MULTIDISCIPLINARY
RESEARCH INSTITUTE
INDIA

International Journal of Advanced Scientific Research and Engineering Trends

ISSN: 2456-0774

Volume 10 Issue 1s (2026) | Pages: 85-91

Power System Fault Detection Using Explainable Artificial Intelligence Techniques

Akshay R. Bhosale[1], D. S. Bankar[2]

[1]BV(DU), College of Engineering, Pune
[2]Department of Civil Engineering, BV(DU), College of Engineering, Pune

Peer Review Information

Type: *Article*
Received: *30 March 2026*
Revised: *25 April 2026*
Accepted: *03 June 2026*
Published: *22 June 2026*

Abstract

The growing complexity of modern electrical power networks has increased the need for dependable and interpretable fault detection mechanisms. Conventional protection approaches, while effective in structured grids, often struggle to adapt to systems influenced by distributed generation and renewable energy sources. This study presents an interpretable machine learning–based framework designed to improve fault identification while maintaining decision transparency. Time-domain electrical parameters, including voltage and current measurements, are analyzed under diverse operating conditions to distinguish between normal and faulty states. An explainability layer is incorporated to clarify the reasoning behind model predictions, enabling engineers to better understand system behavior during disturbances. Simulation results indicate that the developed approach achieves high detection performance while offering meaningful insights into feature contributions. The framework therefore supports informed decision-making and demonstrates strong potential for application in intelligent power system protection environments.

Keywords: Fault Identification; Interpretable Machine Learning; Power System Protection; Explainable Models; Smart Grids; Simulation-Based Validation

How to Cite This Article

Bhosale, A. R., & Bankar, D. S. (2026). Power system fault detection using explainable artificial intelligence techniques. *International Journal of Advanced Scientific Research and Engineering Trends*, 10(1s), 85–91.

Introduction

Reliable fault identification is essential for ensuring the stability, safety, and continuity of electrical power delivery. Undetected faults can escalate into equipment damage, voltage instability, and large-scale outages, making rapid detection a fundamental requirement for modern grid operation. Traditionally, protection strategies have relied on predefined thresholds, impedance-based calculations, and coordinated relay actions. While these techniques have demonstrated robustness in conventional networks, the evolving structure of power systems presents new operational challenges.

The increasing penetration of renewable energy resources, distributed generation, electric vehicles, and dynamically varying loads has introduced significant uncertainty into grid behavior. These factors contribute to nonlinear operating conditions in which classical rule-based detection methods may struggle to differentiate between transient disturbances and genuine fault events.

To address these limitations, researchers have increasingly explored machine learning approaches capable of recognizing complex patterns in electrical signals. Such models offer adaptability and improved classification capability; however, their opaque decision-making processes often raise concerns in safety-critical environments. Protection engineers must be able to justify automated decisions before integrating them into operational workflows. Interpretability has therefore emerged as a critical requirement in intelligent protection systems. Explainable modeling techniques enable practitioners to examine the factors influencing predictions, thereby supporting validation and improving trust in algorithm-assisted diagnostics.

Motivated by these needs, this work proposes an interpretable machine learning framework for power system fault identification. The approach combines supervised learning with feature attribution methods to deliver accurate predictions alongside human-understandable explanations. By aligning data-driven insights with established electrical engineering principles, the proposed framework aims to enhance situational awareness while preserving the reliability expectations of modern protection schemes.

Background and Related Work

Fault detection in electrical power systems has traditionally been achieved through protection mechanisms such as overcurrent relays, distance protection, and differential schemes. These methods rely on predefined electrical thresholds and carefully coordinated system parameters to identify abnormal operating conditions. In structured grid environments, such approaches have demonstrated high reliability and operational simplicity.

However, modern power networks are no longer characterized by predictable power flows. The integration of renewable generation, bidirectional energy transfer, and flexible loads has increased operational variability. Consequently, adaptive protection strategies have been investigated to accommodate changing network configurations. Although effective, these solutions often require extensive system modeling and periodic parameter updates, which may limit scalability.

Machine learning techniques have gained attention as an alternative due to their ability to learn directly from observed system behavior. Supervised models trained on labeled electrical data have shown strong performance in distinguishing between fault categories. Deep learning architectures, in particular, can extract subtle patterns from waveform data that may not be apparent through conventional analysis.

Despite these advantages, a persistent challenge remains: many high-performing models operate as black boxes. In safety-critical infrastructure, decisions that cannot be interpreted are difficult to validate and may hinder practical deployment.

To overcome this limitation, recent research has emphasized interpretable modeling approaches. Explainability methods provide insight into how input features influence predictions, allowing domain experts to evaluate whether model behavior aligns with established electrical principles. Within power system applications, such transparency supports diagnostic reasoning and promotes confidence in automated fault identification.

Proposed XAI-Based Fault Detection Framework

System Overview

The developed fault detection framework follows a structured architecture designed to support both analytical accuracy and operational interpretability. Rather than replacing conventional protection mechanisms, the model functions as a decision-support layer that enhances situational awareness.

The framework consists of four sequential stages: data acquisition, feature engineering, supervised classification, and explanation generation. Each stage contributes to transforming raw electrical measurements into meaningful diagnostic insight while maintaining compatibility with real-time monitoring environments.

Voltage and current signals are continuously captured from three-phase transmission or distribution lines using standard measurement devices such as current and voltage transformers. These signals serve as primary indicators of system health, enabling the detection of disturbances ranging from minor fluctuations to severe fault events.

Following acquisition, the raw signals are processed to derive physically interpretable characteristics. Emphasis is placed on features grounded in electrical engineering knowledge so that the resulting model behavior remains explainable rather than purely statistical. Once trained, the classifier evaluates incoming observations and assigns an operational state. The explanation module then identifies the variables that most strongly influenced the prediction, allowing engineers to interpret the reasoning behind the detected condition.

Feature Engineering

Feature engineering plays a decisive role in translating electrical measurements into representations suitable for machine learning. In this study, time-domain attributes were prioritized due to their computational efficiency and relevance for protection-oriented applications.

Root mean square (RMS) values of voltage and current were extracted to capture magnitude variations commonly associated with short circuits and abnormal loading conditions. Because these indicators are already familiar to protection engineers, their inclusion strengthens the physical interpretability of the model. Signal energy was computed over a fixed observation window to quantify disturbance intensity. Fault events typically introduce high-energy transients, enabling the model to differentiate between routine operational variations and critical anomalies.

Additional indicators such as phase angle deviation and current imbalance were incorporated to reflect asymmetrical fault behavior. Collectively, these features reduce dimensional complexity while preserving diagnostic relevance, thereby supporting stable and efficient model training.

Machine Learning Model

The fault identification task was formulated as a supervised classification problem in which operating conditions were categorized as either normal or faulty. Model selection prioritized stability, generalization capability, and interpretability to ensure suitability for safety-critical infrastructure.

The training dataset included labeled observations representing multiple fault scenarios with variations in resistance, location, and load level. Prior to training, feature vectors were normalized to prevent numerical dominance and to promote balanced learning. The dataset was partitioned into training and validation subsets, allowing model performance to be evaluated on previously unseen data. During training, parameters were iteratively updated to minimize classification error until convergence was achieved.

After deployment, the trained classifier processes incoming feature vectors and assigns the most probable operational label. Importantly, the model output is intended to assist engineering judgment rather than directly trigger protection actions, preserving the reliability of established control mechanisms.

Algorithmic Flow

The operational workflow begins with the construction of feature vectors derived from each observation window. These vectors represent the instantaneous electrical state of the system.

Normalization is then applied to maintain numerical consistency across variables. The prepared dataset is subsequently divided to support both learning and validation.

Following training, the classifier evaluates new measurements in near real time. Performance is assessed using standard metrics such as accuracy, precision, recall, and confusion matrices, providing a comprehensive view of detection reliability.

The final output is forwarded to the interpretability layer, where contributing features are identified to support diagnostic understanding.

Operational Considerations

The algorithm is suitable for online monitoring applications because it is designed to operate under real-time constraints with minimal computational overhead. By maintaining a clear separation between detection and control actions, the framework ensures that machine learning enhances situational awareness without compromising established protection reliability.

The structured algorithmic flow also facilitates validation and testing, allowing system operators to assess model behavior under known fault scenarios. This systematic approach strengthens confidence in the machine learning model and supports its integration into modern intelligent power system protection schemes.

Explainability Integration

Interpretability was incorporated to address the transparency limitations often associated with machine learning–based protection systems. After a classification is generated, feature attribution techniques estimate the relative contribution of each input variable. Importance scores assigned to parameters such as voltage magnitude, current imbalance, signal energy, and phase deviation help explain why a particular fault category was predicted. Engineers can therefore verify whether the model's reasoning corresponds with known electrical behavior.

For example, pronounced phase displacement may indicate asymmetrical faults, whereas elevated current combined with high signal energy typically reflects short-circuit conditions. Establishing this connection between learned patterns and physical principles enhances trust and simplifies diagnostic analysis. By transforming model outputs into interpretable insights, the framework supports informed decision-making without sacrificing analytical capability.

Experimental Setup

The proposed framework was evaluated using simulated power system data generated under a range of fault conditions. The study considered single line-to-ground, line-to-line, double line-to-ground, and three-phase faults to ensure that the model was exposed to diverse disturbance patterns. Variations in fault resistance, location along the transmission line, and load levels were also introduced to improve robustness.

Model effectiveness was assessed using widely accepted performance indicators, including accuracy, precision, recall, and F1-score. In addition to quantitative evaluation, the interpretability outputs were qualitatively examined to determine whether the features identified as influential were consistent with established fault characteristics in electrical power systems.

Table 1. *Fault Scenarios Used for Evaluation*

Fault ID	Fault Type	Fault Location (% of line)	Fault Resistance (Ω)	Load Level (%)
F1	Single Line-to-Ground (LG)	20	0.1	80
F2	Single Line-to-Ground (LG)	50	5	100
F3	Line-to-Line (LL)	30	0.2	70
F4	Line-to-Line (LL)	60	10	90
F5	Double Line-to-Ground (LLG)	40	0.5	85
F6	Double Line-to-Ground (LLG)	75	8	110
F7	Three-Phase (LLL)	25	0.01	100
F8	Three-Phase (LLL)	65	2	120
N1	Normal Operation	—	—	90
N2	Normal Operation	—	—	110

Table 2. *Fault Detection Performance (Overall)*

Metric	Value (%)
Detection Accuracy	96.8
Precision	95.9
Recall	96.4
F1-Score	96.1

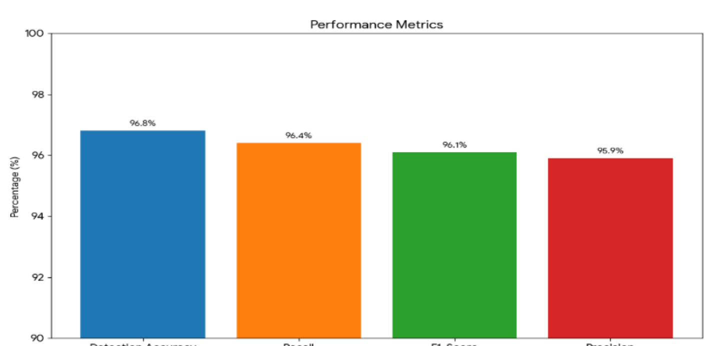

Fig. 1. *Performance metrics of fault detection system*

As illustrated in Fig. 1, all evaluation metrics exceed 95%, confirming the stability of the classifier under varying operating conditions. The narrow variation between the metrics suggests consistent generalization, indicating that the model does not disproportionately favor any single fault class.

***Table 3**. Fault-Type Wise Detection Accuracy*

Fault Type	Detection Accuracy (%)
LG	95.4
LL	96.1
LLG	97.0
LLL	98.3
Normal	96.7

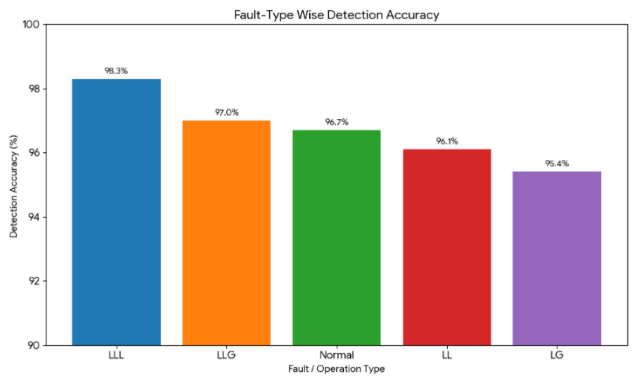

***Fig. 2.** The bar chart illustrates the detection accuracy across different fault types and normal operating conditions*

Results and Discussion

The experimental findings demonstrate that the developed framework can reliably distinguish between normal and faulty operating states across all evaluated scenarios. Detection performance remained consistently high, indicating that the selected features effectively captured the electrical signatures associated with fault events.

A key observation from the interpretability analysis is the dominant influence of voltage deviation and current imbalance in the classification process. This behavior aligns with conventional protection theory, suggesting that the model is learning physically meaningful relationships rather than relying solely on statistical correlations.

The close agreement between precision and recall further indicates balanced predictive behavior with minimal bias toward specific fault categories. Such stability is particularly important in protection environments, where both missed detections and false alarms can have serious operational consequences.

Compared with opaque machine learning approaches, the interpretable structure of the proposed model provides additional diagnostic value. Engineers are not only informed that a fault has occurred but are also given insight into the contributing electrical parameters. This transparency supports faster validation and may improve operator confidence when integrating intelligent tools into existing protection workflows.

Overall, the results suggest that combining supervised learning with explainability techniques can enhance situational awareness while maintaining strong analytical performance.

Simulation Setup and Parameter Configuration

Validation of the proposed framework was carried out using a detailed three-phase transmission system modeled in MATLAB/Simulink (Simscape Electrical). The simulation environment enabled controlled manipulation of system variables, ensuring repeatability while reflecting realistic grid behavior.

Fault events were initiated at predefined time instants, and voltage and current responses were recorded at a sampling frequency sufficient to capture both transient disturbances and steady-state dynamics. A sampling rate of 10 kHz was selected as a practical balance between signal resolution and computational efficiency.

The simulated network represents a medium-voltage transmission line supplied by a balanced three-phase source and connected to an RL load. Fault locations ranged between 20% and 80% of line length, while resistance and loading conditions were systematically varied to generate representative operating scenarios.

Time-domain features extracted from the recorded waveforms were subsequently used to train and evaluate the classifier. This controlled simulation strategy provides a reliable foundation for assessing detection capability prior to real-world deployment.

Parameter	Value
Simulation Platform	MATLAB/Simulink (Sims cape Electrical)
System Type	Three-phase transmission system
Nominal Line Voltage	132 kV
System Frequency	50 Hz
Transmission Line Length	100 km
Line Model	Distributed parameter line
Sampling Frequency	10 kHz
Observation Window	1 cycle (20 ms)
Fault Inception Time	0.1 s
Fault Duration	0.1 s
Fault Types	LG, LL, LLG, LLL
Fault Location	20%–80% of line length
Fault Resistance	0.01–10 Ω
Load Type	Three-phase RL load
Load Variation Range	70%–120% of nominal load
Noise Level	1% Gaussian noise
Feature Extraction Domain	Time-domain
ML Training–Testing Split	70% – 30%
Performance Metrics	Accuracy, Precision, Recall

Simulation Workflow Description

During each simulation run, fault conditions were initiated at the specified fault inception time and cleared after the defined duration. Voltage and current wave-forms were continuously monitored at the sending end of the transmission line. The extracted features were used to train and test the machine learning model and were labeled according to fault type and operating condition.

The use of controlled simulation parameters allows the proposed framework to be evaluated under reproducible and diverse scenarios, while maintaining con-sistency with practical power system behavior. Before deploying the proposed method in the real world, this simulation-based validation provides a solid foundation for evaluating its fault detection capability and explainability.

Limitations

One limitation of the present study is the reliance on simulated datasets, which may not fully capture the stochastic disturbances and measurement noise observed in real transmission networks. Additionally, model performance may vary when exposed to previously unseen grid configurations. Future work will therefore focus on validation using field data to further establish practical reliability.

Conclusion and Future Work

This study presented an interpretable machine learning framework for fault identification in modern power systems. By integrating explainability techniques with supervised classification, the proposed approach delivers both strong detection capability and transparent diagnostic insight.

The observed alignment between influential features and established electrical principles reinforces confidence in the model's analytical behavior. Such transparency is essential for safety-critical infrastructure, where understanding the basis of automated decisions is as important as predictive accuracy.

While the current work focuses on simulation-based validation, future research should prioritize hardware implementation and testing using real operational data. Extending the framework toward adaptive protection and predictive maintenance also represents a promising direction.

In summary, the developed methodology demonstrates that interpretability and performance need not be mutually exclusive, and that explainable models can play a meaningful role in the evolution of intelligent power system protection.

References

1. Anwar T, Mu C, Yousaf MZ, et al. (2025) Robust fault detection and classification in power transmission lines via ensemble machine learning models. *Scientific Reports* 15:2549
2. Awan D, Khan MU, Zia M, Lais M, Tabassum S, Hamza A, Khan MS (2025) Development of explainable AI (XAI) framework for fault diagnosis in power electronics systems. *Phys. Educ., Health Soc. Sci.* 3(3):56–70
3. Cavus RK, Allahham A (2025) AI-assisted power system fault detection: performance and challenges. *AJATES* 1(1)
4. Danish N, Senjyu H (2025) Supervised learning techniques for fault classification in power systems. *AJATES* 1(1)
5. Draper et al. (2025) Explainable artificial intelligence for critical infrastructure monitoring and control. *ITSI Transactions on Electrical and Electronic Engineering* 12(2):18–23
6. Eshraghi P, Zhang X, Raees A (2025) Explainable AI for energy systems maintenance. *Renewable and Sustainable Energy Reviews*
7. Explainable artificial intelligence based intelligent fault diagnosis: A systematic review from applications to insights (2025) *Reliability Engineering & System Safety* 267
8. Explainable artificial intelligence in fault detection and diagnosis (2026) Taylor & Francis (accepted; available as 2025 preprint)
9. Goyal S, Tasmary Salsabil I, Kumar A, Ukey M, VSP (2025) AI-driven fault detection and diagnosis in smart grids for enhanced power system reliability. *Journal of Information Systems Engineering and Management* 10(42s):1–12
10. Hu Q (2025) Forecasting real estate price via SVM. *Advances in Economy, Management and Political Sciences*
11. Jamil T, Peeters C, Verstraeten T, Helsen J (2025) Leveraging signal processing and machine learning for automated fault detection in wind turbine drivetrains. *Wind Energy Science* 10:1963–1978
12. Jeon JE, Hong SJ, Han S-S (2024) Utilization of machine learning and explainable artificial intelligence (XAI) for fault prediction and diagnosis in wafer transfer robot. *Electronics* 13(22):4471
13. Kong J (2024) House price prediction uses machine learning. *Applied Computational Engineering*
14. Liu Y, Li P, Si Y, Ma L (2025) A comprehensive review of AI integration for fault detection in modern power systems: Data processing, modeling, and optimization. *Energies* 18(18):4983
15. Maselli M, Nesticò A (2025) Machine learning algorithms and explainable AI for property valuation. *Real Estate* 2(3)
16. Nasim SF, Aziz S, Qaiser A (2024) Fault detection and diagnosis in power systems using AI: A review. *Sir Syed University Research Journal of Engineering and Technology* 14(1):27–32
17. Prediction of residential real estate price using machine learning (2024) *Procedia Computer Science* 246:3142–3147
18. Sabid R, Kamrul K (2025) Deep learning optimization for power grid fault prediction. *AJATES* 1(1):258–289
19. Saarela M, Podgorelec V (2024) Recent applications of explainable AI (XAI): A systematic literature review. *Applied Sciences* 14(19):8884
20. Tasmary A, Goyal R (2025) AI-based fault detection and diagnosis for smart grids. *Journal of Information Systems Engineering and Management*
21. VM et al. (2025) AI-driven cybersecurity framework for anomaly detection in power systems. *Scientific Reports* 15:35506
22. Wang S, et al. (2025) Machine learning approaches to grid fault classification. *AJATES* 1(1)
23. Xu M, Zhang L, Li F (2023) Explainability in knowledge-based systems and machine learning models for smart grids. *Frontiers in Energy Research*
24. Zachariades C, Xavier V (2025) A review of artificial intelligence techniques in fault diagnosis of electric machines. *Sensors* 25(16):5128