PSIM and Matlab.	Anomaly detection: Normal and abnormal features	No
	[61]	O&M Decision Support System (DSS), Failure Detection Algorithms, Trend-based Loss Routines and Maintenance Strategies Routines (MSRs).	Large-scale 1.8 MW PV system. 7824 PV modules in 326 parallel strings with 24 modules each, and 4 inverters.	Weather data: In-plane irradiance, module back-surface and inverter temperature, wind speed, and direction. Snowfall and rainfall measurements. Electrical data: Inverter DC voltage, power, AC output power, and array/system performance ratios.	Yes	**Real Data**: PV power plant historical data over 6 years in Larissa, Greece. Datasets include 15 measurements averaged on 15 min for four grid-connected inverters covering 2013 to 2018.	MSRs: Corrective actions	Detecting PV's underperformance issues. & No.
	[62]	Semiparametric framework and block bootstrap method	PV system with different array technologies	AC power output data, global horizontal and diffuse horizontal irradiance, Relative Humidity (RH), and ambient temperature	No	**Real Data**: obtained from the DKASC, Australia	Generic remaining Useful life (RUL)	No.
Unsupervised ML	[63]	Monte Carlo Based Pre-Preprocessing and PCA based anomaly detection	Solar cell production plant	Ozone concentration level, station temperature, flow, and pump speed	No	**Real Data**: obtained from Enel Green Power's 3SUN solar cell production plant, in Italy	Anomaly detection: handle outliers and intrinsically deals with outlier substitution considering temporal locality of subsequent samples.	No & Yes—predict equipment's faults ranging almost two weeks.
	[64]	Tree ensemble algorithm based on XGBoost hybrid with unsupervised method	PV Panels	Historical meteorological and PV power datasets	No	**Real data**: from 6.95 MW PV plant	Instantaneous performance monitoring	Yes—Four condition based periods & No

(continued on next page)

Table 4 *(continued)*

Category	Ref	ML methods	Equipment	Description of data used for PdM	Real-time implementation/validation	Data type/source	PdM, Anomaly detection, or Condition Monitoring	Specific fault class & future forecasting
	[65]	Cluster-Based Local Outlier Factor (LOF), simple LOF, KNN, Multi-layer Perceptron (MLP)	Decentralized PV system	DC current, voltage and power of monitored string, global horizontal irradiance, plane of array irradiance and air temperature	No	**Real data**	Anomaly detection: Certain faulty scenarios	No.
	[66]	Reliability block diagram and PCA	Inverters of Grid-connected PV system	Failure and repair rates of PV system components: PV modules, converter, bypass diodes, connectors, DC and AC switch, AC and differential circuit breaker and connector.	No	Data ranges from 100 kW to 2500 kW.	CM: health status or useful life. Avoiding sudden breakdowns and unexpected maintenance.	No.
	[67]	Self-organizing map and KPI.	Three PV plants up to 10 MW installed capacity	Electrical (AC/DC currents, voltages, powers), and environmental temperature (internal inverter, panel, ambient), global tilt, and horizontal irradiance).	Yes	**Real Data**: from six PV plants and more than one hundred inverter modules	Online monitoring of anomalies	No & Yes—Anticipating faults up to seven days
	[68]	Closed O&M based on decision support system	PV grid-connected plant: 10.09 MW, 38,344 polycrystalline silicon PV modules, 10 inverters and 68 string boxes.	Global horizontal and diffuse irradiance, module temperature. DC current, voltage and power, AC output power.	No	**Real Data**: Measurement time: 1-year period Jan. 2021 to Dec. 2021. Historical data from a test PV power plant in the Mediterranean region.	O&M: detecting faults at early stages	No
Software Simulation	[16]	PVSyst: PV software	PV power plant	PV Modules: SC current, I_{mpp}, OC Voltage, V_mpp, Module efficiency. Meteorological data: Meteonorm, NASA, and plant site weather data. Simulation data (no losses and with default losses), plant site conditions, and minimum losses.	No	**Real Data**: obtained from 18 MW PV solar plant, Pakistan	O&M—PdM: Planned outages, internal tripping, and external outages	PV plant internal shutdowns & No.

Remarks: In the realm of PdM, "anomaly detection and CM" goes beyond simply evaluating the equipment's current status. Organizations and researchers aim to gain more profound insights into the equipment condition aiming to identify the underlying causes of problems and predict the remaining operational lifespan. This paper explores methods centered on categorizing unusual behavior in PV system equipment. This table provides valuable insights into PdM practices and focuses on anomaly categories and their future predictions.

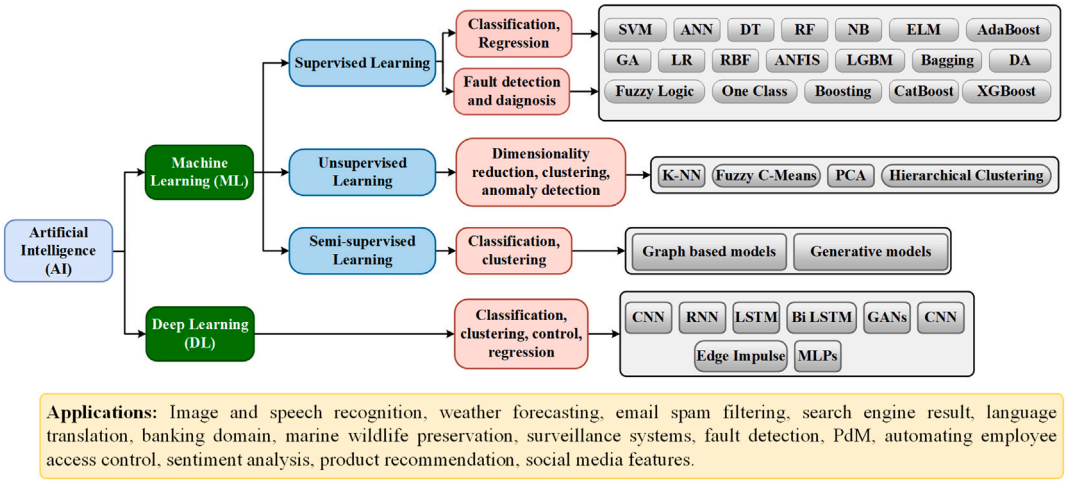


Fig. 5. Major AI (ML and DL) algorithms and applications.

3.1. In-depth analysis and essential insights on PdM

Articles in Table 4 feature a diverse range of AI- and ML-based algorithms for PdM. This research critically evaluates the current landscape of PdM methods in PV plants, demonstrating the necessity to progress beyond the detection of anomalies and condition monitoring to accurate future fault prediction. The key purpose of Table 4 is to highlight the connection between anomaly detection, CM and PdM with specific fault classes and to forecast future prognostics.

- **Anomaly Detection focused papers**: Authors in [39] utilized K-Means clustering and LSTM with data such as solar module electrical current, temperature, and plane global irradiance obtained from the inverter including string modules for seven combiner boxes, a total of 8400 PV modules to detect anomalies in the large PV facility. The method successfully identified anomalies in electrical current at the string level, clearly indicating potential issues. However, it fails to explain the deviation or to specify the actual fault, the severity, or the timing of the occurrence. Similarly, a develop-model transfer DL algorithm has been implemented in [41] to automatically detect PV defective modules based on infrared images. However, this work exclusively identifies the differences between normal and defective modules, without engaging in discussions about the specific types of defects present in the PV modules based on the infrared images. An intelligent IoT platform in [43] monitors multiple PV plants using prediction algorithms such as Bi, convolutional, bidirectional LSTM, and RNN for the next day's power generation and isolation-based forest for anomaly detection. The platform employs a threshold-based anomaly detection method, and every PV plant's humidity sensors identify certain outliers as anomalies. This platform utilized real PV power generation data from 2021 to 2023, with a 24-hour data length. Authors in [46] explored various approaches for anomaly detection in PV systems, including autoencoders, isolation forests, Facebook forest, and AE-LSTM, focusing on identifying healthy versus anomalous system behavior—primarily focusing on anomaly detection. Likewise, [45] applied ResNet-34 CNN to classify images of PV modules as normal or anomalous, while [51] classified anomaly types based on infrared images. However, the study has largely overlooked the future prediction of faults, assessment of severity, and timing of occurrences. Monte Carlo-based unsupervised PCA algorithm is presented in [63] for anomaly detection in PV module production factories. The proposed pre-processing algorithm effectively addresses outliers, surpassing standard methods by managing outlier substitution and considering the temporal locality of samples. Following pre-processing, an anomaly detection model is constructed using PCA, with KPIs defined for each sensor based on model errors. This methodology facilitates the robust isolation of anomalies through KPI monitoring on unseen data streams, triggering alerts when specified thresholds are exceeded. Hamza et al. [47] present a A^2-LSTM method for detecting anomalies in PV plants related to power production issues. The algorithm involves clustering the data, utilizing an attention mechanism for feature extraction, and highlights that the selection of window size plays a crucial role in the model's effectiveness for anomaly detection.

  While these studies categorize anomaly detection as a component of PdM, it is crucial to recognize it as a foundational element rather than a comprehensive solution. To truly harness the full potential of PdM, it is essential to predict specific fault characteristics, the future occurrence of faults, their severity, and the timing of these events—capabilities that anomaly detection cannot achieve. While anomaly detection is useful for identifying potential issues, it lacks the necessary predictive capabilities for effective PdM. It is an important starting point, but not a complete solution.
- **Condition Monitoring focused papers**: Intelligent monitoring system offers an interoperable, scalable, and replicable solution for comprehensive monitoring of PV plants. It efficiently handles data acquisition, storage, pre- and post-processing, as well as malfunctions and failures diagnosis. Furthermore, it assesses performance and energy yield while providing precise output power predictions. This system is implemented in [42] for monitoring and fault classification where LSTM is developed to predict PV output power under different environmental conditions. It also used an IoT platform to handle real PV experimental data, which consisted of 7300 samples with a 20-min timestep. Furthermore, ensemble learning techniques are used to detect and classify the faults. However, the paper focused on the CM of the PV plant. It activates maintenance measures only when a threshold is exceeded, which could be too late to avert a catastrophic failure. Additionally, it lacks information regarding the RUL of the components within the PV plant. Authors in [64] introduced an unsupervised operation and maintenance method that utilizes non-continuous regression models, exploiting the XGBoost algorithm and KMeans clustering. This approach is particularly responsive to indirect faults and delivers immediate alerts regarding degradation. A health state architecture for a test bench PV system is developed in [55] using XGBoost regression models for both power predictions and classifiers for power-related faults. These models facilitate timely fault detection and continuous monitoring of PV systems, irrespective of the availability and quality of historical data.

Similarly, in [61], DSS was presented, which effectively reduced costs and mitigated the energy impacts of underperformance incidents in PV systems. DSS operates exclusively on collected raw field data and employs a robust, automated, data-driven diagnostic framework to optimize PV energy output. Moreover, the proposed DSS was equipped with essential technical asset and financial management features, ensuring prompt remote and real-time failure detection. It also provides clear and actionable maintenance recommendations to address any PV underperformance issues decisively. A semi-parametric degradation path model utilizing multivariate Bernstein bases is presented in [62], capable of accurately modeling nonlinear degradation and interactions of time-varying covariates for different PV technologies. In this study, monocrystalline silicon showed an annual degradation rate of 1.12 %, equating to an RUL of 12.86 years, while polysilicon had a rate of 1.22 % and an 11.39-year RUL, underlining the importance of incorporating time-varying covariates in degradation analysis. Furthermore, authors in [68] developed a cloud-based platform that included functionalities for diagnosing and maintaining PV assets. It featured capabilities for data cleansing and modeling of PV systems, algorithms designed for early fault detection, a method for analyzing energy loss breakdown and assessing the criticality of incidents, as well as a set of automatically generated recommendations to address issues related to underperformance (such as performance losses and failures) and data concerns.

Threshold-based CM and basic monitoring methods are inadequate as these approaches primarily react to current conditions instead of proactively forecasting future failures, and cannot quantify fault severity in predicting RUL. To optimize PV field operations and enhance system performance, it is imperative to incorporate advanced PdM strategies. This includes analyzing long-term trends and accurately predicting future faults or losses, coupled with a robust fault criticality assessment tool. Advanced AI preventive and predictive functionalities for PV asset diagnosis and maintenance require prioritization. This will ensure optimized maintenance scheduling, significantly reduced downtime, and improved safety across the board.

- **PdM focused papers**: An observation mechanism of PV panels through forecasting the daily electrical power curve using the power curves of adjacent panels has been illustrated in [37]. A CNN is utilized to forecast significant deviations in the power curve, signalling potential issues with the malfunctioning panel. Authors in [38] presented a comprehensive mapping of PdM technology, emphasizing the prediction of solar energy output from PV modules. LSTM techniques are utilized to identify degradation in solar modules. A prediction model is developed by [40] utilizing the gathered time series data. LSTM effectively captures the temporal characteristics of sensor data, excelling at feature extraction from time series. An identifying model reconstructs the predicted sequence, where reconstruction error of the input sequence serves as a health indicator to assess the condition of the device. This proposed method identified potential failures several days earlier than traditional approaches. It mainly focuses on anomaly detection but lacks in identifying future fault classes. A model developed in [49] for predicting AC power output utilizes an ANN to estimate power production based on solar irradiance and PV panel temperature data. ANN was trained on a dataset obtained from the monitored plant. Real-time trend data from the PV system is then compared to the model's output, and resulting residuals are analyzed to identify anomalies and generate daily PdM alerts. The residuals are aggregated over one day and processed to detect out-of-threshold samples and signs of system degradation. Trends are identified by calculating the TMA, with an automatically determined window size. This model demonstrated a high anomaly detection rate exceeding 90 %. Additionally, the algorithm identified trends indicating deviations from normal operational behavior, providing PdM alerts to support decision-making for operatives and helping prevent potential failures. However, the shortcomings are the lack of details regarding potential causes behind issues, and predicting the remaining lifespan. Additionally, there is no examination or approach focused on classifying or quantifying the severity of anomalous behavior. Ref. [56] employed ensemble ML algorithms, including RF, XGBoost, CatBoost, and LightGBM to predict the regular maintenance requirements for a PV system. The developed method identifies DC voltage and current as crucial factors influencing the system's regular maintenance needs. Utilizing these vital features, the maintenance schedules are determined. Authors in [57] introduced a new approach to predicting solar power generation by considering maintenance tasks, issues encountered at a power plant, and weather conditions. Power generation prediction is approached as a regression problem to analyze maintenance issues. The processed data set labels are used to train the RF regression model, while future maintenance data serves as test data. The findings indicate that maintenance issues could effectively forecast power generation. Further, authors in [58], explored two distinct pipelines for PdM. The first utilizes a Hidden Markov Model to assess degradation levels on a discrete scale, employing PCA for dimensionality reduction of multivariate time series data. This method provided valuable insights into the degradation process over time. The second pipeline applied an ML approach using RF regression after selecting features from the time series data. Both methods are evaluated based on their ability to predict RUL from a random point before failure. The paper fails to identify and analyze the relevant failure types associated with predictive maintenance, which significantly undermines its applicability and depth. Researchers in [67] present a maintenance strategy that detects potential malfunctions and predicts upcoming faults days in advance. Specifically, it outlined a new and adaptable solution for predicting inverter-level faults using a data-driven approach. Although the model demonstrates the capability to identify or forecast abnormal patterns and faulty operating states, it lacks in predicting the specific class of fault, limiting its effectiveness to recognizing only a generic faulty condition.

The examined PdM studies for PV plants reveal a significant shortcoming: there is a lack of integration of comprehensive fault classification into their prognostic capabilities. While some studies focus on forecasting failures or estimating RUL, they do not effectively identify the specific type of fault that may occur. A successful predictive maintenance approach should not only predict failures but also classify faults accurately, evaluate their severity, recognize recurring trends, and forecast when they will occur. Additionally, RUL estimations should be specific to each fault class, allowing for more detailed and actionable insights. To improve maintenance scheduling, future predictive maintenance strategies should include elements that enable proactive measures based on predicted fault class, severity, patterns, and occurrence timing, ultimately enhancing the reliability and performance of PV plants.

3.2. Evaluation of AI algorithms

This section assesses AI algorithms used for predictive maintenance and fault diagnosis, highlighting their classification accuracy and the specific faults they can detect. It also addresses the advantages and limitations of these algorithms. Table 5 presents a summary of the critical features and constraints associated with AI algorithms relevant to this field.

3.2.1. Artificial neural network (ANN)

ANN consists of layers that use weights, biases, and activation functions as depicted in Fig. 6 to efficiently process data. This makes them ideal for non-linear function approximation and pattern recognition in PV system diagnostics [69]. The most popular and widely used ANN type due to its capability to approximate complex non-linear functions is Multi-layer Perceptron (MLP), which is composed of a variable number

Table 5
Pros and Cons of AI algorithm for solar PV systems.

Algorithms	Features	Limitations
SVM	**(A)**. Ability to analyze and identify nonlinear relationships from multi-dimensional data. Predict and identify a pattern in complex and multI–Variable datasets within the context of the PV system. **(B)**. Effective in high-dimensional spaces, versatile with different kernels. **(C)**. Less vulnerable to overfitting issues as compared to other algorithms. **(D)** Applicable for array performance and anomaly detection in PV system components such as arrays, inverters, and modules.	**(A)**. Limited expressiveness for complex relationships. **(B)**. Very limited ability to identify and predict specific patterns of fault class. **(C)**. Requires deep knowledge in selecting specific kernel functions.
ANN	**(A)**. Applicable for capturing complex features and handling high dimensional data from the inverter. ANN can effectively model nonlinear patterns from the PV System.	**(A)**. Prone to overfitting in case of small PV modules dataset. **(B)**. Needs expensive parameter tuning. **(C)**. Less interoperability than other simpler models.
RF	**(A)**. Can be useful in predicting remaining useful life, identifying deviations in inverter behavior and anomalies in the PV dataset. **(B)**. Robust, handles overfitting, good for high dimensional data **(C)**. RF is capable of managing high-dimensional data and can also capture non-linear relationships in PV datasets.	**(A)**. Less interpretable, can be computationally expensive **(B)**. May require a longer time to train large-scale PV modules and inverter data. **(C)**. Less sensitive to outliers.
DT	**(A)**. Excellent for forecasting PV degradation, modeling inverter's complex non-linear behavior, and complex PV faults. **(B)**. Robust and effective in handling high dimensional PV panels and inverter data. **(C)**. Interpretable, handles non-linear relationships, features importance. **(D)**. Excellent for anomaly detection and can handle high dimensional inverter and PV panel data.	**(A)**. Less interpretable as compared to other methods due to higher computational complexity. **(B)**. May require a longer time to train large-scale PV modules and inverter data. **(C)** Prone to overfitting having insufficient PV panel data and sensitive to noisy inverter data.
GBM	**(A)**. XGBoost, LightGBM, and CatBoost are effective in handling structured/tabular data (typical in PV system monitoring). **(B)**. Handle various data types, including numerical, categorical, and ordinal, making them versatile for analyzing various PV system aspects, such as power output prediction or fault diagnosis.	**(A)**. Optimal performance often requires tuning hyperparameters. This process can be time-consuming and may require extensive experimentation. **(B)**. Gradient boosting models are considered black-box models, meaning they provide little insight into underlying relationships between input features and predictions. Also, they are sensitive to outliers.
DL	**(A)**. Well suited for capturing nonlinear relationships from PV data and handling inverter multi-dimensional data. **(B)**. AE does not require labeled data for training, making it suitable for anomaly detection and unsupervised fault diagnosis in PV systems. **(C)**. LSTM networks excel at capturing temporal dependencies in sequential data, making it suitable for time-series analysis in PV systems, such as predicting power output or detecting recurring patterns. **(D)**. CNNs are adept at extracting spatial features from image-like data, advantageous for analyzing PV system components like solar panel images or thermal maps.	**(A)**. Higher computational complexity due to complex and deep architecture. **(B)**. With insufficient data from the panels and inverter, it can be prone to overfitting. **(C)**. LSTMs may suffer from the vanishing gradient problem, affecting their ability to learn long-term dependencies in data sequences. **(D)**. CNNs require fixed-size input data, which may necessitate re-sizing or cropping PV system images or sensor readings, potentially leading to information loss.

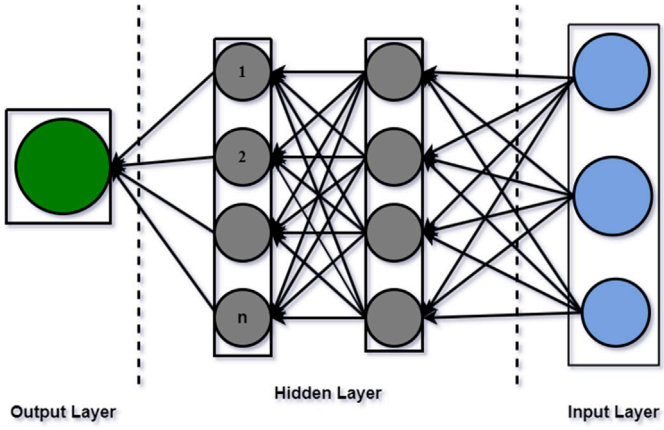


Fig. 6. The structure of ANN.

of neurons. ANN is presented in several PdM works because of its efficiency and ability to handle large, complex datasets effectively. In [49], ANN has been used to detect anomalies in the PV system, providing predictive alerts to the maintenance team for planning PdM intervention. The experimental results indicated that the model achieves an anomaly detection rate exceeding 90 %. Wavelet-based ANN has been proposed in [70] to identify fault location in an ungrounded PV system. The network demonstrated reliable performance despite noise and changing conditions. Likewise, ANNs are widely used to characterize PV system failures. Authors in [71] proposed an ANN model to identify several faults in PV arrays which are Partial Shading (PS), Line to Line (LL), Open Circuit (OC), degradation, bridge, bypass, and hybrid faults with an accuracy of 99.99 %. Furthermore, a new GA-based ANN has been developed for a GCPV system to classify various faults in PV arrays. An accuracy of 88.48 % is obtained with 3 PV arrays of 4 KW each. Further studies employing ANN can be found in Table 4 (Refs. [51–53]) for PdM, refer to Table 6, Refs. ([72–76]) for fault diagnosis.

ANN has high classification and prediction accuracy as well as a good approximation of a nonlinear function. However, ANN faces difficulties in training a large number of weighting parameters. It also requires high computational resources, is prone to overfitting, and lacks physical meaning.

3.2.2. Support vector machine (SVM)

SVM was developed by AT&T Laboratories for tasks like classification and regression [77]. It works by identifying the best hyperplane that maximizes the distance between various classes, as illustrated in Fig. 7. SVM is a very popular ML algorithm for having high precision, implemented for both classification and clustering issues. Several authors

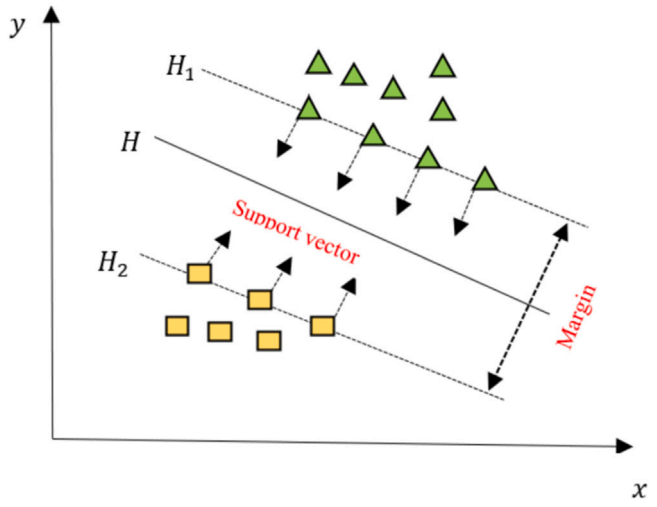


Fig. 7. The basic idea of support vector machine [26].

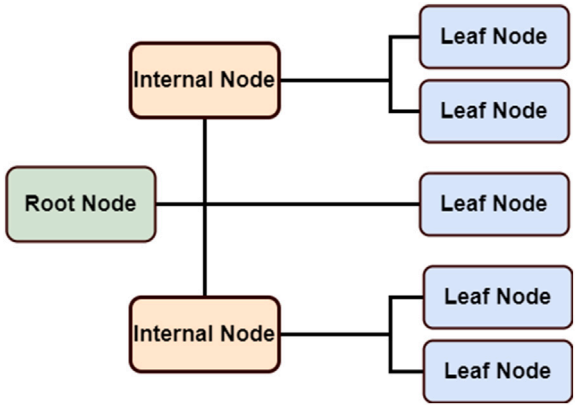


Fig. 8. Attributes of DT.

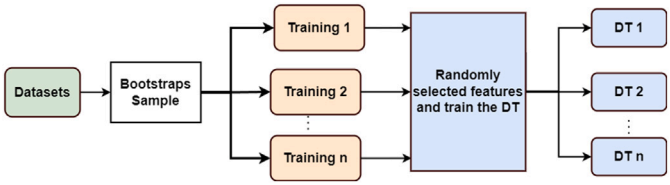


Fig. 9. Random forest generation steps.

implemented SVM for PV system PdM and fault diagnosis. Anomaly detection using a one-class SVM is developed in [60] for monitoring the DC side of a grid-connected PV system and for assessing temporary shading to differentiate between normal and abnormal features. Supervised ML algorithms using SVM are proposed in [78] and [79] for the detection and classification of multiple faults in GCPV and $3 * 4$ PV array, respectively. The obtained training and testing accuracy for SVM in [79] is 98.2 % and 97 %, respectively. Multiple faults are considered including OC, Short Circuit (SC), and inadequate irradiance. For a broader overview, refer to Table 4, which includes additional studies on SVM for predictive maintenance (Refs. [42,54]) and refer to Table 6 for SVM based fault diagnosis algorithms (Refs. [80–83]).

The disadvantage of SVM is its sensitivity to the choice of hyperparameters and kernel functions. SVM performance heavily relies on selecting suitable parameters, such as the regularization parameter and kernel type. Choosing an inappropriate kernel or hyperparameter values results in suboptimal model performance or overfitting. Additionally, SVMs may become computationally expensive and less efficient, especially with large datasets, as they involve solving a quadratic optimization problem.

3.2.3. Decision tree (DT)

The Decision Tree (DT) algorithm divides data into subsets based on feature values, resulting in a hierarchical tree structure for classification through a series of if/else decisions. Each node acts as a decision point that directs samples to various branches until the desired level of class purity is obtained, promoting both accuracy and efficiency in classification [84,85]. DT consists of root, internal, and leaf nodes as depicted in Fig. 8. Simulation results in [86] show that the DT model is capable of correctly predicting faults in a 4 KW GCPV system with 94.7 % accuracy and 1400 observation/sec prediction speed. For additional studies utilizing the DT algorithm, refer to Tables 4 and 6, Refs. [78,86]. DT involves repetitive attribute testing and branching (known as splitting), to identify optimal nodes for classification. This method is recognized for practical accuracy and computational efficiency [85].

A disadvantage of DT is susceptibility to overfitting, particularly when the trees are deep and capture noise in the training data. Deep trees tend to fit the training data too closely, resulting in poor generalization to new, unseen data. Although techniques like pruning and setting depth limits can mitigate overfitting, finding the right balance remains a challenge. Additionally, DTs may struggle with capturing complex relationships and interactions within the data, especially when dealing with nonlinear patterns.

3.2.4. Random forest (RF)

RF method is an ML ensemble technique used mainly for classification and regression tasks. It constructs numerous DTs by utilizing bootstrap sampling and selecting features randomly, resulting in a variety of uncorrelated trees. In making predictions, each tree contributes a vote; for classification, the decision is determined by majority votes, while for regression, the predictions are averaged. This ensemble approach improves both the accuracy and reliability of the predictions [87]. The construction of RF is shown in Fig. 9 starting from datasets to decision trees. Several PdM and fault diagnosis algorithms based on RF are summarized in Tables 4 and 6, Refs. ([57,78,88–90,90–93]).

RF lacks in interpretability compared to simpler models. The ensemble nature of the random forest, combining multiple decision trees, makes it challenging to trace individual predictors' contributions to the model's output. While Random Forest excels in predictive accuracy and robustness, especially with complex datasets and diverse feature types, the trade-off is a somewhat reduced interpretability, which can be a concern in situations where understanding feature importance or explaining model decisions is crucial.

3.2.5. Gradient boosting methods (GBM)

Gradient boosting techniques like XGBoost, CatBoost, and LightGBM enhance PV fault diagnosis and PdM by improving model accuracy and efficiency. These methods utilize features such as regularization, parallelization, and handling of missing data, making them effective for identifying PV system faults.

XGBoost is a highly efficient and scalable algorithm for tree boosting in ML, noted for its remarkable predictive performance relative to RF [94]. It utilizes regularization to curb overfitting, supports parallel computation for improved efficiency, and adeptly handles missing data [95]. These attributes make it exceptionally suitable for large datasets and for

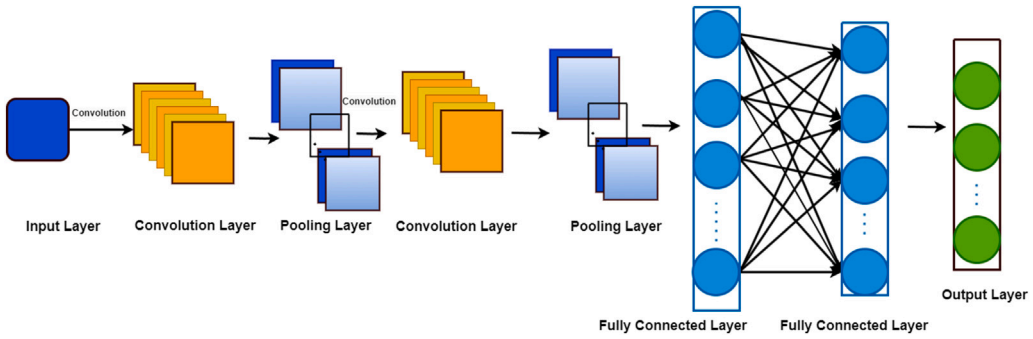


Fig. 10. Architecture of CNN.

a wide range of ML tasks. The effectiveness of XGboost can be used for PV system PdM and fault diagnosis. The XGboost-based intelligent and data-driven O&M framework has been proposed in [64]. It provides O&M suggestions to the engineers regarding fault identification and evaluates various operational statuses of the plant. A health-state architecture for advanced PV system monitoring is demonstrated in [55] where XGboost regressed is used for predicting output power. Fault conditions are detected with an 83.91 % sensitivity for synthetic power-loss events (5 % reduction) and 97.99 % sensitivity for field-emulated failures in the test-bench PV system. A PdM based on XGboost is deployed [56] to predict the maintenance needs of the PV system, with XGBoost proving to be the best model, achieving 98.62 % accuracy and 94.37 % precision. However, it requires careful and time-consuming hyperparameter tuning.

CatBoost is designed to minimize the need for extensive hyperparameter tuning. It introduces novel strategies such as handling categorical features more effectively and implementing a robust method for dealing with missing data. This makes CatBoost less prone to overfitting, more resilient to default parameter settings, and often requires less manual tuning than XGBoost, addressing one of the disadvantages of the latter. CatBoost can handle categorical data during the training phase, eliminating thus the need for separate data preprocessing steps [96]. The disadvantage of Catboost is its relatively longer training time compared to simpler models as it employs a more sophisticated algorithm with additional optimizations. While CatBoost is efficient in handling categorical features and reducing the need for hyperparameter tuning, its training time may be a consideration, especially in real-time or resource-constrained applications. LightGBM shares similarities with CatBoost in terms of being a gradient-boosting algorithm, but it uses a different approach that can help address certain disadvantages. It provides options for handling categorical features and supports parallel and distributed computing, contributing to its efficiency in comparison to CatBoost. While both CatBoost and LightGBM have their strengths, LightGBM's focus on speed and scalability can be advantageous in scenarios where computational efficiency is a priority. Moreover, LightGBM is proficient in handling categorical features, similar to CatBoost, providing an additional advantage over XGBoost. This capability broadens its applicability and enhances its performance in diverse ML scenarios.

Studies on PDM that have implemented Gradient-boosting algorithms are summarized in Table 4 (Refs. [55,56,64]). For fault diagnosis, refer to Table 6.

3.2.6. Deep learning approaches

DL is the unsupervised subset of the ML and is a training approach that uses NN with multiple layers. A neural network with more than two layers is referred to as DL. The main algorithms of the DL are CNN, LSTM, RNN, deep belief networks, generative adversarial networks, and other hybrid techniques. Recently, PdM based on DL algorithms is employed in PV power plants for reducing maintenance-related issues [97]. This subsection discusses the most utilized algorithms such as CNN, LSTM and autoencoders for PdM and fault diagnosis. Further information on the rest of the DL algorithm can be found in [98].

CNN. CNN stands for Convolution Neural Networks. It is a special type of NN for data processing that has a known grid-like topology. Instead of relying on standard matrix multiplication, CNNs incorporate convolution in one of their layers [99]. It consists of input, hidden, and output layers in which pooling or subsampling and convolution layers are used. Fig. 10 shows the structure of a deep CNN. Rectified Linear Unit (ReLU), one of the most popular activation functions is used in the CNN layers helping NN to accelerate the convergence speed. Fully connected layers in the CNN perform classification or regression tasks based on the learned features.

CNN can enhance the PdM and fault diagnosis for the PV System. Authors in [37] deployed CNN-based DL architecture for PdM of PV panels, predicting the daily power curve of individual PV panels in relation to neighboring panels. Additionally, it helps in analyzing the deviation of the power curve between the predicted and observed/actual values to identify the malfunctioning. A hybrid DL model with multiple temporal windows in [48] is developed along with IoT to evaluate estimated output power generation. The YOLOv3 CNN model has been trained on the thermal images of the solar panels [100] to compare the actual and estimated fault values. A methodology based on CNN is developed in [101] to estimate a maintenance strategy's impact and evaluate the potential benefits of adopting a strategy mainly based on predictive actions while considering realistic limitations associated with monitoring using automatic fault detection tools. The results show that a predictive strategy reduces the need for urgent interventions by 10 % while maintaining average performance. CNN has been used in several notable studies for PdM, refer to Table 4, Refs. [38,40,42,43,46] and for fault diagnosis, Table 6, Refs. [74,81,89,102–104].

LSTM. Long Short-Term Memory, acronymized as LSTM is a Recurrent Neural Network (RNN) variant, designed to manage long-term dependencies [105]. It features memory cells and three gates (Forget, Input, Output) and is used in natural language processing (NLP), sequence-to-sequence tasks, and recognition applications. The general architecture of LSTM is illustrated in Fig. 11. Authors in [106] proposed a short-term PV power prediction model based on LSTM and K-nearest neighbors (KNN) to provide predictive data support for the power grid's distributed photovoltaic output fluctuation model. Work in [38] focuses on controlling the degradation of PV modules that are exposed to a variety of climatic loads. This study proposed an LSTM model for PdM involving an elaborative system capable of solving a time series problem. Autoencoder-LSTM, Facebook-Prophet, and Isolation Forest are proposed in [46] to identify PV system's healthy and abnormal behaviors. Authors in [43] developed a method based on BiLSTM architecture to perform multiple PV plant monitoring using an IoT platform along with real-time anomaly detection and next-day power generation. The

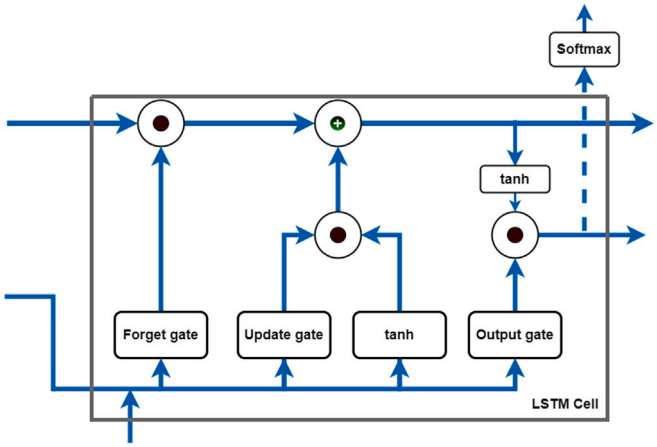


Fig. 11. Architecture of LSTM.

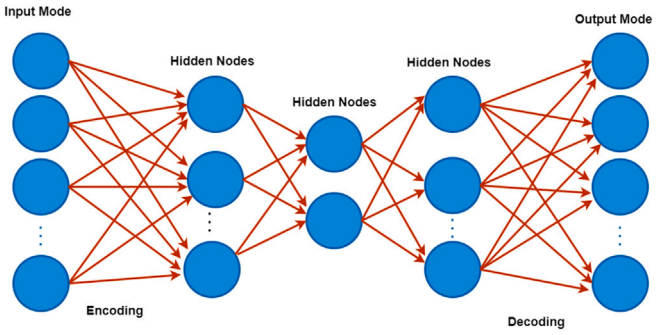


Fig. 12. Architecture of AutoEncoder.

results showed that BiLSTM is the best model with desirable error predictors, such as Mean square error (MSE), Mean absolute percentage error (MAPE), Mean absolute error (MAE), and coefficient of determination (R^2) values of 0.0072, 0.1982, 0.0542, and 0.9664, respectively.

Several studies utilized LSTM for PV system's PdM, refer to Table 4, Refs. [37,41,44,45] and for fault diagnosis, Table 6, Refs. [74,102,107–110].

AutoEncoder (AE). An autoencoder (AE) is an ML algorithm for dimensionality reduction and feature learning. It comprises an encoder that compresses input signals into a lower-dimensional latent space and a decoder that reconstructs the original input as shown in Fig. 12. The AE aims to minimize the difference between the original and reconstructed data. It effectively reconstructs normal data with low error, but for anomalies, it will result in high reconstruction error, which will serve as an identifier for anomaly detection. A data-driven fault prediction model based on LSTM and AE has been developed in [40] for solar arrays. A predictive model is developed by gathering time series data, with the reconstruction residual threshold for normal data determined using AE. During the identification process, the ultimate prediction result is derived by assessing the reconstruction error against the prediction sequence threshold. Studies employing AE for PDM and fault diagnosis are summarized in Tables 4 and 6 (Refs. [40,46,111,112]).

Different AI methodologies come with varying degrees of complexity, accuracy, advantages, disadvantages, and constraints. Table 5 outlines key benefits and drawbacks of SVM, ANN, DT, RF, and gradient boosting methods when it comes to developing PdM and fault diagnosis systems for PV technologies.

3.3. Experimental verification of AI model

Edge computing, commonly known as 'edge AI,' integrates AI into embedded systems situated at the network's periphery. This approach improves PdM and fault diagnosis for PV systems, offering advantages such as decreased latency, better privacy, and reduced data transmission expenses. By facilitating early detection of faults and accurate scheduling of maintenance, ML-driven edge AI aids in minimizing downtime, avoiding expensive repairs, and enhancing the overall lifespan of PV systems. An embedded system designed for detecting and diagnosing PV module faults using thermal images and Deep CNN (DCNN) is presented in [113], where DCNN-based classifiers are embedded into a Raspberry Pi 4. The workflow model consists of Tensorflow (TF) model training and evaluation, TF model conversion to lite TF and deployment on an edge device. The system facilitates real-time analysis to support decision-making such as cleaning, changing PV modules, removing/replacing faults and diodes etc. Moreover, a GSM module (SIM808) is used to notify plant operators by SMS and email about the PV operational status. Likewise, an edge computing framework based on Raspberry Pi is developed in [65] for condition monitoring in decentralized PV Systems. Edge, fog, and cloud layers are integrated into this framework for the effective identification of anomalies through online as well as offline configurations. Authors in [114] used embedded edge devices to detect solar panel faults. Real-time PV panel EL videos are captured by a charged-coupled device (CCD) camera, and processed by the OpenCV and CNN based on an edge device processing unit that detects PV panel anomalies and faults. The rk3399 pro-vision chip is the edge device's main control unit with a central frequency of 1.8 Hz and a CPU with 64 64-bit processor arm. This chip is supported by the NPU NN unit to enable AI hardware acceleration. This chip is supported by a Neural Processing Unit (NPU) to provide AI hardware acceleration.

Deploying AI models on edge computing devices facilitates immediate anomaly detection and system alerts for PdM in solar power plants, improving reliability and reducing downtime. Integrating this technology presents challenges due to limited computational resources and the need for real-time processing. However, it requires the development of optimized AI models and a collaborative effort between hardware and software to ensure effective implementation in maximizing the performance of PV systems.

4. PV data source and analysis for PdM and fault diagnosis

Advancements in data analytics play a crucial role in various PV system aspects, including operation management, load profiling, predictive maintenance, fault diagnosis, and energy forecasting. These systems rely on sensors, control technologies, and methods for data transmission. Therefore, it is essential to understand the sources of PV data to perform effective analytics for maintenance and fault detection. The categorization of photovoltaic data sources includes sensor data, historical data, weather data, IV curve data, load profiles, equipment metrics, and synthetic data. To this end, data loggers are crucial to the PV system as they continually track and log information from different sensors and meters [54]. They collect real-time data on essential factors concerning solar energy production, system efficacy, and environmental conditions.

4.1. Sensor data

Sensory systems play a crucial role in data monitoring and collection for operational and maintenance purposes in PV systems. Key units include phasor measurement units (PMUs), wireless networks, cloud storage, and IoT sensors [115]. Common sensor types include irradiance sensors (pyranometer, pyrheliometer), temperature sensors (thermocouple and thermistor), current/voltage sensors, power meters, anemometers (wind speed and direction), humidity sensors, tilt-angle sensors, and load current sensors. PV industries use station temperature, pump speed, flow speed, and ozone concentration mounted sensors for anomaly detection in the production line [63]. Authors in [49,54]

utilized data loggers to collect measurements from pyranometers and thermometers, along with metering equipment for voltage, current, and generated power. Furthermore, real-time data on AC/DC currents, voltages, ambient temperature, and solar irradiance in [60] is gathered using various sensors including Hall effect, AC adapter, resistive divider and K-type thermocouple. A ZigBee wireless system is developed in [116] to manage remote solar power operations, where a potentiometer is used as a tilt sensor. Moreover, authors in [46] used 22 inverter sensors to measure the generation rate (AC and DC powers), irradiation, and ambient/module temperatures.

A key source for data mining is the IoT and authors in [48] collect real-time PV data via an IoT module and opera digital platform. Sensors are controlled by a microprocessor that wirelessly uploads information on irradiance, temperature, current, voltage, and power. Furthermore, authors in [43] proposed an IoT platform architecture for monitoring multiple PV facilities, tracking power generation and various parameters like DC voltage, power factor, and active/reactive powers. AI models can be trained on the data extracted from the sensors to know when the equipment such as PV arrays or modules is likely to fail. These sensory data offer valuable insights into the health and performance of the PV system components and help predict potential equipment failure.

4.2. Historical data

Historical data encompasses past trends and system details logged over time, helping identify patterns and relationships that enhance understanding of future occurrences. This data includes information on the timing, location, fault type, and maintenance activities within PV systems. Historical Records of equipment failures and maintenance schedules can be used to train ML models to forecast the likelihood of potential component failures in future. Such insights reveal common failure modes, enabling predictions about when similar issues may arise in the future, such as forecasting degradation in PV modules that have shown a history of failure.

Analysis in [52] used historical data of electrical and environmental inputs such as DC and AC electrical tags (current, power, and voltage), temperature tags (ambient, panel, and internal), and irradiance tags (Global tilted irradiance (GTI), Global Horizontal Irradiance (GHI)) extracted on a 5-min average by SCADA system for offline performance assessment and normal training period selection. Author in [37] collected historical time-series power data for the target and neighboring PV panels to identify faulty ones. The target panel's known measurements are utilized in training and test data for the predictor. Furthermore, researchers in [39] collected one-year historical data for string modules output current (from August 2020 to August 2021) and used for anomaly detection using KMeans and LSTM. A total of 8400 PV modules and 7 combiner boxes were monitored. Authors in [117] optimized O&M tasks by developing an AI-enabled power prediction model for individual inverters and used historical production data for training. A total of more than 7583 MPPT data is gathered every 5 min, and at least 6,369,720 batch data inputs for model training. Moreover, PV power forecasting based on maintenance activities reported in [57] used a dataset from a 1 MW capacity PV power plant from 2012 to 2020. Researchers in [118] detected anomalies for a GCPV system based on degradation rate by taking 5-year data from a data logger between 2012 and 2017 with a 5-min interval. The historical dataset considered includes irradiance, module temperature, string voltage/current, and AC power.

4.3. Weather and load profile data

In PV systems, weather data and load profiles are key to understanding the factors affecting performance and helping predict faults. Weather data includes atmospheric conditions like temperature and precipitation. In [119], a model predicts PV energy using data such as irradiation and air temperature. A system in [38] uses weather forecasts to estimate solar module output based on historical rainfall (mL), temperature (°C), and radiation data (W/m^2). Furthermore, authors in [64] investigate weather impact on low-power PV generation using meteorological data (such as rainstorms, blizzards, hail, and sandstorms) to develop reference curves for optimizing plant operations.

The load data, on the other hand, provides insights into power demand at various times, allowing AI models to predict potential failures in PV components [120]. Various datasets are created from PV system components, including transmission lines and circuit breakers for monitoring and management purposes [121]. Likewise, equipment metrics such as power, voltage, frequency, and current can contribute to power quality assessments, such as harmonic distortion and flicker index. Additionally, having a record of past faults, disturbances, trip times, and voltage dips helps in equipment diagnostics and analysis. Moreover, energy management data with remote telemetry units used for renewable energy tracking can also help in monitoring and fault prediction for the PV system [122].

4.4. IV curves data

IV curves can be used for PdM and fault diagnosis by monitoring deviations from baseline data, detecting anomalies indicating potential issues, and identifying specific faults. Monitoring IV curves may provide insights into the overall performance of the PV system and feed directly into AI algorithms. Significant deviations from the expected curves indicate systemic issues such as panel faults, inverter malfunctions or mismatched panels. In [60], fault detection is performed, where unknown parameters of one diode model (ODM) are determined via an efficient heuristic algorithm based on I–V characteristic curves. Fault classification for line-to-line, open circuit, partial shading, and degradation consisting of six operating conditions is investigated in [123]. Typically, IV characteristics are studied for a better understanding of fault patterns in PV modules. Authors in [124] employed IV characteristics curves with 110,250 records wide dataset for fault detection and diagnosis in PV arrays. Each record has the format: temperature, irradiance, V_{mpp}, Maximum Power Point (MPP) current I_{mpp}, OC voltage V_{oc}, SC current I_{sc}, voltage measurement from PV cell V_{Cell}, current measurement from PV cell I_{Cell}, and operational status. Temperature and irradiance values are retrieved directly from the PV Geographical Information System (PVGIS) service. MPP voltage model $V_{mppModel}$, MPP current model $I_{mppModel}$, OC voltage model $V_{ocModel}$, and SC current model $I_{scModel}$ are retrieved from each generated IV curve.

4.5. Synthetic data

Synthetic datasets are simulation-based generated data that mimic the characteristics and behavior of real-world PV systems under a controlled environment. It is essential to recognize that real data is often not easily accessible; therefore, synthetic data is of critical importance. A synthetic database is developed in PdM for fault detection and diagnosis that typically includes simulation of PV modules under normal and faulty conditions such as partial shading, cell short circuits, bypass diode short circuits, etc [50]. Authors in [37] synthetically generated daily power curves of the target panel and the two neighboring panels, along with actual measurements and applied CNN-based methods for predictive maintenance. Researchers in [60] developed a PV array simulation model to describe the PV system's normal condition using a co-simulation between Matlab/Simulink and Physical Security Information Management (PSIM). A distribution network with multiple DGs integration has been modeled in Matlab/Simulink environment in [125], where three-phase voltage and current datasets are generated for fault case scenarios.

4.6. Experimental setup for data collection

An experimental setup of an Analog-to-digital converter (DAQ NI USB 6008) with adjustable resistive load is developed in [50] to emulate the fault scenarios and measurement for PV characteristics curves. A SCADA system is used in [52] for electrical and environmental to

record historical data averaged on 5-min, and inverter's manufacturer electrical parameters for the on-site inverter technology. Likewise, in [37], real measurements of the power curve from three adjacent panels over 157 days are provided by SOLA Sense Ltd. from a solar monitoring pilot system. The authors in [48] utilized an Arduino-based IoT module to gather electrical and environmental data from a PV system. The system includes ambient sensors to measure solar irradiance and temperatures (module and ambient). Additionally, voltage and current sensors are also connected to the Arduino board. These sensors transmit data to the Arduino microprocessor via the Zigbee protocol, allowing for seamless wireless communication in open areas, which is essential for PV system deployment. In [67], a SCADA system is used to log data from various sensors, capturing 10 different signals with a sampling interval of 5 min. These signals comprise both electrical and environmental measurements. In [60], a 34901 A 20-channel multiplexer integrated into the Data Acquisition (DAQ) and switch unit of an Agilent 34970 A device is used to collect measurements from multiple sensors, including irradiance on an inclined plane, ambient temperature, DC current, and voltage. The sampling interval of 1 min is used, and data monitoring is carried out using LabVIEW software. Moreover, a PVPM2540C device is used to gather IV curves and Artificial Bee Colony (ABC) algorithms to identify unknown parameters of a single diode model. The researchers in [116] employed an Arduino, H-bridge motor driver circuit, and a DC motor to adjust the tilt angle of a PV panel in alignment with the sun, keeping the azimuth and elevation angles constant at noon. A high-precision and affordable monitoring system is designed in [42] for monitoring PV plants. The data from temperature, humidity, power, voltage, current, and irradiance sensors is transmitted to data loggers which consist of a microcontroller and ESP8266 for initial processing. The ESP8266 is a mini WIFI cost-effective IoT device that sends data to the server for recording and ultimate processing of the data (cloud computing). Furthermore, research conducted in [126] used an IEEE 13-node test feeder, incorporating distributed generation sources and uncertainties using the RTDS RSCAD software. Phasor measurement units (PMUs) were strategically positioned in optimal locations to capture real-time data. These PMUs deliver real-time three-phase current signals, facilitating precise and prompt fault diagnosis.

In addition to electrical analysis, thermal details of PV systems (including damages) which are not visible to the human eye can be obtained from various sources, especially through thermal infrared (IR) cameras. IR imaging is employed to assess the surface and internal temperature of PV panels. Authors in [41] used SmartView thermal imaging software to analyze the images captured by the IR cameras. Likewise, electroluminescence (EL) imaging of cells/panels captured by EL camera are used in [127] for automatic detection of PV cell defects. Moreover, infrared radiation cameras are used in [51] to capture 20,000 real datasets for solar plant IR images to accurately predict and classify anomalies in solar modules. The temperature distribution on the PV modules is captured remotely by IR cameras and subsequently, ML/DL algorithms predict the normal and faulty modules along with classifying the anomaly type. Unmanned aerial vehicle (UAV) devices with IR cameras are used to collect images due to convenience and easy application across real PV plant sites. Besides, authors in [113] used thermos-camera FLIR T540 type to capture IR thermography images of a PV array (4 mono-crystalline modules) for fault detection and diagnosis.

4.7. Critical analysis of PV data sources

4.7.1. Challenges in data standardization

A significant challenge in using AI for PdM and fault diagnosis arises from the diverse range of data sources and the absence of standardized formats. Fault diagnosis systems typically depend on a mix of different data types, each offering distinct perspectives on the system's health. For example, sensor readings, such as temperature, wind speed and direction, tilt angle, humidity [63], voltage, current, and generated power [49,54], AC/DC currents, ambient temperature, and solar irradiance [60], tracking power generation, DC voltage, power factor, and active/reactive powers [43] analyzes the dynamic behavior of the PV system. These data sets are typically time-dependent with varying samples and units. Visual content, such as thermal and visual inspections, provides spatial insights into the condition of PV components [41,44,81,104]. However, these images vary in resolution, format (like JPEG, PNG, etc.), and colour depth. Even among a single data type such as IV curves, there are considerable differences. These variations can include the quantity of data points that make up the curve (110,250 records [124] vs 5410 and 4612 data samples from the synthetic and real-time irradiation and temperature respectively [123]) and the storage format (CSV, binary). Moreover, the use of synthetic data, such as [60] developed a PV array simulation model to describe the PV system's normal condition, synthetically generated daily power curves of the panels [37] highlights the issue of format diversity. Although these data mimic real-world sensor readings, their unique format, noise characteristics, and discrepancies emphasize the need for standardized data representations.

The lack of standardization in data types and formats, including those for synthetic data, poses several challenges. It complicates data integration and impedes the creation of reliable AI models, as preprocessing and feature engineering become too specific to each application. This also impacts model portability; a model trained on one data format may not transfer easily to another. Most critically, without standardized formats, objectively comparing the performance of different fault diagnosis algorithms is challenging, as each may need unique preprocessing tailored to its training data.

4.7.2. Impact on ML model performance

The absence of standardized data formats in PV PdM and fault diagnosis greatly influences the effectiveness and advancement of AI models. The variety of hardware and data collection methods employed in PV systems, as highlighted in Section 4.6, exacerbates this issue. This section explains how this diversity impacts data integration, feature engineering, model bias, as well as comparison and development costs.

- **Data Integration challenges**: Integrating data from different sources is a major hurdle. Data acquisition systems such as DAQ NI USB 6008 in [50] typically generate time series data with a specific sampling rate of 10 K/S (10,000 per second) and numeric format. Similarly, [52] utilized SCADA system with 5-min averaging for extracting and collecting historical data. Almonacid-Olleros et al. [48] used an IoT module to capture analogue signals within a 0–5 V range, later converted into digital format. Mellit [113] utilizes a Raspberry Pi 4 to retrieve data from Firebase and convert the IR thermal image to the desired dimensions (e.g., 224 × 224). In contrast, [51] used the public dataset released in 2020 by Raptor Maps Inc., which comprises IR images measuring 40 × 24 × 1 pixels (1 channel) taken by aircraft and UAVs with spatial resolutions between 3.0 and 15.0 cm per pixel. The variation among data sources makes data fusion challenging, necessitating tailored scripts and preprocessing procedures for each source. Integrating data from SCADA systems, IoT environmental sensors, and IR thermal images processed via Raspberry Pi, which come in various formats and sampling rates, poses considerable challenges in terms of synchronization and alignment. This makes it very challenging to create a unified dataset for training AI models that are robust and generalizable.
- **Feature Engineering Complexity**: Different types of data formats have a direct effect on feature engineering. The distinction between processing data from a PMU [126], which delivers real-time three-phase current signals for accurate and timely fault diagnosis and localization, and analyzing EL imaging of cells/panels captured by EL camera [127] is undeniable. The PMU provides real-time monitoring of voltage, current, and frequency across multiple points in the grid by precisely measuring the phase angle and magnitude of

the electrical signals, crucial for maintaining system stability, while the EL camera effectively detects failures in PV modules by capturing emitted light when a voltage is applied. This method produces high-resolution digital images that highlight defects such as cracks and degradation, ensuring that potential issues are identified with precision. A thorough understanding of these two powerful technologies is imperative for effective monitoring and optimization of performance. Extracting features from PMU data for fault detection may include determining the magnitude and phase angle of voltage and current phasors, in addition to tracking the frequency and the rate at which these values change. However, detecting faults from EL images entails examining the intensity and distribution of emitted light to pinpoint defects. This requires the development of custom feature engineering processes for each data source, which complicates automation and standardization significantly.

- **Difficulty in Model Comparison and Benchmarking**: Comparing and benchmarking AI models present considerable challenges when data formats differ. For instance, consider two diagnosis models: one trained on data from a DAQ NI USB 6008 [50] and the other from a PMU [126]. The discrepancies in data formats and sampling rates require each model to undergo unique preprocessing steps. This complexity makes direct performance comparisons almost impossible, as variations in these steps can significantly influence the results.
- **Increased Development Time and Costs**: The lack of standardized data formats drastically escalates financial costs and development time for AI-driven PdM and fault diagnosis systems, particularly during hardware implementation and verification. Deploying AI algorithms on embedded systems invariably requires multiple data processing units, significantly inflating the overall system cost. For example, when implementing DCNN classifiers using thermal images on a Raspberry Pi 4 in [113], the workflow's dependency on TensorFlow for training and evaluation, coupled with the necessity of a GSM module (SIM808) for data transmission, not only inflates the bill of materials but also demands complex data pipelines that increase labour costs. Moreover, adapting a model for processing event log videos from a CDD camera using OpenCV on an RK3399 ProVision chip equipped with an NPU [114] requires specialized video processing, data format conversions, and model optimizations, leading to elevated engineering time and potential software tool costs. These disparate hardware and software configurations—driven by non-standardized data formats inevitably result in exhaustive efforts and need for specialized skills, thereby significantly increasing the overall costs of developing and deploying AI-based PdM and fault diagnosis solutions on edge devices.
- **Bias and Inconsistency**: Non-standard data formats undeniably introduce bias into models. When data from a specific sensor type collected through a more standardized interface proves easier to process, the model will inevitably skew towards that sensor type, regardless of its actual usefulness for fault diagnosis. For instance, if data from a particular weather station is consistently easier to handle, the model will disproportionately emphasize this data, neglecting other, more pertinent weather sources. This can create a model that performs well for certain faults while failing significantly with others.

5. Fault diagnosis

PdM provides early fault alerts but lacks detailed information on specific fault types without extra diagnostics. Integrating PdM with fault diagnosis provides a comprehensive approach, allowing for accurate fault prediction, identification, and diagnosis. To complement this process, it is essential to consider both fault severity and fault timing. Fault severity assesses how critical a fault is, allowing for the prioritization of diagnostic actions based on its potential impact, with severe faults warranting immediate attention. Likewise, understanding fault timing, including when faults occur and their duration, helps identify patterns and recurring issues. This leads to improved classification models and greater diagnostic precision. Together, these elements strengthen the overall effectiveness of fault diagnosis and classification, leading to a more proactive and resilient maintenance strategy for PV systems.

Faults in PV systems are of various natures, such as electrical faults (e.g., fuses, DC box, wiring, diode bypass, grounding system), as well as physical and environmental issues, PV inverter failures, and grid-side failures [19]. These faults can be either temporary or permanent depending on their duration and can be due to internal or external factors, resulting in the system's performance degradation. A permanent fault, enduring over an extended period, may result from factors like aging or issues such as loose or disconnected electrical wiring in the system. In contrast, a temporary fault occurs within a specific timeframe, often caused by external factors like the accumulation of dust, dirt, or snow on the PV module surface. Additionally, shadows cast by nearby structures, such as buildings or trees, as well as passing clouds overhead, can also contribute to a transient decline in PV performance [128]. Faults in general are classified according to their general characteristics and stages. This section provides an overview of common types of PV system faults depicted in Fig. 13. This study focuses extensively on faults in PV arrays, providing a detailed examination of their characteristics. While it acknowledges faults in other components, the main aim is to explore in depth the specific types of issues encountered within PV arrays.

Table 6 reports a comprehensive critical review of existing literature on fault diagnosis where papers are compared based on (i) type of ML algorithm used, (ii) PV system, capacity and data type, (iii) inputs and output of ML algorithms for fault diagnosis- faults type, (iv) Fault daignosis (FD) ML accuracy (v) real-time implementation/verification of ML developed algorithms, (vi) indication of fault severity levels, (vii) estimated fault time and (viii) future fault forecasting. The important columns of Table 6 are explained below which will help engineers and researchers to select suitable fault diagnosis methods pertinent to specific requirements. A critical discussion of Table 6 is provided in the subsequent section.

- **Inputs/outputs of ML algorithm**: This column outlines the inputs and outputs of the ML algorithm employed for PV fault diagnosis. Inputs typically consist of various data sources and parameters collected from the PV system, such as sensor readings, environmental conditions, performance metrics, or other relevant data sources. These inputs serve as the basis for training the ML algorithm to recognize patterns indicative of faults or anomalies. Outputs, on the other hand, represent diagnostic results produced by the ML algorithm, which may include classifications of detected faults, confidence scores, or other indicators of system health.
- **FD AI Accuracy**: This column outlines the accuracy of the AI algorithms under consideration. It includes information on validation, testing, and training accuracy percentages, which demonstrate how effectively the algorithm diagnosed and classified faults. Furthermore, Mean Squared Error (MSE) and Normalized Root Mean Square Error (NRMSE) are also provided for certain AI algorithms to further evaluate their performance. These metrics are essential for understanding the reliability and efficiency of the algorithms in use.
- **Fault Severity level and Estimated Fault time**: The columns labeled "Fault Severity Levels" and "Estimated Fault Duration" in Table 6 show whether each paper discusses fault diagnosis with varying severity levels, such as mild, moderate, or extreme, as well as the duration of faults from when they start until resolved. A "Yes" indicates that the paper covers these aspects, which contribute to the development of prompt and efficient maintenance strategies. Conversely, a "No" denotes that these topics are not addressed.
- **Future Fault Prediction**: The papers on fault diagnosis reviewed in the table mainly focus on classifying faults into their respective type classes, paying little attention to predicting future faults,

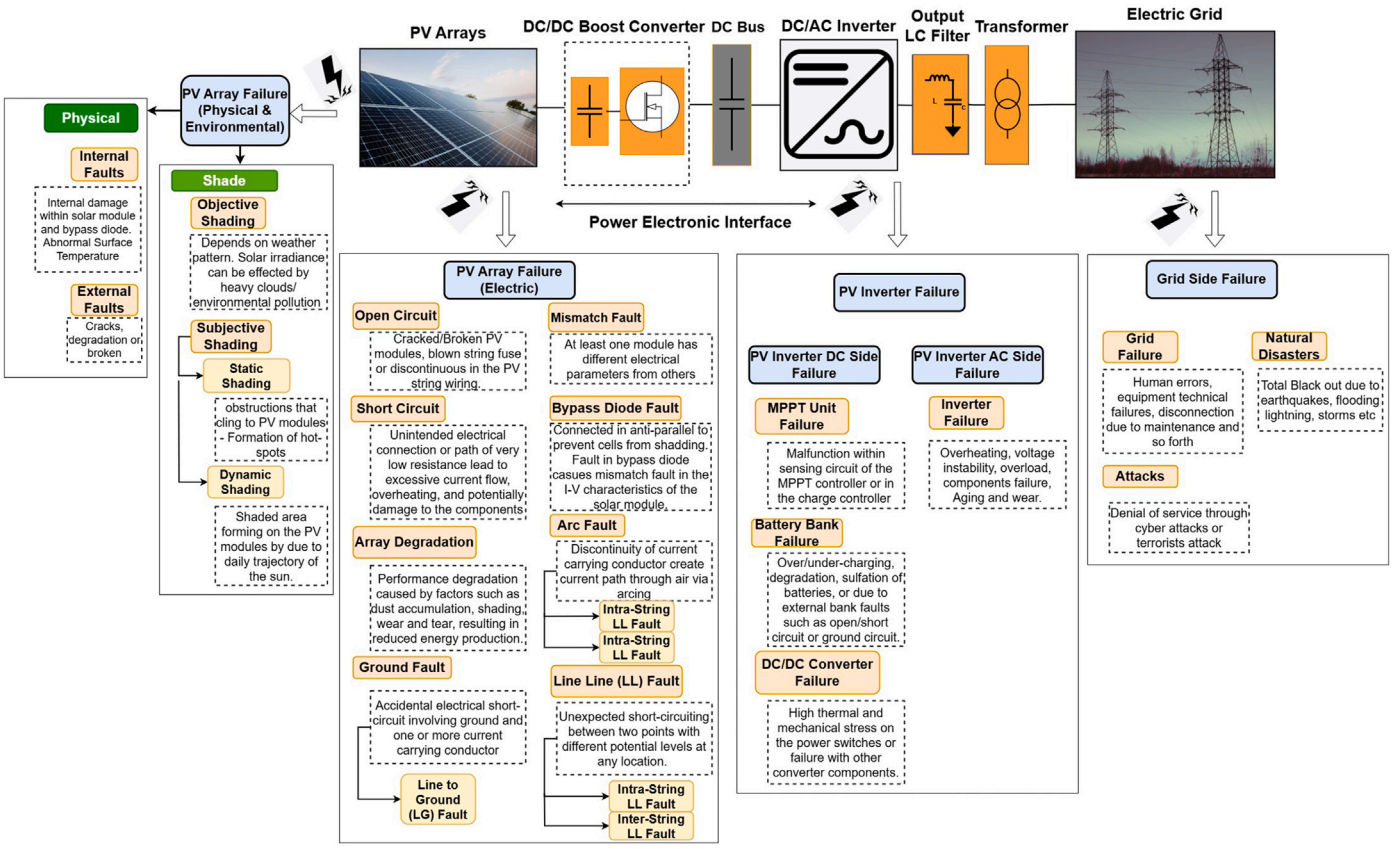


Fig. 13. Classification of major faults in PV system.

assessing their severity, and estimating the time of faults. Integrating these aspects into fault diagnosis systems could greatly improve their predictive and diagnostic effectiveness. Consequently, the relevant column marks "*No*" for studies that overlook this feature and "*Yes*" for those that incorporate it.

In addition to the comprehensive information in Table 6, following points can be observed:

- Among 53 articles reviewed, only three [76,86,129] focused on real-time implementation, i.e., experimental verification of AI model on the embedded system.
- Very few papers focused on the PV system's AC side, which includes inverters, transmission lines, and grids. Most researchers considered major faults that can occur on the DC side of the PV system, including PV modules, strings, and arrays.
- Articles reviewed primarily focus on fault diagnosis, detection, and classification based on historical and current data. This typically involves collecting and preprocessing data, annotating data based on the known fault types and classes, and training the AI algorithm to detect and classify faults. Although this method has proven effective in diagnosing issues, it has limitations. Traditional fault algorithms depend on historical data and do not account for the dynamic nature of systems over time. Historical data often overlooks the temporal dependencies that are vital for understanding system behavioral changes over time and identifying potential faults.
- Predicting faults involves anticipating fault types and their severity based on current and historical data, ranging from a few hours to weeks in advance. This approach enables us to detect faults before they occur, reducing downtime and optimizing maintenance by avoiding unnecessary inspections and concentrating only on the parts where faults are likely to happen. The papers reviewed however do not consider fault anticipating and prediction. Thus, it is crucial to include fault anticipation for modern PdM. Furthermore, to effectively address a component-level fault, it is essential to combine fault prediction with fault classification for guided maintenance pertinent to the specific fault category. Knowing fault types helps prioritize issues and optimize resources for improved PV system performance and longevity.
- Table 6 summarizes the detection and association of common faults presented in Fig. 13 to the respective AI algorithms. Overall, existing AI algorithms with different configurations can detect almost all the faults in the PV system's DC and AC sides. However, some faults such as Pelvic inflammatory disease (PID), back-sheet adhesion loss, intermittent faults, gradual and minor PV components degradation, and environmental influence (soiling, dust accumulation) appear to be challenging for AI algorithms to detect. This may be due to the detection process's difficulty in gathering high-quality and volume data, unpredictable fault nature and unrecognized cybersecurity issues affecting PV systems. SC, OC, PS, LL, bypass diode, and sensor faults are the most studied types because of their stronger signature and easy reproducibility in simulation/real tests. Various training parameters are used in AI algorithms for fault detection and classification such as irradiance, module and ambient temperature, AC and DC values (voltage, current), strings and system level voltages, humidity, IV characteristics of PV arrays, Inverter and grid currents and voltages, load demand data and frequency etc.
- Regarding AI algorithms, visible faults such as cell cracks, cell multi, soiling, vegetation, shadowing, hotspot, discolouration, and delamination are usually detected by DL algorithms. The other ones like OC, SC, PS, degradation faults, etc., are detected by supervised machine learning algorithms. DL algorithms such as CNN are effectively

Table 6
A summary of the most recent published work on PV system fault diagnosis.

Reference	ML methods	PV system, capacity and data type	Input/output setting	FD-AI accuracy	Fault severity levels	Estimated fault time	Real-time implementation	Future fault prediction
[130], 2023	Extreme Learning Machine, SVM, NN	GCPV system, 250 kW and Simulated data	**Input**: Minimum, maximum, average, and range values for currents, voltages, and powers. **Outputs**: Healthy and 3 faults: on-string, string to ground, string to string	Validation: 99.1663 %, Testing: 97.9727 %	No	No	No	No
[91], 2024	RF Classifiers	GCPV system, 9.54 KW and Simulated data	**Inputs**: weather (solar irradiance, module temperature), and PV output (I_{mpp}, V_{mpp}, P_{mpp}) parameters at MPP, **Outputs**: Healthy and 4 faults: Three SC and three shaded modules, OC, LL	99.4 %	No	No	No	No
[71], 2022	ANN	PV array, 5 kW and Simulated data	**Inputs**: P_{mpp}, OC voltage V_{oc}, SC current I_{sc}, V_{mpp}, and I_{mpp}, **Outputs**: Healthy and 7 faults: PS, LL, OC, degradation, bridge, bypass diode, hybrid fault	99.9 %	No	No	No	No
[131], 2023	Fuzzy logic system	PV array, -and Simulated data	**Inputs**: Extracted features, Fuzzy sets for each fault type (fuzzy sets). **Outputs**: Degree of membership for each fault type (membership)	96 %	No	No	No	No
[132], 2023	Adaptive neuro-fuzzy inference system (ANFIS)	PV module, 80 W and laboratory real data	**Inputs**: PV current and voltage characteristics, **Outputs**: Normal and 6 faults: inverter, feedback sensor, grid anomaly, MPPT controller, PV mismatch and boost converter controller fault	95.4 %	No	No	No	No
[88], 2023	Isolation forest and rule based algorithm	Solar power plant, and Real sensor data	**Inputs**: AC and DC power, irridiation. **Outputs**: Anomaly outliers: Normal and anomalous data	98.86 %	No	No	No	No
[111], 2023	AE NN algorithm	PV system, 9.54 KW and Real field data	**Inputs**: I_{mpp}, V_{mpp}, P_{mpp}, irradiance, T_{mod}. **Outputs**: Healthy system and 7 fault cases: different modules (1,234,510) SC, first string disconnected	99–100 %	No	No	No	No.
[123], 2022	CatBoost, LGBM, and XGBoost algorithms	PV arrays, 4.8 KW and Simulated data	**Inputs**: Current, voltage, power ratio, array yield, array efficiency and array capture loss. **Outputs**: Healthy and six fault cases: Intra and Inter string, OC, SC, PS, Degradation fault in string and array.	99.996 %	No	No	No	No.

(continued on next page)

Table 6 *(Continued)*

Reference	ML methods	PV system, capacity and data type	Input/output setting	FD-AI accuracy	Fault severity levels	Estimated fault time	Real-time implementation	Future fault prediction
[78], 2021	Supervised ML algorithms: DT, SVM, KNN, NB, Discriminant Analysis (DA) and RF.	GCPV system, 15 kW and Simulated data	**Inputs**: Squared Prediction Error (SPE), T^2 and Squared Weighted Error (SWE) statistics, first retained principal components. **Outputs**: Healthy and five fault classes: Inverter, grid connection, PV side sensor, PV PS and OC	Testing: 99.64 %, training: 99.87 %	No	No	No	No.
[133], 2022	EL and ML methods	PV array, 960 W and Real Laboratory experimental data	**Inputs**: I_{sc}, V_{oc}, I_{mp}, V_{mp}, P_{mp}, fill factor, I_1, I_2 from IV curves, **Outputs**: normal and five faults cases: PS with degraded PV modules, Dust accumulation (homogeneous dirt) with diode SC, PS (one PV module) with dust accumulation (homogeneous dirt), OC (one PV module disconnected) with PS, LL (between two PV modules) with degradation.	Catboost: 80 %, MLP (ANN): 80.66 %, EL stacking classifier: 81.73 %	No	No	No	No.
[102], 2023	CNN, LSTM, and Bi-LSTM networks	A 10 × 10 PV array test bed with a central inverter topology and Simulated data	**Inputs**: String voltages, system voltage and current, G, and T. **Outputs**: Normal and three fault cases: Module faults at string level.	Bi-LSTM: 99.94 %, 99.54 %, 99.98 %.	No	No	No	No.
[134], 2023	Distributionally robust logistic regression method	PV array- three PV strings, with four PV modules in series, 960 W and Real laboratory test data	**Inputs**: I_{mpp}, V_{mpp}, and P_{max}. **Outputs**: Normal and three fault causes: LL, OC, and PS.	≥ 98 %	No	No	No	No.
[107], 2022	Clustering algorithm and transfer LSTM	GCPV system, PV array with three series parallel column and Real plant data	**Inputs**: Historical data G, T, and humidity, PV power with 30 min time interval, prediction data of T, G, and humidity. **Ouputs**: Prediction of power generation and fault severity diagnosis: quantitative fault severity	–	Quantitative fault severity threshold	No	No	No.
[90], 2023	RF	PV arrays and Simulated data	**Inputs**: Module and ambient temperature, current, DC powers, irradiation, AC power, DC energies, and wind speed. **Outputs**: Two fault cases: DC power faults	≥ 98 %	No	No	No	No.

[135], 2020	EL method based on probabilistic strategy	PV array Simulink model and Simulated data	**Inputs**: fault features extracted from I–V curves. **Outputs**: Normal and LL fault cases: LL faults under low mismatch levels and high fault impedance.	99 % and 99.5 %	No	No	No	No.
[79], 2023	SVM	3 × 4 PV array and Simulated data	**Inputs**: I–V characteristic curves of PV arrays, SC current, OC voltage, maximum power current, and maximum power voltage. **Outputs**: three fault cases and normal condition: OC, SC, irradiation fault	Training: 98.2 % and testing accuracy: 97 %	No	No	No	No
[136], 2021	Domain adaptation and deep convolutional generative adversarial network, lightweight transfer CNN with adversarial data augmentation	Emulator-based grid-tied PV system, 1.5 kW and experimental lab data	**Inputs**: Real-time current data captured by CT **Outputs**: series arc fault detection	99.72 %	No	No	No	No
[72,73,108,109,137,138]- 2021–2024. **Remarks**: Papers by the same authors, implementing different ANN algorithms for the same system and faults	salp swarm algorithm (SSA) with Supervised ML, Bi-LSTM, GA based ANN, One class ML classifiers based on PCA under varying irradiance, Enhanced NN Method-Based Multiscale PCA, LSTM and BiLSTM	GCPV system, 3 PV array system 4 kW each and Simulated data	**Inputs**: Output current and voltage of PV_1 and PV_2, grid DC voltage, grid currents (phase a, b, c). **Outputs**: Healthy and 20 different faults: **PV array 1**: Bypass doide (BD_1), connectivity (Cf_1), LL_1, LG_1 **PV array 2**: (BD_2), (Cf_2), LL_2, LG_2. **Multiple faults**: $LL_1 + LG_1$, $LL_1 + BD_1$, $LL_2 + LG_2$, $LG_2 + Cf_2$. **Mixed faults**: $LL_1 + LL_2$, $LG_1 + LG_2$, $LL_1 + BD_1$, $LG_1 + Cf_1$, $LL_1 + BD_2 + LG_1$, $BD_1 + BD_2 + LG_2$, $BD_1 + BD_2 + LL_2$, $LL_1 + BD_1 + Cf_2 + LG_2$	SSA—DT & SVM: 99 %, Bi-LSTM: 100 %, GA: 88.48 %, One Class ML: 99 %, MSPCA-ANN: Train: 91.49 %, Test: 93.63 %, BiLSTM: Test and Train: 100 %.	No	No	No	No.

(continued on next page)

Table 6 *(Continued)*

Reference	ML methods	PV system, capacity and data type	Input/output setting	FD-AI accuracy	Fault severity levels	Estimated fault time	Real-time implementation	Future fault prediction
[139], 2024	Digital twin and shifted windows (swin) transformer optimized by particle swarm optimization	GCPV system 10 PV array, 49KW and Simulated data	**Inputs**: DTY and current ratios, **Outputs**: five different types of fault and PS conditions: LL, open-module, shorted-module, open-string, shorted-string, PS conditions.	Classification accuracy: 98.55 %	No	No	No	No.
[110], 2023	Bi-LSTM	GCPV System—3 PV array, 4 kW each and Simulated data.	**Inputs**: 11 different levels of solar irradiance, PV output voltage, DC bus voltage, grid currents. **Outputs**: Healthy and nine fault cases: failures in individual PV components, simple faults PV arrays, mixed faults in two arrays, and complex faults.	100 %	No	No	No	No
[140], 2023	Supervised ML with Data Dimensionality Reduction Strategy	GCPV system emulator and Simulated data	**Inputs**: Current and voltage at PV array, DC voltage, inverter phase currents, and voltages, estimated current and voltage magnitudes, estimated current and voltage frequency. **Outputs**: Normal and six fault cases: Inverter, feedback current sensor, grid anomaly, PV array mismatch with PS and OC, MPPT/IPPT controller, boost converter controller	KNN: 97 %	No	No	No	No.
[112], 2023	Deep AE based semi-supervised learning module followed by a hybrid SVM-LR	two stage GCPV system and Experimental lab data	**Inputs**: Historical data samples of $I_{pv}, V_{pv}, GandT$. **Output**: Healthy and six fault cases: LL fault in a string with mismatch level of 10 % and 80 %, OC fault, a cross-string LL, PS of one panel in a string, and LG fault with 30 % mismatch level.	99.67 %	No	No	No	No.
[74], 2023	LSTM, CNN and NN.	GCPV emulator system and Simulated data	**Inputs**: Squared prediction error, T^2. squared weighted error, and first retained principal components. **Outputs**: Healthy and five faulty cases: PV sensor fault, PV array level, three-phase inverter, grid external connection fault	LSTM: 95.12 %, ANN: 94.55 %, CNN: 61.24 %	No	No	No	No.

[129], 2023	Open source ML platform (edge impulse)	PV modules and Real data: IR images	**Inputs**: Database of 2000 infrared thermography image for PV modules. **Outputs**: Normal and 3 fault cases: Dirt, degradation, and dust/sand deposit on PV modules.	Mean accuracy: 93.4 %	No	No	Yes: Development of edge device (Nano 33 LBE sense).	No.
[141], 2023	Density ratio-based batch active learning fault diagnosis method integrated with adaptive Laplaciangraph trimming	PV array—Chroma 62150 H-1000 S solar array simulator and Experimental lab data	**Inputs**: **Outputs**: Normal and 7 fault cases: inverter, feedback sensor, grid anomaly, PV array mismatch with 10 to 20 % nonhomogeneous PS and 15 % OC in PV array, MPPT/IPPT controller, boost converter controller.	≥ 85 %	No	No	No	No.
[142], 2022	EL-based FDD paradigms	GCPV System and Simulated data	**Inputs**: PV current, three-phase currents, PV voltage, and three phase voltages, **Output**: Healthy and five faulty cases: Inverter (OC on one switch at a time), grid connection (switch to the standalone operation for protection reasons), output PV current sensor (poor connection and/or erroneous reading), PV panel (10 % to 20 % PS), PV panel connection (OC, SC, sudden disconnection).	100 %	No	Yes	No	No.
[103], 2022	CNN	PV panels and Simulated data	**Inputs**: PV string voltage and current, T and G level, fault label. **Outputs**: Normal and four faulty conditions: Degradation, LL, OC, PS.	Training accuracy: 97.64 %, Testing accuracy: 95.20 %	No	No	No	No.
[92], 2022	RF classifier	GCPV System and Simulated data	**Inputs**: V_{pv}, I_{pv}, V_{dc}, i_a, i_b, i_c, v_a, v_b, v_c, $\|I_{abc}\|$, $\|V_{abc}\|$, f_I, f_V, fault label. **Outputs**: Normal and six fault cases: Inverter, feedback sensor, grid anomaly, PV array mismatch based on PS and OC, controller, and boost converter.	100 %	No	No	No	No.

(continued on next page)

Table 6 *(Continued)*

Reference	ML methods	PV system, capacity and data type	Input/output setting	FD-AI accuracy	Fault severity levels	Estimated fault time	Real-time implementation	Future fault prediction
[143], 2022	Ensamble-based supervised and unsupervised ML	10 series connected PV modules, 2.2 kW & 32 modules, 4.16 kW—Simulated and lab experimental data	**Inputs**: G, total output power. **Outputs**: Normal and 3 cases of faults: Low and high percentage PV faults, faulty PV string.	Naive Bayes: 94 %–100 %	No	No	No	No.
[144], 2024	Radial basis function	PV array—PV modules connected in series with three parallel strings—Experimental lab data	**Inputs**: irradiation level and temperature, **Outputs**: Normal and eleven fault cases: PS, LL including intra and cross string, OC, and hybrid faults.	MSE for whole data: 0.110, for test data: 0.065	No	No	No	No
[80], 2024	SVM	N/A—Experimental lab data	**Inputs**: PV array output voltage and current, output power, irradiance, and temperature. **Outputs**: Normal and four fault cases: Random and fixed shading, and aging degradation	faults with known characteristics: 99.5 % and with unknown characteristics: 95.2 %.	No	No	No	No
[89], 2022	Isolation forest algorithm with Continuous Wavelet Transform and CNN	IEEE 34-bus distribution test feeder—3 solar PV installations, 200 kW each—N/A	**Inputs**: three-phase current, ground mode current and statistical features. **Ouptus**: Normal and 3 fault cases: Single-line-to-ground, LL, and three-phase faults.	classification accuracy: 96 %–99 %	No	No	No	No.
[145], 2022	EL approaches: boosting and bagging and Double Exponentially Weighted Moving Average chart.	PV plants, 9.54 kW and Real experimental data	**Inputs**: Ambient and cell temperature, tilt global irradiance, PV array DC current and voltage, maximum dynamic PV power, inverter AC current, AC power, grid AC voltage. **Outputs**: Normal and five anomaly conditions: PV string OC, circuit breaker, inverter disconnection, PS (pylons), 2 PV modules SC	≥ 90	No	No	No	No.
[81], 2022	CNN combine with SVM	PV modules and Real data: Dataset 1: 2624 defective and good EL images and Dataset 2: 1028 images of good and corroded/cracked.	**Inputs**: Electroluminescene images. **Outputs**: Good and defective, crooked/cracked	Classification accuracy for Dataset 1: 99.49 % and Dataset 2: 99.46 %	No	No	No	No

[146], 2022	Supervised ML algorithms	Small scale GCPV system-2 Parallel strings of six PV modules—Real data: data measurement time: 16 June 2020 to 16 Sep 2020	**Inputs**: five features: $\frac{I}{I_{exp}}$ (normalized current), V/V_{exp} (normalized voltage), $\frac{P}{P_{exp}}$ (normalized power), $\frac{V}{Voc_{ref}}$, and PV module condition under experimental test. **Outputs**: Normal and eight faulty conditions: Connector, PID, PS, pole shading and building shadow condition, SC bypass diode, soiling, glass breakage	NN: 93 %, RF: 92.5 %, Nearest Neighbor: 92.3 %, SVM: 91.7 %, DT: 89.8 V, Linear SVM: 89.2 %, LR: 76.6 %	No	No	No	No.
[124], 2023	LR with cross-validation	One diode model—Real T, GHI and simulated data of other features	**Inputs**: T, GHI, V_{mpp}, I_{mpp}, V_{oc}, I_{sc}, V_{cell}, I_{cell}, and operational status. **Ouputs**: Normal and four cases of faults: OC, SC, mismatch, and unidentified fault	97.11 %	No	No	No	No.
[147], 2022	Adaptive variational mode decomposition and deep minimum variance random vector functional link network	PV based DC microgrid and Simulated data	**Inputs**: Input $I_d c$ with AVMID-CSCFA algorithm and EWKU Index: significant modes features. **Outputs**: Normal and 8 classes of faults: Pole-to-pole, pole-to-ground, series arc, shunt cross-string arc, shunt intra-string arc, load-switching, PS, changes in PV irradiance	100 %	No	No	No	No.
[148], 2022	Multiple ML algorithms from SK-Learn library of python and a DL model	two PV strings of three panels, 1.8 kW and experimental lab data	**Inputs**: Current and voltage sensor output from the first and second string, T and G for different seasons—summer and winter. **Outputs**: Normal and faulty conditions	Overall classification accuracy: 99.2 %, F1-Score of ML models: Adaboost: 0.59, DT: 1, Navie Bye: 0.88, SVM: 0.994.	No	No	No	No.

(continued on next page)

Table 6 *(Continued)*

Reference	ML methods	PV system, capacity and data type	Input/output setting	FD-AI accuracy	Fault severity levels	Estimated fault time	Real-time implementation	Future fault prediction
[93], 2022	RF and modified independent component analysis.	GCPV system and Simulated data	**Inputs**: PV output current and voltage, DC bus voltage, grid phase currents, and voltages, I_f, V_f, target. **Outputs**: Normal and Four faults cases: Inverter, PS, voltage sag, and OC	99.88 % and 99.43 % for scenarios 1 (SMOTE) and 2 (Random under sampler)	No	No	No	No.
[86], 2022	DT	GCPV system, 4 kW and Laboratory data	**Inputs**: PCC current and voltage signals features from Wavelet transform. **Outputs**: Normal and four fault cases: Shade, inverter (AC and DC side), array (LG, LL, and OC), and panel (external and internal).	94.7 %	No	No	Yes—Field programmable gate array (FPGA).	No.
[149], 2023	PSO with back propagation NN	4*3 PV array with DC load and Simulated data	**Inputs**: V_{oc}, I_{sc}, P_m, V_m. **Outputs**: Normal and five faulty conditions: PS, aging cells, temperature and PS combined, temperature and cell aging combined.	95 %	No	No	No	No.
[75], 2022	ANN	Four PV parks, Greece, 99.84 kW and Real data—Jan. 2013 to Dec. 2016 at 15 min intervals.	**Inputs**: In-plane irradiance, panel back sheet temperature, ambient temperature.**Outputs**: AC power. Faults: String fault behavior, PV panel and shading fault.	Normalized root mean square error (nRMSE): Below 5 %	No	No	No	No
[150], 2024	Gradient Boosting techniques	6*6 TCT PV Array- Grid-connected PV system—7.2KW	**Inputs**: PCA based features: current, voltage and power ratio, array and total array yield. **Outputs**: Healthy and seven fault cases: OC, SC, PS, Array and string degradation, Inter and Intra fault.	**Training accuracy**: Catboost: 98.90 %, LGBM: 99.06 %, Adaboost: 97.01 %, **Model accuracy**: Catboost: 98.45, LGBM: 99.23 %, Adabost: 97.44 %.	No	No	No	No.
[76], 2023	ANN and Stacking Ensemble Machine Learning-based algorithm	PV array—three PV modules in parallel connection, each 60 W—Real Experimental lab test data	**Inputs**: First dataset: Solar irradiance, air and cell temperature, and PV output power. Second dataset: I–V curves. **Outputs**: Normal and five fault conditions: **Single fault**: PS and dust deposit. **Multiple faults**: OC and dust accumulation, PS and dust accumulation with shunted diode in shaded PV module.	Classification Accuracy: 96.8 %, ANN RMSE (W): 0.05	No	No	Yes—Embedded ML system ESP8266 microcontroller for real-time deployment.	No.

[104], 2023	CNN and fine-tuned model based on Visual Geometry Group (VGG-16)	Unit for Developing Solar Equipment and Real data—Thermal images	**Inputs**: Binary and multi-class classification of fragmented thermal images. **Ouputs**: Normal and five faulty cases: Bypass diode failure, partially covered PV module, shading effect, SC and dust deposit.	VGG-16 average accuracy: 99.91 % and fault diagnosis: 99.90 %.	No	No	No	No.
[151], 2023	Deep Stack-based Ensemble Learning approach	GCPV system and Simulated dataset	**Inputs**: G, T, OC voltage, SC current, form factor, maximum current, maximum voltage, maximum power, and boost converter output power. **Ouputs**: Normal and five fault cases: OC, SC, PS, bridge and degradation faults.	Fault detection accuracy without noise data: 98.62 % and with noisy data: 94.87 %.	No	No	No	No.
[82], 2023	XGboost, LGBM, LR, Support Vector Regression, Relevant Vector Machine	150 PV power stations, 54 MW and Real data—from 240 PV station projects with 120 sites.	**Inputs**: Solar irradiation, rated capacity, output power, voltage, and current of each inverter under MPPT. **Outputs**: Power prediction, alerts about different faults: Shadowing, inverter thermal degradation, fuse burnt, site outages, and other faults: equipment maintenance, broken module, inverter shut-down by miss operation, and so on.	Fault detection: Precision 99.2 %, failure mode classification: 92.3 %	No	No	No	No.
[83], 2023	Exemplar efficient model with Neighborhood Component Analysis and SVM	PV modules	**Inputs**: 20,000 infrared images—Real online data—Infrared solar modules datasets **Outputs**: Normal condition and 11 fault classes: Cell and cell-multi, hotspot and hotspot-multi, soiling, vegetation, shadowing, offline module.	Accuracy: 93 %, F1-score: 89.80 %, precision: 91.55 %, and sensitivity: 88.38 %	No	No	No	No.
[152], 2022	Wavelet analysis and Ensembled KNN ML classifier	6*6 Series parallel PV array and Simulated data	**Inputs**: Decompose of current signals through Wavelet Transform **Outputs**: Normal and three fault cases: SC, OC and PS faults.	Classification Accuracy: 96.29 %	No	No	No	No.

preferred for images and two-dimensional data, RNN-LSTM is for temporal dependencies-based fault detection and SML models are suitable for one-dimensional (1D) data.

6. Discussions and recommendations: challenges and solutions

The future of AI algorithms for PdM and fault diagnosis in PV systems appears promising, as they can greatly enhance efficiency and reliability. Some key challenges, solutions and recommendations for PV systems PdM and fault diagnosis are provided in this section as follows:

- **PdM integrated fault diagnosis**: A key unit for consideration is PdM, which involves monitoring and analyzing data from the entire system to predict potential faults. However, it may not always provide detailed information about specific fault types without additional diagnostic efforts. PdM integrated fault diagnosis enhances the understanding and root cause of equipment health deterioration. Fault diagnosis provides a systematic way to identify and categorize specific issues within the system. By adding this feature, PdM predicts potential failures and provides insights into their nature. This knowledge is instrumental in proactively addressing and mitigating the root causes, allowing for more targeted and effective maintenance strategies. Understanding fault types can inform decision-making, prioritize critical issues, and optimize resource allocation for better overall PV system performance and longevity.
- **Fault Severity, Pattern and Estimated Time**: A promising approach for ML in fault diagnosis and PdM of PV systems involves the integration of metrics like fault severity, patterns, frequency, and estimated time to critical levels. By utilizing these indicators, ML models can facilitate severity-based predictions and establish threshold-driven maintenance alerts, ensuring that maintenance is conducted proactively based on predictive data. These algorithms can anticipate the need for maintenance well ahead of serious issues by predicting faults and their categories anywhere from minutes to days in advance. Recognizing fault types, along with their severity, allows for the prioritization of critical issues. This ensures that immediate attention is given to high-severity faults, while routine maintenance can address less critical ones. By incorporating fault time, recurring patterns or trends can be identified, helping to distinguish between transient issues and long-term degradation. Not only does this holistic approach optimize resources, but it also enhances the performance and longevity of PV systems, facilitating targeted interventions and reducing downtime.
- **Weather Forecasting**: Conventional forecasting methods for weather parameters in PdM primarily utilize linear models neglecting the nonlinear factors influencing solar irradiance due to their unpredictable and time-dependent nature [29]. While linear statistical approaches are effective in analyzing historical data, particularly for short-term predictions, they struggle with nonlinearities and fluctuating weather conditions. Therefore, accounting for these irregular behaviors is crucial for developing an effective PdM model that can accurately estimate weather parameters to enhance the performance of PV systems.
- **Correlation model**: The cost of gathering data from the PV panels poses significant challenges in terms of data capturing, transmission as well as the expenses and maintenance (including false data) of measurement equipment (on-site sensors). To address the expenses associated with measurement equipment, utilizing nearby weather stations to predict climatic factors appears to be a viable solution. There is a possibility to develop a correlation model between weather parameters and PV power generation for regenerating PV system dynamical behavior, i.e., array voltage, current, temperature, and solar irradiation.
- **Consideration of various environmental conditions**: Deterministic fault detection models for PV systems can sometimes yield false predictions when environmental conditions are not considered. Non-faulty PV plants may show outputs similar to faulty ones under certain weather, leading to incorrect alarms. It's crucial to differentiate between actual faults and variations due to weather. Therefore, effective fault detection should incorporate insights from PV system performance across various weather conditions (like winter, cloudy, summer, and sunny) into the PdM and fault diagnosis algorithms.
- **Prediction horizon and window lengths**: The fault diagnosis model predicts fault type within a defined horizon, typically analyzing each fault configuration separately without considering transitions. By including transitional data—beginning with fault-free input and later incorporating instances of short or open circuit faults, the model can respond more swiftly to faults, enhancing its sensitivity and robustness.
- **Data's nature and temporal dependencies**: The quality of the data in PdM and fault diagnosis is critically important as the performance of AI algorithms is often limited to data representation [153]. Lack of attribute importance and data with redundant attributes generates poor ML performance. The predictions of AI algorithms become less reliable without consideration of temporal dependencies. Consequently, the AI algorithm can enhance fault detection accuracy by leveraging both specific attribute attention and the temporal relationships between historical and real-time data.
- **Data Standardization**: The lack of standardization in data types and formats, including those for synthetic data, poses several challenges. It complicates data integration and hinders the creation of reliable AI models, as preprocessing and feature engineering become too specific to each application. To effectively tackle the intricate challenges in standardizing PV data, immediate and focused industry-wide initiatives must be undertaken, utilizing frameworks like ISO 13374–1:2003 [154] and ISO 13373–1:2002 [155] for CM. Establishing uniform data formats and protocols is imperative, particularly for the integration of diverse data sources such as IV curves, load, and weather parameters. This will decisively resolve data integration issues. Standardized formats will not only streamline feature engineering but also ensure consistent preprocessing pipelines, thus eliminating bias and inconsistencies in AI model training. The creation of openly accessible benchmark datasets, complete with comprehensive metadata, is essential for enabling objective model comparisons and evaluations, while simultaneously reducing development costs by eliminating the need for customized preprocessing for each dataset. Future research must prioritize benchmarking studies across various data formats, develop automated data integration techniques, and validate synthetic data to guarantee robust AI training. This approach will foster a more consistent, reliable, and cost-effective strategy for PV system analysis.
- **Real-time monitoring and experimental hardware**: One approach to enhancing ML-based PdM applications for PV systems is by enabling real-time fault detection and diagnosis. Currently, only a few prototypes utilizing ML for fault diagnosis in real-time applications have been developed in laboratory settings, with no commercial devices available to date. Developing an experimental hardware setup for PdM and fault diagnosis using AI on an affordable chip is a potential future direction. ML algorithms are adept at analyzing real-time sensor data and identifying patterns that may indicate potential faults, thus facilitating maintenance and minimizing the risk of power outages. Consequently, the future aim is to build a system that can handle real-world disruptions caused by environmental factors and noise from measurement devices.
- **Economic Considerations in AI-driven PdM and Fault Diagnosis in PV Systems**: While AI-driven PdM and fault diagnosis methods have proven effective in identifying and forecasting faults in PV systems, their economic viability is still not well-studied. A majority of current research emphasizes the development of engineering and mathematical models aimed at improving fault detection accuracy and PdM capabilities. However, there is a lack of systematic studies evaluating the cost-effectiveness and return on investment (ROI)

of these methods. While large-scale PV plants can take advantage of economies of scale, making AI-driven PdM and fault diagnosis financially viable, small- and medium-sized PV facilities face a more intricate challenge regarding their financial feasibility. The considerable initial investment, necessary infrastructure, and high computational needs linked to implementing AI can create major obstacles for smaller PV system operators. When it comes to AI-driven PdM, it's essential to focus not only on technical precision but also on cost-effectiveness, especially for medium- and small-sized PV plants, where budget limitations may hinder adoption. A comprehensive economic evaluation framework is essential for analyzing the cost-benefit ratio of AI-based fault diagnosis in various sizes of PV systems, providing a balanced view of the feasibility of implementation. Moreover, future studies should connect the advancements in engineering-focused AI with their economic practicality, making sure that PdM solutions are both technically effective and financially sustainable for PV operators of all sizes. Additionally, a thorough evaluation should go beyond basic ROI calculations to include aspects like risk mitigation, enhanced long-term planning abilities, and potential environmental advantages. By bridging the gap between technological progress and economic viability, the full potential of AI in PV fault diagnosis can be achieved, facilitating its widespread and sustainable adoption across a range of PV system sizes.

- **Fault isolation and 5G communication**: Generally, fault detection and diagnosis algorithms focus solely on fault labelling and classification. The isolation of malfunctioning parts and the impact of isolating faulty components from the PV system on the overall system generation is however not considered. The underlying question is how a power shortage to demand will be delivered at the consumer end in case of isolating faulty components from the PV System. 5 G communication technology can be used for efficient and effective communication between multiple power system entities for demand response (application of networks to power systems).
- **Digital twin for PdM**: A digital twin is a virtual counterpart of a physical object, process, or system, operating as a real-time digital version. It is created by gathering extensive data from a system, allowing for safer, cost-efficient, and effective management of intricate processes. In the context of PV systems, digital twins can simulate components like PV arrays, modules, and inverters, utilizing sensor data to monitor performance, identify issues, and refine maintenance efforts. By considering factors such as fault severity, frequency, patterns, and advanced predictions for fault classification (ranging from minutes to days), digital twins improve PdM capabilities. These models can anticipate maintenance requirements, initiate prompt actions based on predictive insights, and support a more accurate and proactive maintenance approach. Consequently, fault detection, diagnosis, and forecasting can be carried out with greater precision, reducing downtime and prolonging the lifespan of PV systems. For example, [156] proposed a digital twin model for proactive maintenance of the manufacturing industry.

7. Conclusion

The paper presented a comprehensive review of PdM system architecture with a fault diagnosis perspective and associated data preferences. First, we presented an overview of PdM architecture, providing an excellent foundation for researchers and engineers interested in gaining insights into PdM architecture for PV systems. The recommendations for integrating PdM with fault diagnosis algorithms and future fault prediction will lead to the avoidance of future equipment failures in the PV system, increasing thereby the operational efficiency and resilience of the network. The following summaries emphasize the main conclusions and comparisons for a quick overview.

- *Critical evaluation of existing review articles:* Table 1 provides an overview of existing review papers that focus on the application of AI to PV system condition monitoring and fault diagnosis analysis, and compares these studies to our work, clearly highlighting our notable contributions to this area.
- *Maintenance techniques within PV system framework:* The advantages and disadvantages of each maintenance method for a PV system are outlined in Table 2, emphasizing the important characteristics of each technique related to operational effectiveness.
- *AI driven PdM framework for PV system components:* Table 3 illustrates the overview of designing a PdM framework for PV systems based on AI models, and maps the potential applications of specific AI algorithms to certain data types.
- *Critical summary of most recent work on PV system PdM:* Table 4 reports the most popular ML/DL algorithms used for PdM, condition monitoring, and anomaly detection for the PV system.
- *AI algorithm pros and cons for solar PV systems:* Table 5 outlines the advantages and disadvantages of using SVM, ANN, DT, RF, and gradient boosting methods in the development of PdM and fault diagnosis systems for PV systems.
- *Critical summary of most recent work on PV system fault diagnosis*: Table 6 reports the most popular ML/DL algorithms used for fault diagnosis in a PV system.

Furthermore, key contributions are summarized below:

- Our work focuses on the integration of PdM with fault diagnosis of PV systems. By adding this feature, PdM can predict potential failures and provide insights into their nature. This knowledge is instrumental in proactively addressing and mitigating root causes, allowing for more targeted and effective maintenance strategies. To the best of our knowledge, none of the previous literature on PdM has explored this idea.
- While numerous review papers have concentrated on fault diagnosis for PV systems, this work extends the literature by systematically incorporating fault severity based on time-dependent degradation patterns. By considering this concept, we establish severity-based predictions and threshold-driven maintenance alerts, ensuring that maintenance is conducted proactively based on predictive data. This approach optimizes maintenance scheduling. To the best of our knowledge, no previous review or technical paper has considered such an analysis.
- We also developed an innovative framework combining predictive maintenance with fault diagnosis, outlining how it will function. Based on this framework, we created Table 3, where we map which AI techniques will utilize specific data for individual components to execute PdM, taking into account fault diagnosis and other indicators such as fault severity and time duration. Additionally, our framework emphasizes the prediction of faults well in advance of their occurrence, a consideration that is typically absent in the existing fault diagnosis literature.
- We further discussed the nature of the data used for PdM and fault diagnosis, as well as the challenges it presents to AI performance. Additionally, fault diagnosis varies across industries, creating further complications, which raises the need for standardization across the entire fault diagnosis sector. This type of work has not been previously reported in the literature.
- We also shed light on the cost-effectiveness of AI-based fault diagnosis techniques for PV systems of different sizes. Large PV facilities often have the necessary resources and funding for such analyses, while medium and small PV systems tend to lag. We explore how to create a framework to support these smaller systems.

By considering all these points and aiming to enhance the entire PdM and fault diagnosis process, we aim to prolong the lifespan of PV systems. The integration of these concepts into a cohesive framework represents an innovative contribution to this field. Future research studies should concentrate on developing a digital twin for the PdM model, as well as on forecasting, fault detection and classification, conducting

real-time monitoring, and experimentally validating PdM and fault diagnosis algorithms in real-time scenarios. Furthermore, consider factors such as integrating weather forecasting for PdM, fault isolation, and the introduction of 5G communication between faulty PV systems and neighboring microgrids, these can enhance the precision of identifying potential failures in individual components within the PV system.

CRediT authorship contribution statement

Ali Hamza: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zunaib Ali:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Sandra Dudley:** Writing – review & editing, Supervision. **Komal Saleem:** Writing – review & editing, Visualization, Investigation, Formal analysis. **Muhammad Uneeb:** Visualization, Investigation, Formal analysis. **Nicholas Christofides:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] International Energy Agency. Net zero by 2050. OECD Publishing; 2021.
[2] Satya Bharath Kurukuru V, et al. A novel fault classification approach for photovoltaic systems. Energies 2020;13(2):308.
[3] Bataineh K, Eid N. A hybrid maximum power point tracking method for photovoltaic systems for dynamic weather conditions. Resources 2018;7(4):68.
[4] Dhibi K, et al. A hybrid approach for process monitoring: improving data-driven methodologies with dataset size reduction and interval-valued representation. IEEE Sens J 2020;20(17):10228–39.
[5] AlSkaif T, et al. A systematic analysis of meteorological variables for PV output power estimation. Renew Energy 2020;153:12–22.
[6] Zhao Y, et al. Hierarchical anomaly detection and multimodal classification in large-scale photovoltaic systems. IEEE Trans Sustain Energy 2018;10(3):1351–61.
[7] Karunathilake H, et al. Project deployment strategies for community renewable energy: a dynamic multi-period planning approach. Renew Energy 2020;152:237–58.
[8] Xiaoxia L, et al. Deep learning based module defect analysis for large-scale photovoltaic farms. IEEE Trans Energy Convers 2018;34(1):520–29.
[9] Jidong L, et al. How to make better use of intermittent and variable energy? A review of wind and photovoltaic power consumption in China. Renew Sustain Energy Rev 2021;137:110626.
[10] Kang M, et al. A hybrid feature selection scheme for reducing diagnostic performance deterioration caused by outliers in data-driven diagnostics. IEEE Trans Ind Electron 2016;63(5):3299–310.
[11] Mansouri M, Numan Nounou M, Numan Nounou H. Multiscale kernel PLS-based exponentially weighted-GLRT and its application to fault detection. IEEE Trans Emerg Top Comput Intel 2017;3(1):49–58.
[12] Badr MM, et al. Intelligent fault identification strategy of photovoltaic array based on ensemble self-training learning. Sol Energy 2023;249:122–38.
[13] Kumar A, Babu Chinnam R, Tseng F. An HMM and polynomial regression based approach for remaining useful life and health state estimation of cutting tools. Comput Ind Eng 2019;128:1008–14.
[14] Omotola Aremu O, et al. A relative entropy based feature selection framework for asset data in predictive maintenance. Comput Ind Eng 2020;145:106536.
[15] Zonta T, et al. Predictive maintenance in the industry 4.0: a systematic literature review. Comput Ind Eng 2020;150:106889.
[16] Iftikhar H, Sarquis E, Costa Branco PJ. Why can simple operation and maintenance (O&M) practices in large-scale grid-connected PV power plants play a key role in improving its energy output? Energies 2021;14(13):3798.
[17] Mansouri M, et al. Deep learning-based fault diagnosis of photovoltaic systems: a comprehensive review and enhancement prospects. IEEE Access 2021;9:126286–306.
[18] Navid Q, et al. Fault diagnostic methodologies for utility-scale photovoltaic power plants: a state of the art review. Sustainability 2021;13(4):1629.
[19] Mellit A, Marco Tina G, Kalogirou SA. Fault detection and diagnosis methods for photovoltaic systems: a review. Renew Sustain Energy Rev 2018;91:1–17.
[20] Sohani A, et al. Using machine learning in photovoltaics to create smarter and cleaner energy generation systems: a comprehensive review. J Clean Prod 2022;364:132701.
[21] Mellit A, Kalogirou S. Artificial intelligence and internet of things to improve efficacy of diagnosis and remote sensing of solar photovoltaic systems: challenges, recommendations and future directions. Renew Sustain Energy Rev 2021;143:110889.
[22] Baojie L, et al. Application of artificial neural networks to photovoltaic fault detection and diagnosis: a review. Renew Sustain Energy Rev 2021;138:110512.
[23] Venkatakrishnan GR, et al. Detection, location, and diagnosis of different faults in large solar PV system—a review. Int J Low Carbon Technol 2023;18:659–74.
[24] Et-Taleby A, et al. Applications of machine learning algorithms for photovoltaic fault detection: a review. Stat Optim Inf Comput 2023;11(1):168–77.
[25] Yuan Z, Xiong G, Xiaofan F. Artificial neural network for fault diagnosis of solar photovoltaic systems: a survey. Energies 2022;15(22):8693.
[26] Toche Tchio GM, et al. A comprehensive review of supervised learning algorithms for the diagnosis of photovoltaic systems, proposing a new approach using an ensemble learning algorithm. Appl Sci 2024;14(5):2072.
[27] Hong Y-Y, Pula RA. Methods of photovoltaic fault detection and classification: a review. Energy Rep 2022;8:5898–929.
[28] El-Banby GM, et al. Photovoltaic system fault detection techniques: a review. Neural Comput Appl 2023;35(35):24829–42.
[29] Ramirez-Vergara J, et al. Review of forecasting methods to support photovoltaic predictive maintenance. Clean Eng Technol 2022;8:100460.
[30] Berghout T, et al. Machine learning for photovoltaic systems condition monitoring: a review. In: IECON 2021–47th annual conference of the IEEE industrial electronics society; IEEE; 2021. p. 1–5.
[31] Osmani K, et al. A review on maintenance strategies for PV systems. Sci Total Environ 2020;746:141753.
[32] Hammoudi Y, et al. Review on maintenance of photovoltaic systems based on deep learning and internet of things. Indones J Electr Eng Comput Sci 2022;26(2):1060–72.
[33] Pal Singh U, Chandra S. A predictive maintenance scheme for solar PV system. In: Control applications in modern power systems: select proceedings of EPREC 2021; Springer; 2022. p. 189–95.
[34] Williams JH, Davies A, Drake PR. Condition-based maintenance and machine diagnostics. Springer Science & Business Media; 1994.
[35] Keith Mobley R. An introduction to predictive maintenance. Elsevier; 2002.
[36] Bernard Ang. Understand how data acquisition systems work. 2024. https://www.keysight.com/blogs/en/tech/educ/2024/data-acquisition-system [Online Accessed 07-March-2025].
[37] Huuhtanen T, Jung A. Predictive maintenance of photovoltaic panels via deep learning. In: 2018 IEEE data science workshop (DSW). IEEE; 2018. p. 66–70.
[38] Vicente-Gabriel J, et al. LSTM networks for overcoming the challenges associated with photovoltaic module maintenance in smart cities. Electronics 2021;10(1):78.
[39] Adam Zulfauzi I, et al. Anomaly detection using K-means and long-short term memory for predictive maintenance of large-scale solar (LSS) photovoltaic plant. Energy Rep 2023;9:154–58.
[40] Xue Q, et al. Fault prediction for solar array based on long short-term memory and autoencoder. In: 2021 40th Chinese control conference (CCC); IEEE; 2021. p. 4567–72.
[41] Waqar Akram M, et al. Automatic detection of photovoltaic module defects in infrared images with isolated and develop-model transfer deep learning. Sol Energy 2020;198:175–86.
[42] Emamian M, et al. Cloud computing and IoT based intelligent monitoring system for photovoltaic plants using machine learning techniques. Energies 2022;15(9):3014.
[43] Bagus Krishna Yoga Utama I, et al. Intelligent IoT platform for multiple PV plant monitoring. Sensors (Basel) 2023;23(15):6674.
[44] Bassam A, May Tzuc O. Predictive maintenance of photovoltaic system based on deep learning. 2022.
[45] Bommes L, et al. Anomaly detection in IR images of PV modules using supervised contrastive learning. Prog Photovoltaics Res Appl 2022;30(6):597–614.
[46] Ibrahim M, et al. Machine learning schemes for anomaly detection in solar power plants. Energies 2022;15(3):1082.
[47] Hamza A, et al. Optimizing PV array performance: a 2 LSTM for anomaly detection and predictive maintenance based on machine learning. In: 2024 IEEE energy conversion congress and exposition (ECCE); IEEE; 2024. p. 1681–88.
[48] Almonacid-Olleros G, et al. A new architecture based on IoT and machine learning paradigms in photovoltaic systems to nowcast output energy. Sensors (Basel) 2020;20(15):4224.
[49] De Benedetti M, et al. Anomaly detection and predictive maintenance for photovoltaic systems. Neurocomputing 2018;310:59–68.
[50] Sarquis Filho EA, Santos FC, Costa Branco PJ. Development of predictive maintenance algorithms for photovoltaic systems using synthetic datasets. In: 37th European photovoltaic solar energy conference and exhibition; 2020. p. 1584–89.
[51] Minhhuy L, et al. Remote anomaly detection and classification of solar photovoltaic modules based on deep neural network. Sustain Energy Technol Assess 2021;48:101545.
[52] Betti A, et al. Predictive maintenance in photovoltaic plants with a big data approach. arXiv preprint 2019. arXiv:1901.10855.
[53] Shapsough S, Zualkernan I, Dhaouadi R. Deep learning at the edge for operation and maintenance of large-scale solar farms. In: International conference on smart grid and internet of things; Springer; 2020. p. 27–44 .
[54] Hanif Jufri F, Seongmun O, Jung J. Development of photovoltaic abnormal condition detection system using combined regression and support vector machine. Energy 2019;176:457–67.
[55] Livera A, et al. Photovoltaic system health-state architecture for data-driven failure detection. Solar 2022;2(1):81–98. MDPI.

[56] Bin Mofidul R, et al. Predictive maintenance in photovoltaic systems using ensemble ML empirical analysis. In: 2023 Fourteenth international conference on ubiquitous and future networks (ICUFN); IEEE; 2023. p. 636–38.
[57] Vyas S, et al. Forecasting solar power generation on the basis of predictive and corrective maintenance activities. arXiv preprint 2022. arXiv:2205.08109.
[58] Sildnes Gedde-Dahl G. Optimising maintenance operations in photovoltaic solar plants using data analysis for predictive maintenance. MA thesis. Norwegian University of Life Sciences, Ås; 2022.
[59] Kumar SR, Gayathri RG, et al. Supervised machine leaning based anomaly detection and diagnosis in grid connected photovoltaic systems. In: Proceedings of the first international conference on combinatorial and optimization, ICCAP 2021, December 7-8, 2021; Chennai, India. 2021.
[60] Harrou F, et al. An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class support vector machine. Sol Energy 2019;179:48–58.
[61] Livera A, et al. Operation and maintenance decision support system for photovoltaic systems. IEEE Access 2022;10:42481–96.
[62] Liu Q, et al. Remaining useful life prediction of PV systems under dynamic environmental conditions. IEEE J Photovolt 2023.
[63] Arena E, et al. Anomaly detection in photovoltaic production factories via Monte Carlo pre-processed principal component analysis. Energies 2021;14(13):3951.
[64] Yao S, et al. Intelligent and data-driven fault detection of photovoltaic plants. Processes 2021;9(10):1711.
[65] Ait Abdelmoula I, et al. Towards a sustainable edge computing framework for condition monitoring in decentralized photovoltaic systems. Heliyon 2023;9(11).
[66] Kumari Sarita RS, Khan B. Reliability, availability, and condition monitoring of inverters of grid-connected solar photovoltaic systems. IET Renew Power Gener 2023;17(7):1635–53.
[67] Betti A, et al. Fault prediction and early-detection in large PV power plants based on self-organizing maps. Sensors (Basel) 2021;21(5):1687.
[68] Livera A, et al. Cloud-based platform for photovoltaic assets diagnosis and maintenance. Energies 2022;15(20):7760.
[69] Ashok V, Yadav A, Kumar Naik V. Fault detection and classification of multi-location and evolving faults in double-circuit transmission line using ANN. In: Soft computing in data analytics: proceedings of international conference on SCDA 2018; Springer; 2019. p. 307–17.
[70] Man Karmacharya I, Gokaraju R. Fault location in ungrounded photovoltaic system using wavelets and ANN. IEEE Trans Power Del 2017;33(2):549–59.
[71] Al-Katheri AA, et al. Application of artificial intelligence in PV fault detection. Sustainability 2022;14(21):13815.
[72] Hichri A, et al. Genetic-algorithm-based neural network for fault detection and diagnosis: application to grid-connected photovoltaic systems. Sustainability 2022;14(17):10518.
[73] Attouri K, et al. Enhanced neural network method-based multiscale PCA for fault diagnosis: application to grid-connected PV systems. Signals 2023;4(2):381–400.
[74] Hajji M. Deep learning based faults diagnosis in grid-connected photovoltaic systems. Technical Report, EasyChair. 2023.
[75] Roumpakias E, Stamatelos T. Health monitoring and fault detection in photovoltaic systems in Central Greece using artificial neural networks. Appl Sci 2022;12(23):12016.
[76] Mellit A, et al. An embedded system for remote monitoring and fault diagnosis of photovoltaic arrays using machine learning and the internet of things. Renew Energy 2023;208:399–408.
[77] Yin Z, Hou J. Recent advances on SVM based fault diagnosis and process monitoring in complicated industrial processes. Neurocomputing 2016;174:643–50.
[78] Hajji M, et al. Multivariate feature extraction based supervised machine learning for fault detection and diagnosis in photovoltaic systems. Eur J Control 2021;59:313–21.
[79] Wang J, et al. Fault diagnosis method of photovoltaic array based on support vector machine. Energy Sources Part A Recovery Util Environ Effects 2023;45(2):5380–95.
[80] Zhong Y, et al. Fault diagnosis of PV array based on time series and support vector machine. Int J Photoenergy 2024;2024.
[81] Et-Taleby A, et al. A combined convolutional neural network model and support vector machine technique for fault detection and classification based on electroluminescence images of photovoltaic modules. Sustain Energy Grids Netw 2022;32:100946.
[82] Chang M, et al. Developments of AI-assisted fault detection and failure mode diagnosis for operation and maintenance of photovoltaic power stations in Taiwan. In: 2023 IEEE/IAS 59th industrial and commercial power systems technical conference (I&CPS); IEEE; 2023. p. 1–6.
[83] Bala Duranay Z. Fault detection in solar energy systems: a deep learning approach. Electronics 2023;12(21):4397.
[84] Benkercha R, Moulahoum S. Fault detection and diagnosis based on C4. 5 decision tree algorithm for grid connected PV system. Sol Energy 2018;173:610–34.
[85] Lian Han H, Hong Ying M, Yang Y. Study on the test data fault mining technology based on decision tree. Procedia Comput Sci 2019;154:232–37.
[86] Alam M, et al. Condition monitoring and maintenance management with grid-connected renewable energy systems. Comput Mater Continua 2022;72(2):3999–4017.
[87] Zhang D, et al. A data-driven design for fault detection of wind turbines using random forests and XGBoost. IEEE Access 2018;6:21020–31.
[88] Kabir S, Shufian A, Md SRZ. Isolation forest based anomaly detection and fault localization for solar PV system. In: 2023 3rd international conference on robotics, electrical and signal processing techniques (ICREST); 2023. p. 341–45. https://doi.org/10.1109/ICREST57604.2023.10070033
[89] Paul S, et al. Knowledge-based fault diagnosis for a distribution system with high PV penetration. In: 2022 IEEE power & energy society innovative smart grid technologies conference (ISGT); IEEE; 2022. p. 1–5.
[90] Apoorva Bhat A, Koothenparambil Joy J. Fault detection in PV system using machine learning technique. 2023.
[91] Faris Amiri A, et al. Faults detection and diagnosis of PV systems based on machine learning approach using random forest classifier. Energy Convers Manag 2024;301:118076.
[92] Wali S, Khan I. Explainable signature-based machine learning approach for identification of faults in grid-connected photovoltaic systems. In: 2022 IEEE Texas power and energy conference (TPEC); IEEE; 2022. p. 1–6.
[93] Yang N-C, Ismail H. Robust intelligent learning algorithm using random forest and modified-independent component analysis for PV fault detection: in case of imbalanced data. IEEE Access 2022;10:41119–30.
[94] Chen T, Guestrin C. XgBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 785–94.
[95] Brownlee J. A gentle introduction to XGBoost for applied machine learning. 2021. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
[96] Veronika Dorogush A, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv preprint 2018. arXiv:1810.11363.
[97] Murat Çnar Z, et al. Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0. Sustainability 2020;12(19):8211.
[98] Alpaydin E. Machine learning. MIT press; 2021.
[99] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.
[100] Kayci B, Erdem Demir B, Demir F. Deep learning based fault detection and diagnosis in photovoltaic system using thermal images acquired by UAV. Politeknik Dergisi 2023:1.
[101] Sarquis Filho EA, Müller B, Costa Branco PJ. Understanding the consequences of switching to a predictive O&M strategy. 2025.
[102] Mustafa Z, et al. Fault identification for photovoltaic systems using a multi-output deep learning approach. Expert Syst Appl 2023;211:118551.
[103] Ali Memon S, et al. A machine-learning-based robust classification method for PV panel faults. Sensors (Basel) 2022;22(21):8515.
[104] Kellil N, Aissat A, Mellit A. Fault diagnosis of photovoltaic modules using deep neural networks and infrared images under Algerian climatic conditions. Energy 2023;263:125902.
[105] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[106] Dong T, et al. Aggregate distributed photovoltaic power joint prediction method based on LSTM. In: 2021 13th international conference on measuring technology and mechatronics automation (ICMTMA); IEEE; 2021. p. 155–58.
[107] Guo H, et al. A novel method for quantitative fault diagnosis of photovoltaic systems based on data-driven. Electr Power Syst Res 2022;210:108121.
[108] Hajji M, et al. Fault detection and diagnosis in grid-connected PV systems under irradiance variations. Energy Rep 2023;9:4005–17.
[109] Hichri A, et al. Fault diagnosis in a grid-connected photovoltaic systems based on hierarchical clustering. Technical Report, EasyChair, 2021.
[110] Jincan L, et al. Fault detection and diagnosis in grid-connected PV systems under irradiance variations. In: 2023 3rd international conference on energy, power and electrical engineering (EPEE); IEEE; 2023. p. 331–36.
[111] Seghiour A, et al. Deep learning method based on autoencoder neural network applied to faults detection and diagnosis of photovoltaic system. Simul Modell Pract Theory 2023;123:102704.
[112] Kumar U, Mishra S, Dash K. An IoT and semi-supervised learning-based sensorless technique for panel level solar photovoltaic array fault diagnosis. IEEE Trans Instrum Meas; 2023.
[113] Mellit A. An embedded solution for fault detection and diagnosis of photovoltaic modules using thermographic images and deep convolutional neural networks. Eng Appl Artif Intel 2022;116:105459.
[114] Dong M, et al. ISEE: industrial internet of things perception in solar cell detection based on edge computing. Int J Distrib Sens Netw 2021;17(11):15501477211050552.
[115] Arafat MY, Hossain MJ, Morshed Alam M. Machine learning scopes on microgrid predictive maintenance: potential frameworks, challenges, and prospects. Renew Sustain Energy Rev 2024;190:114088.
[116] Kyi S, Taparugssanagorn A. Wireless sensing for a solar power system. Digit Commun Netw 2020;6(1):51–57.
[117] Chang M, et al. PV O&M optimization by AI practice. In: Proceedings of the 36th European photovoltaic solar energy conference and exhibition; Marseille, French. 2019. p. 9–13.
[118] Zhafran Hussin M, et al. Anomaly detection of grid connected photovoltaic system based on degradation rate: a case study in Malaysia. Pertanika J Sci Technol 2021;29(4).
[119] Ledmaoui Y, et al. Forecasting solar energy production: a comparative study of machine learning algorithms. Energy Rep 2023;10:1004–12.
[120] Gamarra C, Guerrero JM, Montero E. A knowledge discovery in databases approach for industrial microgrid planning. Renew Sustain Energy Rev 2016;60:615–30.
[121] Kang Y, et al. Research on unified information model for big data analysis of power grid equipment monitoring. In: 2017 3rd IEEE international conference on computer and communications (ICCC); IEEE; 2017. p. 2334–37.
[122] Suciu G, et al. Big data processing for renewable energy telemetry using a decentralized cloud M2M system. Wirel Pers Commun 2016;87:1113–28.

[123] Adhya D, Chatterjee S, Kumar Chakraborty A. Performance assessment of selective machine learning techniques for improved PV array fault diagnosis. Sustain Energy Grids Netw 2022;29:100582.
[124] Voutsinas S, et al. Development of a machine-learning-based method for early fault detection in photovoltaic systems. J Eng Appl Sci 2023;70(1):27.
[125] Rai P, Londhe ND, Raj R. Fault classification in power system distribution network integrated with distributed generators using CNN. Electr Power Syst Res 2021;192:106914.
[126] Nazrul Islam Siddique M, et al. Fault classification and location of a PMU-equipped active distribution network using deep convolution neural network (CNN). Electr Power Syst Res 2024;229:110178.
[127] Waqar Akram M, et al. CNN based automatic detection of photovoltaic cell defects in electroluminescence images. Energy 2019;189:116319.
[128] Ramakrishna Madeti S, Singh SN. A comprehensive study on different types of faults and detection techniques for solar photovoltaic system. Sol Energy 2017;158:161–85.
[129] Mellit A, Blasuttigh N, Massi Pavan A. TinyML for fault diagnosis of photovoltaic modules using edge impulse platform. In: 2023 11th international conference on smart grid (icSmartGrid); IEEE; 2023. p. 01–05.
[130] Sabah Mutashar H, Mahmoud Shakir A. Robustness analysis of ELM-based fault detection in PV systems. Int J Smart grid 2025.
[131] Balakrishnan D, et al. An IoT-based system for fault detection and diagnosis in solar PV panels. In: E3S web of conferences; vol. 387. EDP Sciences; 2023. p. 05009.
[132] Mary P, Nasir Uddin M, Rezaei N. An adaptive neuro-fuzzy model-based algorithm for fault detection in PV systems. IEEE Trans Ind Appl; 2023.
[133] Mellit A, Kalogirou S. Assessment of machine learning and ensemble methods for fault diagnosis of photovoltaic systems. Renew Energy 2022;184:1074–90.
[134] Wang M, Xiaoyuan X, Yan Z. Online fault diagnosis of PV array considering label errors based on distributionally robust logistic regression. Renew Energy 2023;203:68–80.
[135] Eskandari A, Milimonfared J, Aghaei M. Line-line fault detection and classification for photovoltaic systems using ensemble learning model based on IV characteristics. Sol Energy 2020;211:354–65.
[136] Shibo L. Intelligent DC series arc fault detection using deep learning in photovoltaic systems. 2021.
[137] Hichri A, et al. Supervised machine learning-based salp swarm algorithm for fault diagnosis of photovoltaic systems. J Eng Appl Sci 2024;71(1):12.
[138] Yahyaoui Z, et al. One-class machine learning classifiers-based multivariate feature extraction for grid-connected PV systems monitoring under irradiance variations. Sustainability 2023;15(18):13758.
[139] Hong Y-Y, Pula RA. Diagnosis of photovoltaic faults using digital twin and PSO-optimized shifted window transformer. Appl Soft Comput 2024;150:111092.
[140] Chokr B, et al. Feature extraction-reduction and machine learning for fault diagnosis in PV panels. Sol Energy 2023;262:111918.
[141] Jiang X, et al. A novel density ratio-based batch active learning fault diagnosis method integrated with adaptive Laplacian graph trimming. Can J Chem Eng 2023;101(11):6471–81.
[142] Dhibi K, et al. Interval-valued reduced ensemble learning based fault detection and diagnosis techniques for uncertain grid-connected PV systems. IEEE Access 2022;10:47673–86.
[143] Hussain M, Al-Aqrabi H, Hill R. Statistical analysis and development of an ensemble-based machine learning model for photovoltaic fault detection. Energies 2022;15(15):5492.
[144] Reza Parsa H, Sarvi M. Online fault diagnosis, classification and localization in photovoltaic systems. IEEE Trans Instrum Meas 2024.
[145] Harrou F, et al. Ensemble learning techniques-based monitoring charts for fault detection in photovoltaic systems. Energies 2022;15(18):6716.
[146] Hojabri M, et al. IoT-based PV array fault detection and classification using embedded supervised learning methods. Energies 2022;15(6):2097.
[147] Kumar Jalli R, et al. Fault analysis of photovoltaic based DC microgrid using deep learning randomized neural network. Appl Soft Comput 2022;126:109314.
[148] Siva Prasad Machina V, Sriranga Suprabhath K, Madichetty S. Fault detection in solar photovoltaic systems during winter season-A deep learning approach. In: 2022 IEEE Texas power and energy conference (TPEC); IEEE; 2022. p. 1–6.
[149] Shaban Eldeghady G, Ahmed Kamal H, Moustafa Hassan MA. Fault diagnosis for PV system using a deep learning optimized via PSO heuristic combination technique. Electr Eng 2023;105(4):2287–301.
[150] Hamza A, et al. Enhancing solar farm operations: machine learning for equipment fault detection and classification. In: 2024 IEEE energy conversion congress and exposition (ECCE); IEEE; 2024. p. 1626–33.
[151] Lodhi E, et al. A novel deep stack-based ensemble learning approach for fault detection and classification in photovoltaic arrays. Remote Sens 2023;15(5):1277.
[152] Ramana Kumar Joga S, et al. Fault diagnosis in PV system using DWT and ensembled k-NN machine learning classifier. In: 2022 International conference on intelligent controller and computing for smart power (ICICCSP); IEEE; 2022. p. 1–5.
[153] Jiang Y, et al. A2-LSTM for predictive maintenance of industrial equipment based on machine learning. Comput Ind Eng 2022;172:108560.
[154] ISO. 13374-1: 2003 condition monitoring and diagnostics of machines—data processing. Communication and presentation—part 1; 2025.
[155] Monitoring C. Diagnostics of machines—vibration condition monitoring—part 1: general procedures. Standard ISO 2002:13373–1.
[156] Zhong D, et al. Overview of predictive maintenance based on digital twin technology. Heliyon 2023.