Comparative Study of Machine Learning for Fault Prediction in Solar Water Pumps

Yogesh Kakasaheb Shejwal, Jayraja U. Kidav, Lakshaman Korra, and Manjiri Lavadkar

Abstract Solar water pumps play a crucial role in sustainable agriculture, particularly in rural regions where reliable access to water and energy is limited. These devices, fueled by renewable solar energy, provide an environmentally sustainable and economical solution for irrigation. However, like any technological system, solar water pumps are susceptible to various faults, including electrical, physical, and environmental issues, which can compromise their efficiency and reliability. This research article examines the utilization of machine learning methodologies to forecast and identify malfunctions in solar water pump systems. This paper examines five machine learning models: Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, and Random Forest—are evaluated for their effectiveness in predicting three key fault types: electrical anomalies, including open circuits; physical faults, including module degradation and cleaning needs; and environmental faults, influenced by factors like wind velocity and temperature. This research study uses sensor data collected from solar water pump systems, including parameters such as current (I), voltage (V), temperature, and solar power. The results show that random forest and logistic regression always achieve the best accuracy in all fault parameters, which is suitable for fault prediction and preventive maintenance. By accurately predicting potential failures, the intelligent system aims to reduce downtime, lower operating costs, and increase the efficiency of solar water pumps. This research paper provides important insights into the optimization of machine learning models for fault prediction in renewable energy systems and advances the overarching objective of fostering sustainability agriculture through technology.

Keywords Solar powered · Solar water pump · Logistic regression · Machine learning · Intelligent monitoring · Maintenance · Power electronics · KNN · SVM

Y. K. Shejwal (✉) · J. U. Kidav · L. Korra · M. Lavadkar
National Institute of Electronics and Information Technology, Aurangabad, Maharashtra, India
e-mail: yogeshshejwal83@gmail.com

P. Shivakumara et al. (eds.), *Proceedings of the NIELIT's International Conference on Communication, Electronics and Digital Technologies*, Lecture Notes in Networks and Systems 1522, https://doi.org/10.1007/978-981-96-9932-2_17

1 Introduction

1.1 Introduction

Solar water pumps have revolutionized modern agriculture offering a sustainable and economical solution for irrigation in rural areas with limited access to reliable energy sources. By harnessing solar energy through photovoltaic panels, these pumps provide several advantages, such as financial savings, less ecological footprint, and energy autonomy. As climate change leads to unpredictable weather patterns, solar water pumps provide resilience to farmers, enabling year-round irrigation and improving food security. However, challenges remain, particularly in maintenance and fault diagnosis. By implementing predictive maintenance using machine learning, the reliability of these systems can be improved, ensuring optimal performance and longevity. The integration of solar water pumps into agricultural systems represents a crucial intersection between clean energy technology and sustainable farming, with the potential for long-term benefits for the environment and the economy. Solar water pumps are seen as a crucial tool in promoting the global shift toward renewable energy and supporting rural livelihoods [1, 2].

Solar water pumps have become crucial in agriculture, offering an environmentally friendly and economically viable solution for irrigation in areas with limited access to electricity or fuel. By harnessing solar energy through photovoltaic panels, these pumps draw water from various sources, reducing energy costs for farmers. The environmental and financial benefits of solar water pumps include cost reduction, environmental benefits by reducing carbon emissions, and increased productivity in agriculture [3].

However, these pumps face maintenance challenges that can lead to downtime and increased repair costs. The integration of machine learning for predictive fault detection offers a enhance resolution the reliability and longevity of these systems. By utilizing real time sensor data, machine learning models can accurately predict and prevent potential system failures, ensuring continuous operation and reducing maintenance costs. Research is ongoing to determine the most effective machine learning techniques for predicting faults in solar water pumps, aiming to enhance their role in promoting sustainable agriculture and renewable energy adoption globally [4, 5].

1.2 Problem Statement

Solar water pump systems play a vital role in supporting agriculture in remote areas, facing various operational challenges categorized as electrical, physical, and environmental issues. These challenges exert a considerable influence on the reliability and efficiency of the systems. Frequent electrical faults, such as open circuit faults, short circuits, and inverter failures, can reduce operational efficiency and disrupt power flow, affecting the overall pump performance. Physical failures, including solar panel degradation, pump component malfunctions, and loose electrical connections, are often a result of regular use and inadequate maintenance, leading to decreased performance, increased maintenance costs, and shortened system lifespan. Environmental factors like dust accumulation, wind speed, and extreme temperatures also contribute to system failures, diminishing efficiency and power output. It is essential to address these operational failures through regular cleaning, proper maintenance, and managing environmental conditions to ensure consistent power delivery, maintain pump performance, reduce maintenance costs, and extend system longevity. Efficient fault detection and prediction mechanisms, including machine learning methods for real-time fault prediction and maintenance, are crucial to improving the reliability of solar water pump systems in agriculture. By addressing these faults, the benefits of solar water pumps can be optimized, promoting their widespread use in sustainable agricultural practices [6, 7].

1.3 Objective

This study utilizes machine learning to predict faults in solar heat pumps, aiming to improve fault detection, efficiency, and minimize downtime. By analyzing real data, the research aims to develop accurate models for detecting various types of faults. Early detection can reduce maintenance time and prevent serious damage, while predictive error detection can optimize system performance. Implementing proactive maintenance strategies based on predictive analytics can reduce system downtimes, ensuring a reliable water supply for agricultural applications. Additionally, the study aims to support data-driven decision-making and promote sustainable agricultural practices by enhancing the reliability and efficiency of solar water pumps.

1.4 Scope

Solar water pump systems require quick identification and troubleshooting to ensure efficiency. This study focuses on three types of failures: electrical, physical, and environmental. Electrical faults are a major concern and can disrupt system operation. Common faults include open circuit, short circuit, and inverter failure. Machine learning methodologies such as support vector machine and random forest are used for fault detection based on historical data. Physical defects result from mechanical degradation over time, such as solar panel degradation, wear and tear of pump parts, and loose electrical connections. Machine learning techniques like k-nearest neighbors and decision trees analyze maintenance data and operating conditions detect physical faults. Environmental factors like dust, wind speed, and extreme temperatures can impact system performance. Machine learning techniques such as logical regression and artificial neural networks can model probability of faults occurring based on environmental variables, providing insight into optimizing system performance under changing conditions [10].

This study focuses on detecting faults in solar water pump systems using machine learning methodologies. By categorizing faults into electrical, physical, and environmental categories, the research aims to improve predictive maintenance and operational efficiency. Utilizing machine learning for fault prediction addresses challenges faced by solar water pumps, promoting their use in agriculture. This research highlights the importance of technological advancements in optimizing renewable energy systems for modern agriculture [11].

2 Literature Review

2.1 Defect Identification and Prediction in Solar Systems

Defect identification and prediction in solar energy systems are crucial for ensuring system reliability, efficiency, and minimizing operational costs. Traditional methods of fault detection, such as physical inspection and routine maintenance, are time-consuming and limited in providing real-time monitoring. Data-driven approaches, including statistical analysis and trend monitoring, have been utilized to identify anomalies in system performance. Machine learning techniques, such as random forests and support vector machines (SVM), have shown promise in accurately predicting faults in solar systems.

Predictive maintenance frameworks combining machine learning with predictive strategies are essential for improving system reliability. Hybrid models that integrate various machine learning algorithms have been explored to enhance accuracy of fault predictions. Quality of input data and feature selection play a significant role in the effectiveness of machine learning models for fault prediction. Challenges in this area include data quality and availability, as well as the real-time implementation of machine learning models in solar systems.

Future research directions could focus on integrating Internet of Things (IoT) sensors for continuous data streams and real-time analysis to improve the effectiveness of predictive maintenance strategies. Overall, fault detection and prediction in solar energy systems are essential for ensuring the reliability and efficiency of renewable energy technologies.

Sr. no	Focus area	Methodology	Key findings
1	Electrical and physical faults [12]	Analyzed fault types in solar water pumping systems	• Identified common electrical and physical faults, • emphasizing the need for real-time monitoring

(continued)

(continued)

Sr. no	Focus area	Methodology	Key findings
2	Electrical faults in PV systems [13]	Utilized Support Vector Machine (SVM) for fault classification	• SVM effectively classifies electrical fault types with high accuracy, highlighting the need for predictive maintenance
3	Degradation of solar panels [14]	Performance analysis under harsh environmental conditions	• Degradation of solar panels can reduce efficiency by up to 20%, underscoring the importance of regular maintenance
4	Predictive maintenance [15]	Review of machine learning applications	• Machine learning significantly enhances predictive maintenance capabilities, with data-driven approaches yielding improved fault detection
5	Fault detection techniques [16]	Combination of decision trees and neural networks	• Hybrid models improve prediction accuracy and robustness
6	Fault diagnosis solar water pumps [17]	Systematic review of machine learning applications	• Identified key machine learning techniques for fault diagnosis, emphasizing the importance of data quality
7	Environmental factors impact [18]	Analyzed the effects of environmental factors on solar panel efficiency	• Environmental conditions significantly affect solar panel performance, requiring robust monitoring strategies
8	Temperature effects on PV efficiency [19]	Reviewed global studies on temperature influence	• Temperature variations impact solar panel efficiency, with significant drops observed in extreme conditions

This presents different approaches used in fault analysis and prediction in solar systems. Each study provides a thorough understanding of various uncertainties and mechanisms and demonstrates the power of machine learning techniques increase reliability and efficiency of solar applications. Collectively, these findings emphasize the importance of continuous monitoring and predictive maintenance strategies to enhance the efficiency of solar power systems.

2.2 *Utilization of Machine Learning in Predictive Maintenance for Renewable Energy Technologies*

Sr. no	Technology	Machine learning techniques	Application details	Benefits	Challenges
1	Solar energy [20, 8]	Support vector machines (SVM)	Used to classify faults in solar panels and inverters based on operational data (voltage, current)	Enhances fault detection and enables timely maintenance	Data quality and availability can be an issue
2		Neural networks (ANNs)	Predicts performance degradation in solar panels by analyzing historical performance data	Increases efficiency by reducing downtime through proactive maintenance	Requires large datasets for effective training
3		Random forest	Identifies patterns of failures by processing large datasets with multiple features	Improves accuracy in fault prediction	Complexity in feature selection
4	Wind energy [21]	K-nearest neighbors (KNN)	Analyzes vibration and temperature data to predict failures in wind turbine components	Reduces downtime by up to 30% through timely interventions	Data collection and sensor placement can be challenging
5		Deep learning (LSTM)	Processes time-series data to forecast mechanical failures in turbine gearboxes	Captures sequential patterns for better predictive accuracy	Real-time data processing can be computationally intensive
6		Decision trees	Provides a visual representation of decision rules based on operational parameters	Simplifies fault diagnosis and decision-making processes	May not perform well with highly non-linear data
7	Hydropower [22]	Artificial Neural Networks (ANNs)	Monitors conditions of turbines and generators, predicting wear and tear through historical performance analysis	Enables proactive maintenance, minimizing operational disruptions	Complexity in model training and integration with existing systems
8		Regression analysis	Analyzes historical data to model the relationship between operational parameters and component lifespan	Supports decision-making for maintenance scheduling	Requires accurate historical data for effective modeling

(continued)

(continued)

Sr. no	Technology	Machine learning techniques	Application details	Benefits	Challenges
9		Ensemble learning	Integrates many forecasting models to enhance precision in fault detection for hydropower systems	Enhances robustness of predictions under varying operational conditions	May increase computational complexity

Machine learning is being applied to predictive maintenance for renewable energy technologies to improve system reliability and efficiency. Techniques used in solar, wind, and hydropower systems can detect faults early, reduce downtime, and support sustainable power systems. Challenges like data quality and processing requirements need to be addressed for machine learning to reach its full potential in this field. Advances in data collection, algorithm development, and system integration are key for future success in predictive maintenance for renewable energy technology.

2.3 *Existing Studies on Solar Water Pumps and Related Technologies*

Solar water pumps have gained a lot of gained prominence in recent years due to their capacity to offer sustainable and efficient water solutions for agriculture and rural areas. This combines research findings on different aspects of solar water pumps, including performance optimization, fault diagnosis, integration with IoT technologies, economic feasibility, and case studies.

Research focus	Key findings
Performance optimization [23]	• Investigated the effects of different solar panel configurations and pump types on efficiency • Conducted a comparative analysis of various solar water pumps, concluding that submersible pumps are generally more efficient in deep well applications • Emphasized the importance of selecting appropriate pump sizes and configurations to maximize energy efficiency
Reliability and fault detection [24, 9]	• Identified common electrical and physical faults, emphasizing the need for enhanced monitoring and maintenance strategies • Reviewed machine learning applications for predictive maintenance in solar water pumping systems • Developed a framework for monitoring faults in solar water pumps using IoT, enabling timely maintenance actions

(continued)

(continued)

Research focus	Key findings
Integration with IoT [25]	• Explored IoT-enabled solar water pumping systems for real time monitoring and performance analysis • Analyzed the execution of intelligent irrigation systems using solar pumps and IoT sensors for improved water management
Economic viability [26]	• Conducted a life cycle assessment of solar water pumping systems, revealing significant environmental benefits over diesel pumps • Provided a cost–benefit analysis comparing solar pumps and traditional pumps, demonstrating long-term savings with solar systems
Case studies [27]	• Documented the adoption of solar water pumps in rural India, discussing their socio-economic impacts on local communities • Presented case studies of solar pump installations in Africa, highlighting increased access to water and improved agricultural productivity • Examined the deployment of solar water pumps in Pakistan, reporting on the operational challenges and community benefits observed

Current research on solar water pumps shows their potential to provide sustainable and efficient solutions for irrigation and water supply. Research findings highlight the importance of optimizing system performance, increasing reliability through predictive maintenance, and using IoT technologies for real-time monitoring. The cost of solar water pumps, supported by case studies, shows their long-term benefits for agriculture and community development. Ongoing research and development in this domain is essential to address the problems and enhance the adoption of solar water pump technologies.

Focus area	Earlier work	Your work
Fault types	Mostly electrical faults	Electrical, physical, and environmental faults
ML techniques	Neural networks, SVM, random forest	Random forest, SVM, KNN, Decision Tree, logistic regression,
Accuracy metrics	Partial, not all models compared	Accuracy and F1-scores reported for all models
Advancement in prediction	Emphasis on individual models	Comparative analysis of multiple models

The research focuses on fault prediction in solar water pump systems using machine learning models like SVM, Random Forest, and Neural Networks. It analyzes three distinct fault types and compares their accuracy for specific categories. The study also introduces logistic regression for capturing linear relationships across fault parameters, highlighting its effectiveness. The research confirms previous studies' findings on Random Forest's strong performance and introduces logistic regression as a key performer.

3 System Overview

3.1 Introduction

Solar power generation has emerged as a key player in the renewable energy landscape, offering a clean and sustainable source of electricity. However, like any technological system, solar power generation systems are prone to various faults that can affect their efficiency and reliability. Understanding and predicting these faults is critical to optimizing power output and reducing system downtimes.

The major classes of faults in a solar power generation system can be divided as electrical, environmental, and physical. Open circuits and line faults are some common types of electrical faults that can disrupt the normal functioning of the system, thereby incurring losses. These types of faults are generally caused by damage, deterioration, or aging in the solar modules. Environmental factors, such as high wind velocities, can lead to structural failures or require preventive maintenance.

In this direction, by using the data coming from the solar power generation equipment, the proposed system will predict faults before they cause severe losses in power or even damage. The system focuses on the four significant faults: electrical faults (open circuit and line faults), physical faults (related to circuit maintenance and module cleaning), and environmental faults (due to structural maintenance caused by wind velocity). After considering the characteristics of temperature, current, voltage, power, sunlight, and wind velocity, the system under consideration aims to build a fault detection system that would complement the reliability of solar power generation systems.

3.2 Working of Proposed System

Figure 1 depicts the flowchart of the process for troubleshooting electrical faults in solar inverters. The first step of the process is data acquisition from sensor, followed by preprocessing of the acquired data and its usage in exploring the data along with feature selection in building the model.

1. If there are no electrical faults, then it goes to check if there is an environmental fault. If there were no faults in the environment, then it goes to check for physical faults.

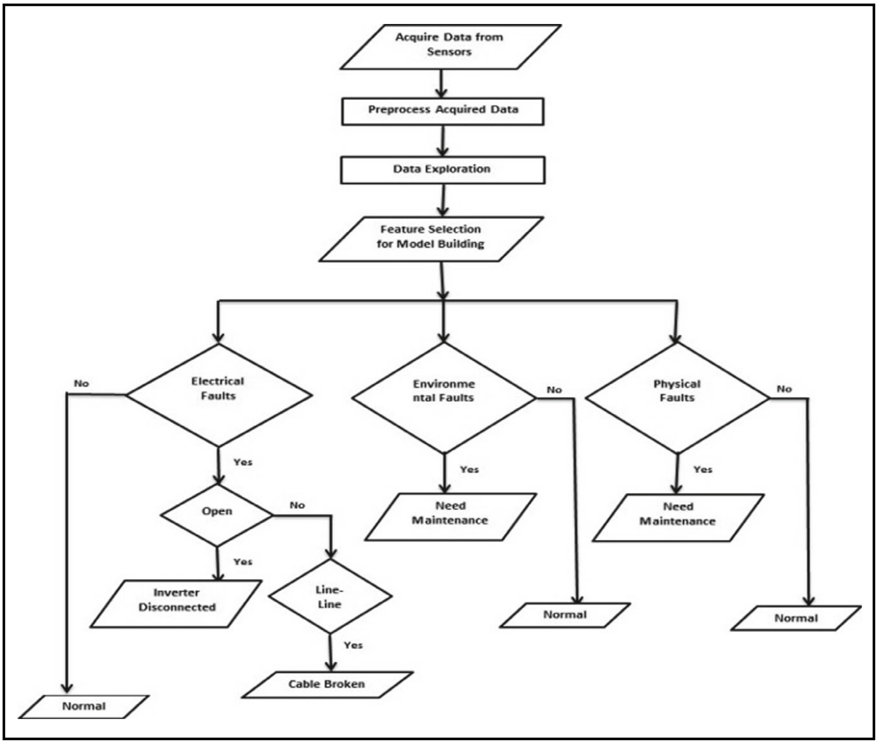

Fig. 1 Working flow of proposed system

2. If there are electrical faults, then it goes to check if the inverter is cut off. If it is cut off, then line is checked whether it is broken or not. If the line is broken, then cable is replaced.
3. If the line is not broken, replace the inverter.
4. If environmental faults are present, the inverter is checked for overheating. The inverter is then allowed to cool off.
5. If the inverter is not hot, heat sources in the environmental area are checked. After eliminating the sources of heat, the inverter is replaced.
6. If some physical faults are detected, then tests are carried out to determine if the inverter has signs of being physically damaged.
7. If there is existence of physical damage, then it calls for the replacement of the inverter. That is, in the absence of signs of physical damage, then the inverter is tested for connections. In this regard, if there exists signs of loose connections then connections are tightened.
8. In case there are no signs of loose connections, then it will call for the replacement of the inverter.

4 Methodology

The proposed fault prediction system for solar power generation uses machine learning algorithms to classifying and predicting electrical, physical, and environmental types of faults based on the input features collected from solar equipment. It predicts three major faults: electrical faults like open circuit and line faults; physical faults concerning the degradation or cleaning of modules; and environmental faults influenced by factors of wind velocity.

4.1 Collection of Data

This system collects data from attached sensors in the solar power generation equipment on the parameters that include temperature, current, voltage, output power, wind velocity, and intensity of sunlight. All these help in identifying the typical patterns that indicate various faults.

1. **Temperature**: Higher temperatures can affect solar panel performance and the efficiency of the system.
2. **Current**: Abnormalities in current flow can signal electrical issues such as open circuit or line-line faults.
3. **Voltage**: Voltage fluctuations are often indicative of faults in the solar panels or wiring.
4. **Sunlight Intensity**: Variations in sunlight affect the power output and may signal the need for cleaning or maintenance of the solar panels.
5. **Wind Velocity**: High wind velocity can lead to physical damage or structural issues, making it a crucial environmental factor to monitor.

These parameters provide valuable input data that reflect the operational health of the solar water pump system and are essential for detecting electrical, physical, and environmental faults.

4.2 Preprocessing Data

The data collected by sensors before its processing is done to ensure quality and reliability in the same. In this step, data is cleaned addressing missing values, normalization to ensure that the same scale is maintained for features and the data is separated for training and test sets. Because there are three different targets to be predicted, it's important to decide which feature is the relevant feature for a target of interest. Feature selection is also done to identify the most pertinent variables for the prediction of each type of fault. 3. In this study, several machine learning algorithms are trained for the fault prediction model. These include logistic regression, support vector machine

(SVM), k nearest neighbors (KNN), decision tree and random forest. Each model is chosen based on the ability of the given algorithm to handle varied data distributions or relationships between input features and output classes.

Before feeding the data into machine learning models, it must undergo preprocessing to ensure its quality and relevance. Key preprocessing steps include

1. **Handling Missing Values**: Sensor data can often be incomplete, with missing readings or irregularities. Techniques such as imputation (using the mean, median, or other statistical methods) are employed to fill in these gaps, ensuring no important information is lost.
2. **Normalization**: Since the different features (e.g., current, voltage, temperature) have varying scales, normalization is applied to bring them to a common scale. This step ensures that one feature does not disproportionately influence the model's predictions.
3. **Feature Selection**: Not all features may be relevant for every type of fault. For instance, electrical faults might rely more heavily on current and voltage, while environmental faults depend on wind velocity. Feature selection helps focus on the most important variables for each fault category, improving model performance and reducing noise in the data.
4. **Splitting the Dataset**: The dataset is partitioned into training and testing subsets, commonly in an 80:20 ratio. The training set instructs the model, whereas the test set assesses its performance on unfamiliar data.

4.3 Training the Models

Feeding the processed data into each of machine learning algorithms trains the models on respective relationships between input features and fault categories. To optimize the hyperparameters for the cross-validation of the models, overfitting is prevented, and the models will generalize well to new, unseen data. In training, all models are evaluated based on accuracy in the performance metrics on a training set and, therefore, aim at minimizing the errors of prediction.

4.4 Testing of the Models on Test Set

The performance of models is checked on test set after training. This is required to determine how accurately faults are predicted by the models. In this discussion, the accuracy of the models is used as the primary measure to compare the performances of various models developed above three prediction targets: electrical faults, physical faults, and environmental faults. The accuracy of the models for each of these targets decides what algorithm can perform best for that fault category.

Process	Details		
Training process			
Data preparation	• Preprocess fault data: normalize features handle missing values, encode categorical variables • Split dataset into training and test sets (e.g., 80/20 or 70/30)		
Model training	**Logistic regression**	• Fit the model to find the best-fit line minimizing logistic loss • A statistical model employing a logistic function to represent binary dependent variables. It assesses the likelihood of an event transpiring	• **Simplicity**: Easy implement and interpret • **Efficiency**: Works well with large datasets • **Linear Relationships**: Suitable for cases where the relationship between the dependent and independent variables is approximately linear
	Support vector machine (SVM)	• Identify the ideal hyperplane that optimizes the margin. • A supervised learning model that identifies the optimal hyperplane for distinguishing various classes inside the feature space	• **High Dimensionality**: Effective in high-dimensional spaces • **Flexibility**: Can use various kernel functions to represent non-linear relationships • **Robustness**: Works well with both linear and non-linear data distributions
	K nearest neighbors (KNN)	• Store training instances without a traditional training phase • A non-parametric algorithm that classifies data points based on the classes of their nearest neighbors in the feature space	• **Intuitive**: Easy to understand and implement • **Non-Linear Relationships**: Effective for data that does not fit a linear model • **Adaptability**: Can adapt to the distribution of the data without requiring a parametric form

(continued)

(continued)

Process	Details		
	Decision tree	• Iteratively partition data according to feature values to reduce impurity measurements • A hierarchical model that determines outcomes through a sequence of feature-oriented inquiries. It partitions the data into subsets according to feature values.	• **Interpretability**: Easy to visualize and interpret • **Non-Linearity**: Handles non-linear relationships well • **Feature Importance:** Provides insights into which features are most important for predictions
	Random forest	• Construct multiple trees using bootstrapped samples; aggregate predictions • An ensemble learning technique that integrates numerous decision trees to enhance predictive accuracy and mitigate overfitting.	• **Robustness**: Reduces the risk of overfitting compared to single decision trees • **Handling Complexity**: Effectively captures complex interactions between features • **Versatility**: Applicable for both classification and regression tasks
Cross-validation	• Use k-fold cross-validation to minimize overfitting and validate model efficacy • Divide dataset into k subsets; train and validate k times using different folds		
Hyperparameter optimization	• Combine cross-validation with hyperparameter tuning for robust validation • Employ methodologies such as Grid Search or Random Search to identify optimal hyperparameter values.		
Testing process			
Evaluation of performance	• Test each model on test set (unseen data) to assess predictive performance • Make predictions and compare them to actual outcomes		
Accuracy calculation	• Calculate accuracy as $\text{Accuracy} = \frac{\text{True Positives}+\text{True Negatives}}{\text{Total Instances}}$ • High accuracy indicates good performance in identifying faults. Consider other metrics (precision, recall) for comprehensive evaluation.		

4.5 *Fault Prediction*

Once the best model is found for each type of fault, which can predict whether a fault is likely going to happen in the system, the final model is deployed to make real time predictions based on live sensor data from the solar power generation system. The model predicts whether a fault is likely to occur, allowing for proactive maintenance and system optimization.

5 Results and Discussion

The machine learning models used for predicting faults in solar power generation systems demonstrate varying levels of accuracy across different fault types.

Tables 1, 2, and 3 show accuracy for each model.

Table 1 Accuracy for target 1

Algorithm	Accuracy	F1_score
Logistic regression	95.2	80.32
SVM	93.6	77.47
KNN	80.1	45.69
Decision tree	92.2	76.33
Random forest	94.2	79.2

Table 2 Accuracy for target 2

Algorithm	Accuracy	F1_SCore
Logistic regression	96.2	81.1
SVM	91.6	75.5
KNN	82.4	47.22
Decision tree	93.5	77.23
Random forest	94.2	79.3

Table 3 Accuracy for target 3

Algorithm	Accuracy	F1_score
Logistic regression	95.9	80.85
SVM	92.6	76.33
KNN	88.4	60.3
Decision tree	94.5	79.8
Random forest	95.5	80.2

For **Target 1**, which focuses on electrical faults like open circuit and line-line faults,

1. **Logistic Regression** achieves the highest accuracy at 95.2% and F1 Score is 80.32%. This indicates that the relationship between features such as temperature, current, voltage, and power is largely linear, making logistic regression an ideal fit.
2. **Random Forest**, with an accuracy of 94.2% and F1 Score is 77.47%, also performs well owing to its capacity to manage intricate complex patterns by aggregating multiple decision trees.
3. **SVM** follows closely at 93.6% and F1 Score is 45.69%, though it's slightly lower performance suggests that it may struggle slightly with the linearity of the data.
4. **Decision Tree** offers decent accuracy at 92.2% and F1 Score is 76.33%, though it is prone to overfitting.
5. **KNN** lags behind with an accuracy of 80.1% and F1 Score is 79.2%, likely due to its sensitivity to noisy data, making it less suitable for this fault type.

For **Target 2**, which involves physical faults such as maintenance and cleaning requirements,

1. **Logistic Regression** again proves to be the most effective model with an accuracy of 96.2% and F1 Score is 81.1%, indicating that physical faults are also well captured by a linear model.
2. **Random Forest** performs robustly here as well with 94.2% and F1 Score is 75.5%, offering consistency across fault types,
3. **Decision Tree** improves slightly with an accuracy of 93.5% and F1 Score is 47.22%, likely due to its ability to capture more intricate relationships between the features.
4. **SVM** scores slightly lower at 91.6% and F1 Score is 77.23%, reflecting its sensitivity to non-linearities or noise in the data,
5. **KNN** improves marginally to 82.4% and F1 Score is 79.3%, though it remains less effective for physical fault prediction.

For **Target 3**, which focuses on environmental faults like structural maintenance due to wind velocity.

1. **Logistic Regression** continues to perform strongly with an accuracy of 95.9% and F1 Score is 80.85%, demonstrating that environmental factors also exhibit a linear relationship with fault occurrences.
2. **Random Forest** excels with an accuracy of 95.5% and F1 Score is 76.33%, showcasing its strength in handling more complex decision boundaries in the data.
3. **Decision Tree** performs similarly well, with an accuracy of 94.5% and F1 Score is 60.30%
4. **SVM** achieves 92.6% and F1 Score is 79.8%, reflecting its effectiveness in handling some non-linearities present in environmental fault data.,

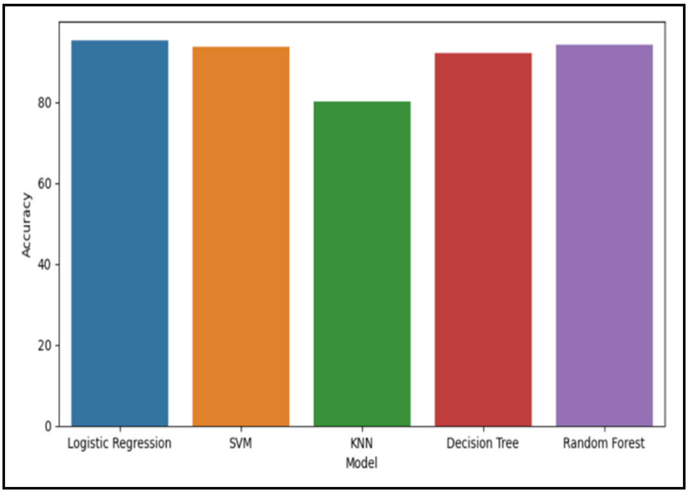

Fig. 2 Accuracy for target 1

5. **KNN** performs better for this target with an accuracy of 88.4% and F1 Score is 80.20%, likely because environmental data like wind velocity forms clearer clusters that KNN can more easily capture.

Overall, **Logistic Regression** and **Random Forest** are the top performers across all three fault types, with **Logistic Regression** benefiting from the largely linear relationships in the data and **Random Forest** excelling due to its ability to generalize across complex patterns while avoiding overfitting. **SVM** performs well but tends to lag slightly, particularly in cases of linearity. **Decision Tree** shows good accuracy but is more prone to overfitting compared to **Random Forest**, and **KNN**, while generally weaker, performs better when dealing with environmental fault data where clearer clustering is present (Figs. 2, 3 and 4).

The study shows that logistic regression outperforms other methods in identifying fault categories, with a maximum accuracy of 95.2% in electrical failures, 96.2% in physical imperfections, and 95.9% in environmental defects. Random Forest's accuracy of 95.5% is also noteworthy. The study suggests incorporating high-performing models for real-time application and averting system disruptions.

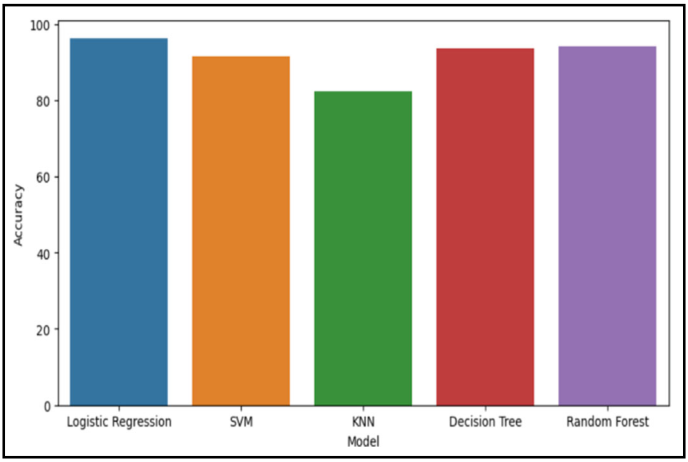

Fig. 3 Accuracy for target 2

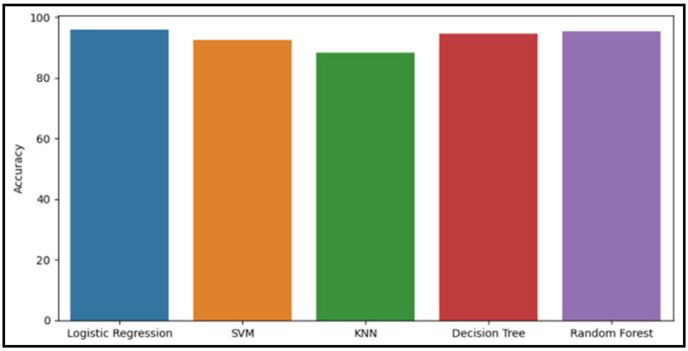

Fig. 4 Accuracy for target 3

6 Conclusion

The integration of Machine learning techniques employed for failure prediction in solar power systems. Generation system proves to be highly effective for improving reliability and system performance. Data collected based on the key parameters, which includes current, voltage, and temperature, sunlight intensity, and wind velocity, results in an accurate electrical, physical, and environmental faults prediction. Among the accuracy of the tested machine learning algorithms, Logistic Regression is one that generally tends to high so it performs well across all the faults, indicating the degree to which many of the underlying relationships are linear in nature. In addition to the capability of capturing complex patterns and handling non-linear data, Random Forest stands out well for application in real-world applications. The proposed system intended for real-time fault prediction for the early detection of issues that actually minimize downtime, enabling proactive maintenance. This not only improves efficiency of the entire solar electricity production systems but reduces costs involved in the operation due to prevention from failures before the situation escalates.

References

1. Ministry of New & Renewable Energy–Government of India. mnre.gov.in. https://mnre.gov.in/
2. Solar Energy Corporation of India Limited(SECI) A government of India enterprise, under ministry of new and renewable energy. www.seci.co.in. https://www.seci.co.in/
3. Shejwal YK, Kidav JU, Korra L (2024) Maintenance and performance of solar water pump in a rural agricultural community: a case study. In: Lecture notes in networks and systems. pp 245–254. https://doi.org/10.1007/978-981-97-3604-1_18
4. Stellbogen D (2002) Use of PV circuit simulation for fault detection in PV array fields. https://doi.org/10.1109/pvsc.1993.346931
5. Chandel SS, Nagaraju Naik M, Chandel R (2015) Review of solar photovoltaic water pumping system technology for irrigation and community drinking water supplies. Renew Sustain Energy Rev 49:1084–1099. https://doi.org/10.1016/j.rser.2015.04.083
6. Hassan YB, Orabi M, Gaafar MA (2023) Failures causes analysis of grid-tie photovoltaic inverters based on faults signatures analysis (FCA-B-FSA). Sol Energy 262:111831. https://doi.org/10.1016/j.solener.2023.111831
7. Impact of electromagnetic field on the conversion efficiency of solar PV panel. Periodico 91(4). https://doi.org/10.37896/pd91.4/91412
8. Yang X, Sun L, Yuan Y, Zhao X, Cao X (2018) Experimental investigation on performance comparison of PV/T-PCM system and PV/T system. Renew Energy 119:152–159. https://doi.org/10.1016/j.renene.2017.11.094
9. Ahmad T et al (2021) Artificial intelligence in sustainable energy industry: Status Quo, challenges and opportunities. J Clean Prod 289(289):125834. https://doi.org/10.1016/j.jclepro.2021.125834
10. Lu S, Phung BT, Zhang D (2018) A comprehensive review on DC arc faults and their diagnosis methods in photovoltaic systems. Renew Sustain Energy Rev 89:88–98. https://doi.org/10.1016/j.rser.2018.03.010
11. Zaki SA, Zhu H, Fakih MA, Sayed AR, Yao J (2021) Deep-learning–based method for faults classification of PV system. IET Renew Power Gener 15(1):193–205. https://doi.org/10.1049/rpg2.12016
12. Waqar Akram M, Li G, Jin Y, Chen X (2022) failures of photovoltaic modules and their detection: a review. Appl Energy 313:118822. https://doi.org/10.1016/j.apenergy.2022.118822
13. Kuo CL, Chen J-L, Chen S-J, Kao C-C, Yau H-T, Lin C-H (2017) Photovoltaic energy conversion system fault detection using fractional-order color relation classifier in microdistribution systems. IEEE Trans Smart Grid 8(3):1163–1172. https://doi.org/10.1109/tsg.2015.2478855
14. Sayyah A, Horenstein MN, Mazumder MK (2014) Energy yield loss caused by dust deposition on photovoltaic panels. Sol Energy 107:576–604. https://doi.org/10.1016/j.solener.2014.05.030
15. Feng Y, Hao W, Li H, Cui N, Gong D, Gao L (2020) Machine learning models to quantify and map daily global solar radiation and photovoltaic power. Renew Sustain Energy Rev 118:109393. https://doi.org/10.1016/j.rser.2019.109393
16. Haque A, Bharath KVS, Khan MA, Khan I, Jaffery ZA (2019) Fault diagnosis of photovoltaic modules. Energy Sci Eng. https://doi.org/10.1002/ese3.255

17. Zhao Y, Li T, Zhang X, Zhang C (2019) Artificial intelligence-based fault detection and diagnosis methods for building energy systems: advantages, challenges and the future. Renew Sustain Energy Rev 109:85–101. https://doi.org/10.1016/j.rser.2019.04.021
18. Blumer LS (2020) Solar panel electricity, efficiency, and environmental impacts. Adv Biol Lab Educ. https://doi.org/10.37590/able.v41.art4
19. Popovici CG, Hudişteanu SV, Mateescu TD, Cherecheş N-C (2016) Efficiency improvement of photovoltaic panels by using air cooled heat sinks. Energy Procedia 85:425–432. https://doi.org/10.1016/j.egypro.2015.12.223
20. Flicker J, Johnson J (2016) Photovoltaic ground fault detection recommendations for array safety and operation. Sol Energy 140:34–50. https://doi.org/10.1016/j.solener.2016.10.017
21. Hsu J-Y, Wang Y-F, Lin K-C, Chen M-Y, Hsu JH-Y (2020) Wind turbine fault diagnosis and predictive maintenance through statistical process control and machine learning. IEEE Access 8:23427–23439. https://doi.org/10.1109/access.2020.2968615
22. Bernardes J, et al (2022) Hydropower operation optimization using machine learning: a systematic review. AI 3(1):78–99. https://doi.org/10.3390/ai3010006
23. Khurana D, Dhingra S (2019) A comparative evaluation on different types of recommender systems. J Adv Res Dynam Control Syst 11(11):97–105. https://doi.org/10.5373/jardcs/v11i11/20193173
24. Ahmad MW, Reynolds J, Rezgui Y (2018) Predictive modelling for solar thermal energy systems: a comparison of support vector regression, random forest, extra trees and regression trees. J Clean Prod 203:810–821. https://doi.org/10.1016/j.jclepro.2018.08.207
25. IoT and neural network-based water pumping control system for smart irrigation. Inform Sci Lett 9(2):107–112. https://doi.org/10.18576/isl/090207
26. Al-Karaghouli A, Kazmerski LL (2013) Energy consumption and water production cost of conventional and renewable-energy-powered desalination processes. Renew Sustain Energy Rev 24:343–356. https://doi.org/10.1016/j.rser.2012.12.064
27. Bhave AG (1994) Potential for solar water-pumping systems in India. Appl Energy 48(3):197–200. https://doi.org/10.1016/0306-2619(94)90008-6