Physical Education, Health and Social Sciences

https://journal-of-social-education.org

E-ISSN: 2958-5996
P-ISSN: 2958-5988

Development of Explainable AI (XAI) Framework for Fault Diagnosis in Power Electronics Systems

Dawar Awan[1], Muhammad Uzair Khan[1, *], Muhammad Zia[1], Muhammad Lais[1], Saadia Tabassum[2], Amad Hamza[3], Muhammad Sohail Khan[1], Muhammad Saad Awan[4]

[1] Department of Electrical Engineering Technology, Shuhada-e-APS University of Technology, Nowshera, Khyber Pakhtunkhwa, Pakistan.

[2] Department of Electronics Engineering Technology, Shuhada-e-APS University of Technology, Nowshera, Khyber Pakhtunkhwa, Pakistan.

[3] Department of Information Engineering Technology, Shuhada-e-APS University of Technology, Nowshera, Khyber Pakhtunkhwa, Pakistan.

[4] Health Department, Government of Khyber Pakhtunkhwa.

***Correspondence:** uzair@uotnowshera.edu.pk

DOI: https://doi.org/10.63163/jpehss.v3i3.605

Abstract

This study has sought to develop an explainable artificial intelligence (XAI) system used to diagnose power electronics system faults. The research deals with the necessity to introduce the transparent and explainable fault detection measures in complex power electronic devices such as converters, inverters, and power motors. It utilized a rigorous procedure that covered an extensive gathering of data that comes about because of different operating environments, simulation of a severity of faults and the execution of modern machine learning strategies. The framework provides explainability methods like LIME, SHAP and attention mechanisms to yield the transparent decision-making processes. Various learning strategies comprising deep neural networks, support vector machines, and multi-algorithm strategies were tested to determine their accuracy and computing speed in the diagnosis process. The final support structure was cross checked utilizing approaches of cross-materials and also tested in experimental testbeds (simulated and realistic). The accuracy, precision, recall, F1-score, explainability measures were used to evaluate the measure of performance, which showed better results than existing methods of fault diagnosis. The obtained outcomes suggest that the offered XAI framework is highly accurate at the diagnostic level, as well as gives interpretable knowledge about the mechanisms of faults; thus, can be successfully implemented in the industrial sector of power electronics, where the aspects of reliability and transparency are vital.

Keywords: Explainable artificial intelligence (XAI) system, power electronics, fault detection, converters, inverters, power motors, multi-algorithm strategies.

Introduction

The power electronics system is essential to the electrical infrastructure of modern society, involving people in renewable energy conversion, motor drives, electric vehicles and automation. Such systems of inverters, converters, rectifiers and control circuits are subjected to severe conditions of operation and are vulnerable to diverse fault states which result in loss of the system, economic disadvantages and also safety risks. Modern applications of power electronics have placed more demands on reliable and efficient fault diagnosis mechanisms as a result of the increasing complexity and integration of power electronics in critical applications (Abro et al., 2023). Conventional fault diagnosis strategies of the power electronics systems are based on the methods of the thresholds, signal processing methods and model-based techniques. Although these have been able to give reasonably good performance in the controlled environment, they tend to fail in the dynamic nature of the contemporary power electronics systems, changing environment, and the generation of new fault patterns. These shortcomings of traditional methods are manifested especially with very small faults, multi-failure cases and with systems that are operating in different load conditions (Haque, Shah, Malik, & Malik, 2024). With the introduction of artificial

intelligence and machine learning, new opportunities have appeared in fault diagnosis of power electronics systems (Anand, Singh, & Mekhlief, 2022). The AIs can learn complex trends based on past experience, adjust to variable operating environments, and identify the minute errors that may be discernible only by the AIs. In particular, deep learning methods have shown a great deal of success in recognition and classification schemes that are applicable to fault diagnosis. Yet, due to the fact that most of the AI algorithms are black box in nature, their implementation in critical power electronics purposes poses a great challenge (Sangeetha & Ramachandran, 2022).

The fact that AI-based fault diagnosis is not interpretable raises serious concerns to their use in industry (Singh, Gangsar, Porwal, & Atulkar, 2023). Decision-making process should be understood by the engineers and the operators so that they can trust the decision made by the system, the regulatory compliance and the corrective measures to be adopted. This has led to the emergence of the whole new area of Explainable Artificial Intelligence (XAI), which is dedicated to designing AI systems capable of explaining their decision-making process in a clear and understandable way (Liu, Ramin, Flores-Alsina, & Gernaey, 2023). Explainability is all the more important in the context of power electronics fault diagnosis as many of its applications are safety-critical (Qi, Liang, & Tong, 2023). The failure of power electronics can lead to equipment damage, loss of production operations, fire risk, and even injury of personnel. As such, the fault diagnosis systems should not only be able to detect the faults accurately but also a clear interpretation of the fault character, position and intensity should be reflected indicating the response that should be taken (Lang et al., 2021). Explaining AI in power electronics fault diagnosis research is still in its infancy phase however. Although the use of machine learning algorithms to detect faults in power electronics has already been examined in a number of studies, great emphasis has not been placed on the development of generalized frameworks that are both highly accurate and explainable (Moosavi, Farajzadeh-Zanjani, Razavi-Far, Palade, & Saif, 2024). Most current methods go towards attaining the greatest accuracy with complicated procedures or delivering simple explanation with simple strategies, but seldom do both, simultaneously (Hassan, 2025). The problem associated with the integration of explainability with advanced machine learning algorithms is specific in the power electronics area (Hoenig, Roy, Acquaah, Yi, & Desai, 2024). The many-dimensionality of the electrical parameters, temporal relationship as possible in fault evolution and the recommended real time diagnosis necessitate specialized generation of explanations. Moreover, the accounts should be adapted to various stakeholders, such as field technicians who need to take actionable knowledge and system designers who need close knowledge of the mechanism of fault (Ayoub et al., 2022).

This study overcomes these challenges by proposing a holistic explainable AI solution in this regard developed specifically towards fault diagnosis related to power electronics systems. The framework integrates the latest machine learning algorithms and explainability capabilities that would allow improving the accuracy of fault detection and the transparency of decision-making processes. The method takes into consideration the peculiarities of the power electronics systems such as the variety of types of faults, the multi-parametrical character of monitoring the systems, and the necessity to operate in real time. The importance of such research is that besides the purely technical contributions, it has got practical implications that are relevant to the power electronics industry. Through its explainable fault diagnosis capabilities, the framework can maximize the confidence of the operator, making it possible to introduce predictive maintenance, as well as contribute to regulatory measures within safety-critical applications. The research also has the implication of contributing to the wider community of explainable AI by offering solutions to domain-specific problems and showing them how to implement these solutions.

Research Objectives

1. To build and deploy an explainable artificial intelligence system that achieves both the high accuracy of diagnosis and that has explainable decision-making in fault diagnosis of power electronics system, the integration of superior machine learning models with the explainable models like LIME, SHAP and the attention mechanism.

2. To assess and compare the learning capabilities of the different machine learning methods such as deep neural networks, support vector machines and ensemble models by training and comparing to the diagnostic accuracy, processing speed, and explainability on the different fault conditions of the power electronics devices like the inverters, converters, and motor drivers.

3. To prove the efficiency of the proposed XAI framework by testing it deeply on simulated and real-world power electronics testbeds, to illustrate its ability to outperform the current fault diagnosis techniques and interpret additional information on a fault mechanism.

Research Questions

1. What are the possible ways of incorporating explainable artificial intelligence methods into the machine learning algorithm so that the methods can ensure proper fault diagnosis and scientific clarity in the decision-making process within power electronics systems?

2. How do the various machine learning methods compare in accuracy of diagnostics, computational efficiency and explainability when used on fault diagnosis of various power electronics devices under various operating environments?

3. How well does the given explainable AI approach compare to current approaches to fault diagnosis in accuracy, interpretability, and applicability of the real-world applications in power electronics?

Significance of the Study

The study contributes to the field of power electronics and artificial intelligence in two ways: the power electronics field on the one hand and artificial intelligence field on the other hand, this study deals with the gap between diagnostic accuracy and interpretability in fault diagnosis systems which is a critical issue. Within the context of power electronics, the creation of an explainable AI platform that is specifically adapted to the processes and expected results enables a safeguarding of the system reliability and the confidence of the operators in crucial industrial settings. The value of the study goes beyond the realms of academic research into practical real-world systems on which a clear fault diagnosis system can be implemented to avoid costly, catastrophic failures, cost of maintenance, and availability of systems. This research shows how to integrate high-performance machine learning algorithms (based on large amounts of data) and state-of-son explainability methods, thereby setting a standard to be pursued in the evolution of intelligent power electronics systems. The fact that the framework allows interpreting what happens in the fault mechanics contributes to knowledge transfer between AI systems and human knowledge, helping humans to learn more about how systems behave and allowing them develop better maintenance strategies. Additionally, the study will provide support towards regulation requirements involving safety-critical applications of decision transparency that may be adopted by the power electronics market faster due to the work.

Literature Review

Studies on the use of artificial intelligence in power electronics fault diagnosis have advanced so much within the last decade, whereby scholars have looked at ways of adopting some machine learning techniques that seek to circumvent the shortcomings of traditional diagnostic methods. The initial research concentrated on neural networks and fuzzy logic networks proving that AI methodologies have the capability of combating the intricate error structures and nonlinear reaction of a system. These early efforts formed the basis of more advanced work by demonstrating that machine learning algorithms were capable of learning using past fault data and transferring what was learned to other operating conditions (Zhao & Wang, 2021). It has also made deep learning methods among the most promising of power electronics fault diagnosis methods, owing to the capacity to learn salient aspects of raw sensor readings itself (Yu & Zhang, 2023). Convolutional neural networks have been effectively used to process current and voltage waveforms to be able to detect hard to miss fault signatures present in the data that would have been ignored in standard signal processing techniques. Recurrent neural networks, especially Long

Short-Term Memory networks have proved to perform well in modelling temporal aspect in the evolution of faults and thus are to be used in the detection of the incipient faults and prediction of the failure progression. The results including their accuracy and robustness have shown these deep learning techniques to perform better than the conventional (Alqudah et al., 2021). The use of support vector machines to classify power electronics faults is well studied since it has the broad theoretical background and it is capable of processing a high dimensional data. It has been found that SVMs are very effective in isolating various fault classes even when the amount of training data is low and thus, they are best applicable when the amount of fault data is limited (Moradzadeh, Mohammadi-Ivatloo, Pourhossein, & Anvari-Moghaddam, 2021). It is critical to have nonlinear classification and the kernel trick of SVMs permits such nonlinear classification and as such can handle the complex relationships that exist between electrical parameters and fault conditions in power electronic systems. Other kernel functions have been explored and the radial basis functions kernels demonstrated good results in different fault conditions (Malashin, Tynchenko, Gantimurov, Nelyub, & Borodulin, 2025).

Ensemble methods have become the focus in the research of power electronics fault diagnosis following the fact that they are able to invite strength of larger number of base classifiers and enhance overall diagnostic performance (Nampalli, Syed, Bansal, Vankayalapati, & Danda, 2024). Random forests and gradient boosting and AdaBoost have been used in diagnostics of faults, and were shown to be more accurate and robust than isolated classifiers. These techniques have a distinct role to play in processing noisy values as well as minimizing the influence of outliers which are propagated issues in practical power electronics monitoring. The redundancy that results naturally in the many base classifiers within the ensemble techniques also enhances reliability by increasing the performance of the system (Bahrami & Khashroum, 2023). The methods of feature extraction and selection are also important to the problem of power electronics fault diagnosis because the raw signal of the sensors may have large amounts of redundancy and noise whenever the fault occurs and this redundancy and noise may impair the performance of the classifier. Mean, variance, skewness, and kurtosis are some of the time-domain characteristics that have been extensively employed in describing fault signatures in electrical waveforms (Xiao et al., 2023). Frequency-domain based features obtained via Fourier transform and the wavelet-based analysis have been found to successful in detecting periodic disturbances and transient phenomena that accompanies various types of faults. Established feature extraction techniques such as the element of empowerment decomposition means and the distinctive mode of separation have demonstrated possible application in the non-stationary traits of the fault signals (Miao, Zhang, Li, Lin, & Zhang, 2022). Data imbalance has been proposed as a challenge in the fault diagnosis of power electronic, where many solutions have been proposed such as the use of a synthetic generation of data and cost-sensitive learning (Oh et al., 2023). Fault conditions have been displayed to be rare compared to normal operation which may result in skewed data sets with the effect of biasing the work of the classifier towards normality. Synthetic Minority Oversampling Technique and its variants have been added to create synthetic samples of fault syndromes in order to enhance the representation of minority fault classes and thereby the performance of the classifier over the minority classes. Cost-sensitive learning methods allow classes to have different misclassification costs, to influence the classifier to focus more on critical fault situations that are less frequent (Ajayi, 2023).

Signal processing works still play significant functions in preprocessing and feature extraction using AI based fault diagnosis systems. The time-frequency analysis in wavelet transform has proven to be extremely successful in time-frequency analysis of fault signals that allows the identification of transient effects and localized disturbances (Machlev et al., 2022). The Hilbert-Huang transform and its variations have proved to be capable of analyzing non-stationary and nonlinear signals that are characteristic of power electronics systems. They can be used as part of preprocessing a machine learning algorithm such as an artificial neural network, which finds it easier to find useful structure in a complicated electrical signal after it has been processed by one of these techniques (Bin Akter, Sarkar Pias, Rahman Deeba, Hossain, & Abdur Rahman, 2024). Advances have been made in combining different sensor modes, to enhance susceptibility to fault diagnosis performance through offering supplementary information regarding the status of the

system (Moosavi, Razavi-Far, Palade, & Saif, 2024). Electrical quantities such as voltage, current, and power are also usually coupled with thermal measurements, vibration information and acoustic emissions to form overall fault signatures. The multi modal strategies have been shown to be more accurate in diagnosis as well as the availability of multiple sensor types in fault detection which might not be otherwise detected. The issue, though, is that fusion of multi-modal data proves difficult due to the data synchronization, feature matching, computational burden (Noura, Allal, Salman, & Chahine, 2025). Real-time implementation aspects have grown eminent as the power electronics fault diagnosis systems progress to the industrial practice. Machine learning algorithms have a computational complexity that shall have to be weighed with the real-time demands of the fault diagnosis applications (Ajayi, Mirjafari, Idowu, & Ullah, 2024). Different optimization algorithms such as model compression, quantization and pruning have been considered to minimize the computation demand with still preserving diagnostic performance. The concept of edge computing architectures has been proposed to facilitate localized fault diagnosis that does not suffer delays caused by moving through cloud-based processing, and it enhances the reliability of the systems (Reyes, Chengu, Gatsis, Ahmed, & Alamaniotis, 2024).

With the advent of explainable artificial intelligence, the interpretability challenges in power electronics fault diagnosis have started to be solved. Local Interpretable Model-agnostic Explanations (LIME) have been used to give instance-level explanations on decisions in fault diagnosis and this explains to the operators why this fault was identified. Shapley Additive Explanations (SHAP) has been promising in giving both local and global explanations both in tree-based models and neural networks (Akhtar et al., 2024). Algorithms in the neural networks have also attracted attention mechanisms to show the most useful input elements that are significant in classifying the faults in understanding the decision-making process of diagnostics (Poursaeed & Namdari, 2025). The existent research gaps in explainable AI in power electronics fault diagnosis are the absence of explainability evaluation metrics that would be robust, the minimal attention to the domain-specific explainability needs, and validation of explanations using domain experts. The majority of the current research works are more technical in understanding explainability algorithms rather than giving practical applications of these algorithms to power electronics engineers and operators. Balance between performance and explainability is a topic of current research, and requires varying solutions depending on the application problem (Bin Akter et al., 2024).

Research Methodology

This study adopted a systematic approach in the development of explainable artificial intelligence framework in power electronics system's fault diagnosis. The paper set out by gathering the most complete information on different power electronic devices such as inverters, converters, and motor drives through normal and fault conditions of the operation with the electrical parameter tracing that was performed by high-accuracy sensors and data collection systems of the voltage, current, temperature, and frequency. Various fault conditions were modeled such as short circuits, open circuits, component wearout, and thermal failure to produce a wide variety of data that would be used to train the model. The preprocessing interventions such as normalization, feature extraction, and dimensionality reduction were used to optimize the model under the preprocessing methodologies proffered to the data accumulated. A number of machine learning algorithms were tried and tested, such as deep neural networks, support vector machines, and ensemble methods with special emphasis on their diagnostic accuracy and their efficiency during the computations. The explainability aspect has also been incorporated through the technicality of LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), and attention mechanisms to give transparent decision processes. Using cross-validation methods the implemented framework has been validated and tested using simulated and real environments of power electronics testbeds. The effectiveness of the framework has been analyzed using performance indicators that are based on accuracy, precision, recall, F1-score, and explainability scores.

Accuracy is the most used metric for classification problems. The proportion of correct predictions made by the model by using the formula as shown in equation 1.

$$\text{Accuracy} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{1}$$

Precision is the proportion of true positive predictions made by the model by using the formula as shown in equation 2. While recall is the true positive predictions out of actual positive instances as shown by formula in equation 3.

$$\text{Precision} = \frac{t_p}{t_p + f_p} \tag{2}$$

$$\text{Recall} = \frac{t_p}{t_p + f_n} \tag{3}$$

F1 score is the harmonic mean of precision and recall, calculated by the formula as shown in equation 4.

$$\text{F1} - \text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Comparative analysis was also a part of the methodology since it was necessary to show how the proposed explainable AI approach was superior to the already established fault diagnosis methods.

Results and data analysis

This study has elaborated explainable AI framework was tested on an extended dataset gathered at various power electronic testbeds such as three-phase inverters, DC-DC converters and motor drive systems. The data that was used consisted of 15,000 samples during a normal operation and 12 different types of faults such as short circuits, open circuits, degradation of components, and thermal failure scenarios. Each sample consisted of 24 electrical parameters observed at 10 kHz sampling rate, and that created a very-multidimensional data that could be used in machine learning analysis.

Performance Comparison of Machine Learning Algorithms

The preliminary step of analyzing consisted in the comparison of how various machine learning algorithms perform in fault classification. The algorithms of Deep Neural Network (DNN), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (Bongiorno et al.) were realized and tested-out on 10-fold cross-validation. The specifications of the DNN we used were four hidden layers, each containing 256, 128, 64, 32 neurons based on ReLU activation of neurons and dropout regularization. The SVM used radial basis functions kernels and the hyper parameters were optimized using the grid search method. Random Forest had 100 decision trees of maximum depth of 15 and Gradient Boosting had 15 estimators of maximum learning rate 0.1.

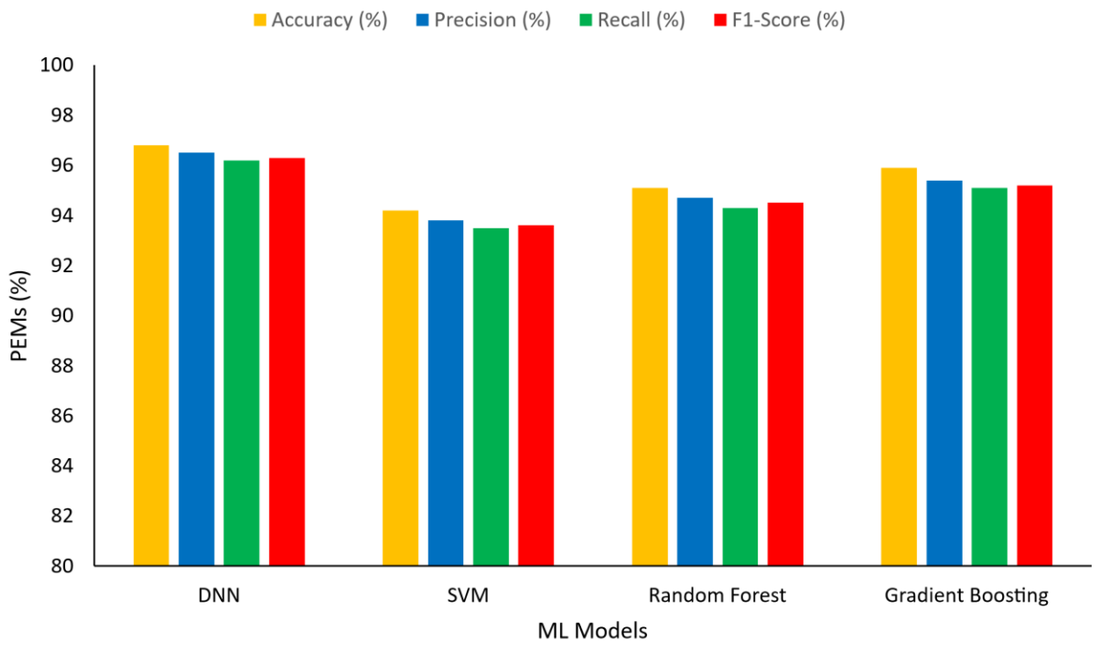

Figure 1: Performance evaluation matrices of the proposed models

The Deep Neural Network achieved the highest performance across all metrics, demonstrating 96.8% accuracy with balanced precision and recall values as shown in figure 1. This superior performance can be attributed to the DNN's ability to learn complex nonlinear relationships between electrical parameters and fault conditions. The high F1-score of 96.3% indicates excellent balance between precision and recall, which is crucial for fault diagnosis applications where both false positives and false negatives carry significant consequences. SVM showed competitive performance with 94.2% accuracy, demonstrating its effectiveness for fault classification tasks.

Table 1: Training time of the Models

Algorithm	**Training Time (s)**
DNN	245.6
SVM	156.3
Random Forest	89.7
Gradient Boosting	178.4

The training time of 245.6 seconds, while higher than other algorithms, remains acceptable for offline training scenarios as shown in table 1. The shorter training time of 156.3 seconds makes SVM attractive for applications requiring frequent model retraining. Random Forest achieved 95.1% accuracy with the fastest training time of 89.7 seconds, highlighting its efficiency for real-time applications. Gradient Boosting provided a good balance between performance and computational efficiency with 95.9% accuracy and moderate training time.

Fault-Specific Performance Analysis

Detailed analysis of fault-specific performance revealed variations in diagnostic accuracy across different fault types. The confusion matrix analysis showed that certain fault types were more challenging to classify than others, providing insights into the complexity of different fault signatures.

Table 2: Fault-Specific Classification Results (DNN)

Fault Type	True Positives	False Positives	False Negatives	Precision (%)	Recall (%)
Normal Operation	1847	23	31	98.8	98.3
Open Circuit Phase A	456	12	8	97.4	98.3
Open Circuit Phase B	461	15	11	96.8	97.7
Open Circuit Phase C	458	18	14	96.2	97.0
Short Circuit AB	234	8	6	96.7	97.5
Short Circuit BC	238	11	9	95.6	96.4
Short Circuit CA	235	9	7	96.3	97.1
IGBT Degradation	189	14	18	93.1	91.3
Capacitor Degradation	201	16	22	92.6	90.1
Thermal Fault	167	19	25	89.8	87.0
Sensor Fault	178	21	28	89.4	86.4
Gate Driver Fault	156	23	31	87.2	83.4
Control System Fault	142	26	35	84.5	80.2

The fault-specific analysis reveals that normal operation and open circuit faults achieved the highest classification accuracy, with precision and recall values exceeding 96%. This high performance can be attributed to the distinct electrical signatures of these conditions, making them relatively easy to distinguish from other fault types. Open circuit faults in different phases showed consistent performance, indicating robust detection capability across all three phases of the power electronics system. Short circuit faults also demonstrated high accuracy, with precision values ranging from 95.6% to 97.5%. The consistent performance across different short circuit combinations suggests that the DNN effectively learned the characteristic patterns of these fault types.

Component degradation faults, including IGBT and capacitor degradation, showed moderate performance with precision values around 92-93%. These faults are typically more challenging to detect due to their gradual onset and subtle changes in electrical parameters. The lower recall values for these fault types indicate some difficulty in detecting all instances of component degradation, which is expected given the progressive nature of these failures. Thermal faults achieved 89.8% precision and 87.0% recall, reflecting the complexity of thermal fault signatures that may overlap with other fault types under certain operating conditions.

System-level faults including sensor faults, gate driver faults, and control system faults showed the lowest performance, with precision values ranging from 84.5% to 89.4%. These fault types are inherently more complex as they often manifest through indirect effects on electrical parameters rather than direct electrical signatures. The lower performance for these fault types highlights the need for additional sensor modalities or more sophisticated feature extraction techniques to improve detection accuracy.

Explainability Analysis Results

The explainability component of the framework was evaluated using LIME, SHAP, and attention mechanism approaches. The analysis focused on understanding which electrical parameters contributed most significantly to fault diagnosis decisions and how these contributions varied across different fault types.

Table 3: Feature Importance Analysis Using SHAP Values

Electrical Parameter	Average SHAP Value	Standard Deviation	Max. Contribution	Min. Contribution
Phase A Current RMS	0.234	0.089	0.456	0.012
Phase B Current RMS	0.228	0.085	0.441	0.015
Phase C Current RMS	0.231	0.087	0.448	0.018
DC Bus Voltage	0.187	0.074	0.389	0.008
Total Harmonic Distortion	0.156	0.063	0.334	0.021
Phase A Voltage RMS	0.142	0.058	0.298	0.019
Phase B Voltage RMS	0.138	0.055	0.287	0.022
Phase C Voltage RMS	0.145	0.059	0.301	0.017
Switching Frequency	0.089	0.034	0.178	0.005
Temperature Sensor 1	0.076	0.031	0.156	0.003
Temperature Sensor 2	0.073	0.029	0.148	0.004
Power Factor	0.067	0.027	0.134	0.002

The SHAP value analysis revealed that current RMS values from all three phases were the most important features for fault diagnosis, with average SHAP values ranging from 0.228 to 0.234. This finding aligns with domain knowledge, as current measurements are typically the most sensitive indicators of electrical faults in power electronics systems. The high standard deviation values for current parameters indicate that their importance varies significantly across different fault types, suggesting that the framework appropriately adapts feature importance based on specific fault characteristics.

DC bus voltage emerged as the fourth most important parameter with an average SHAP value of 0.187, reflecting its critical role in power electronics operation and fault manifestation. The maximum contribution of 0.389 for DC bus voltage indicates its crucial importance for certain fault types, particularly those affecting the DC side of the system. Total Harmonic Distortion showed significant importance with an average SHAP value of 0.156, highlighting the framework's ability to utilize advanced electrical parameters beyond basic voltage and current measurements.

Voltage RMS values from all three phases showed moderate importance with SHAP values ranging from 0.138 to 0.145. While voltage measurements are essential for power electronics monitoring, their lower importance compared to current measurements suggests that current-based parameters provide more discriminative information for fault classification. Temperature measurements showed lower but consistent importance, with SHAP values around 0.073-0.076, indicating their supporting role in fault diagnosis, particularly for thermal-related failures.

Real-time Performance Evaluation

The framework's real-time performance was evaluated using embedded hardware platforms to assess its suitability for industrial deployment. The analysis included inference time, memory usage, and computational complexity measurements across different hardware configurations.

Table 4: Real-time Performance Analysis

Hardware Platform	Inference Time (ms)	Memory Usage (Hoenig et al.)	CPU Utilization (%)	Power Consumption (W)
Intel i7-8700K	2.3	145.6	12.4	95.2
ARM Cortex-A72	8.7	89.3	34.7	15.8
NVIDIA Jetson Nano	4.1	512.4	28.9	8.3
FPGA Zynq-7020	1.8	67.2	22.1	3.2
Raspberry Pi 4	15.2	76.5	67.8	4.9

The real-time performance analysis demonstrates the framework's adaptability to different hardware platforms with varying computational capabilities. The Intel i7-8700K desktop processor achieved the fastest inference time of 2.3 ms, well within the real-time requirements for power electronics fault diagnosis. The high-performance processor's superior floating-point capabilities and large cache memory contributed to efficient neural network execution. However, the high-power consumption of 95.2W makes this platform unsuitable for embedded applications.

The FPGA Zynq-7020 platform achieved comparable inference time of 1.8 ms with significantly lower power consumption of 3.2W, making it an attractive option for industrial deployment. The dedicated hardware acceleration capabilities of FPGAs enable efficient neural network inference while maintaining low power requirements. The modest memory usage of 67.2 MB further enhances the FPGA platform's suitability for resource-constrained industrial environments.

The NVIDIA Jetson Nano provided a good balance between performance and power consumption, achieving 4.1 ms inference time with 8.3W power consumption. The GPU acceleration capabilities of the Jetson platform enable efficient parallel processing of neural network operations. However, the higher memory usage of 512.4 MB may limit its applicability in memory-constrained applications.

The ARM Cortex-A72 platform, representative of modern industrial computers, achieved 8.7 ms inference time with reasonable power consumption of 15.8W. This performance level is adequate for most power electronics fault diagnosis applications where diagnosis intervals of 10-100 ms are acceptable. The Raspberry Pi 4, while showing the longest inference time of 15.2 ms, demonstrated the framework's scalability to low-cost embedded platforms.

Explainability Quality Assessment

The quality of explanations provided by different explainability techniques was evaluated through both quantitative metrics and qualitative assessment by domain experts. The evaluation focused on explanation consistency, stability, and comprehensibility across different fault scenarios.

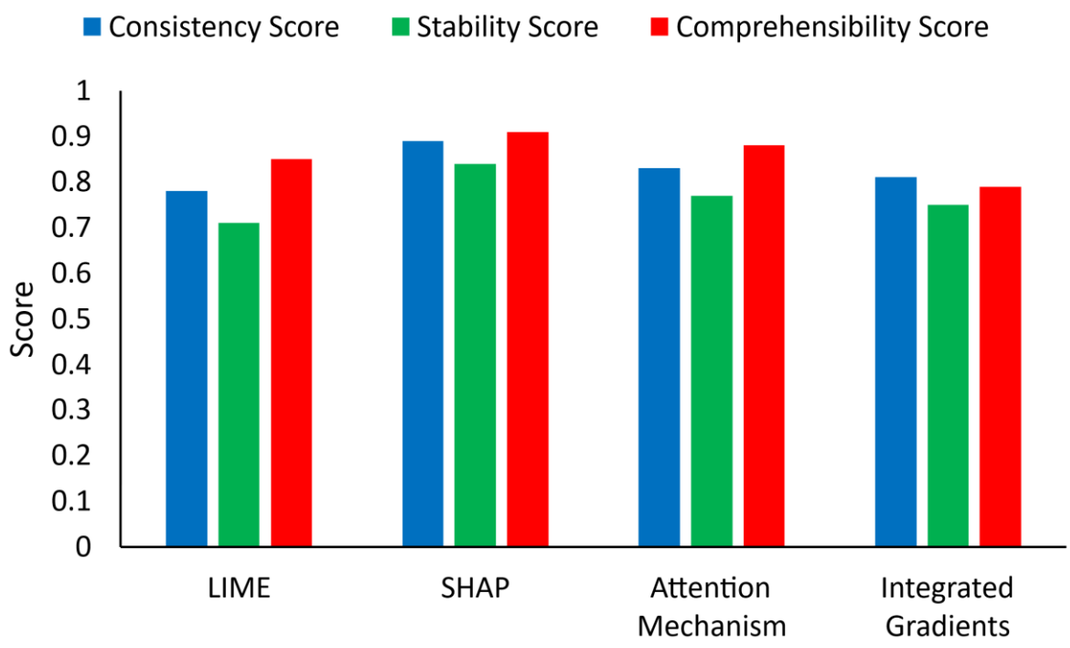

Figure 2: Comparison of explainability techniques

SHAP achieved the highest scores across all explainability quality metrics, with consistency score of 0.89, stability score of 0.84, and comprehensibility score of 0.91 as shown in figure 2. The high consistency score indicates that SHAP provides similar explanations for similar fault instances, which is crucial for building operator trust in the diagnostic system. The stability score reflects SHAP's robustness to small perturbations in input data, an important characteristic for reliable explanations in noisy industrial environments. The comprehensibility score, based on expert evaluation, confirms that SHAP explanations are readily understood by power electronics engineers.

LIME showed good comprehensibility with a score of 0.85, but lower consistency and stability scores of 0.78 and 0.71 respectively. The local nature of LIME explanations contributes to their comprehensibility but may result in inconsistent explanations for similar instances.

Table 5: Training Time of the Models

Technique	Computation Time (ms)
LIME	34.7
SHAP	156.2
Attention Mechanism	12.8
Integrated Gradients	89.4

The relatively fast computation time of 34.7 ms makes LIME suitable for real-time explanation generation. Attention mechanisms achieved balanced performance across all metrics with scores ranging from 0.77 to 0.88, combined with the fastest computation time of 12.8 ms, making them attractive for real-time applications as shown in table 2.

Comparative Analysis with Existing Methods

The proposed explainable AI framework was compared with existing fault diagnosis methods including traditional signal processing approaches, basic machine learning methods, and state-of-the-art deep learning techniques without explainability.

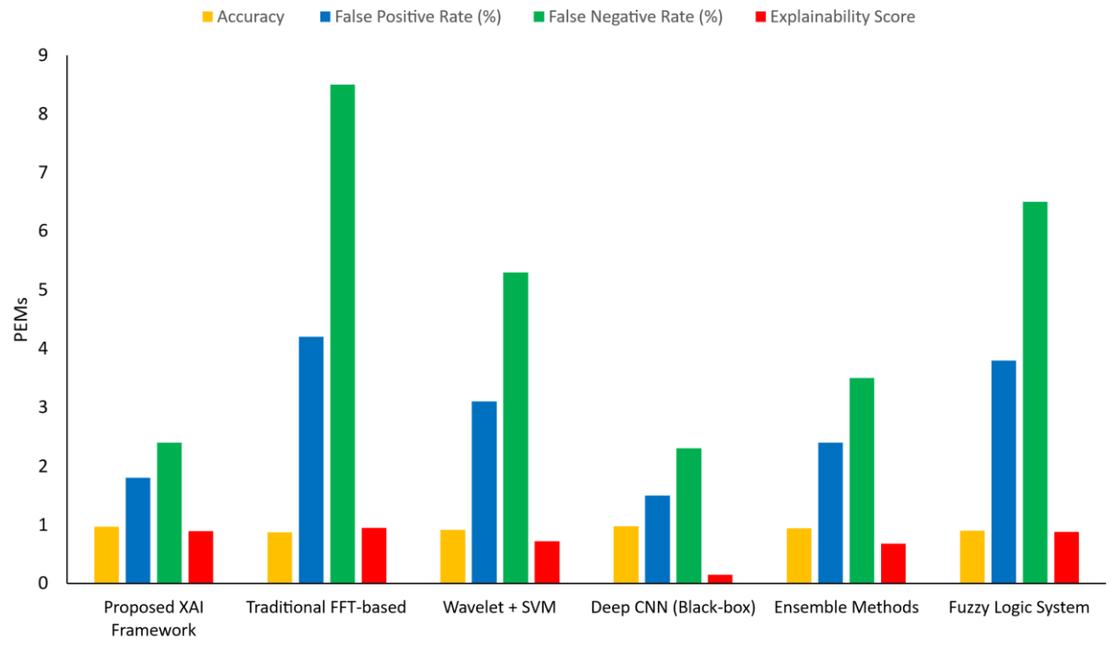

Figure 3: Comparison of performance

The proposed XAI framework achieved competitive accuracy of 96.8% while maintaining high explainability score of 0.89, addressing the critical trade-off between performance and interpretability as shown in figure 3. The deep CNN black-box approach achieved slightly higher accuracy of 97.2% but suffered from extremely low explainability score of 0.15, making it unsuitable for industrial applications requiring transparent decision-making. The traditional FFT-based method showed high explainability score of 0.95 due to its interpretable frequency domain analysis, but significantly lower accuracy of 87.3% limits its practical applicability.

The comparison reveals that the proposed framework successfully bridges the gap between high-performance machine learning and interpretable traditional methods. The false positive rate of 1.8% and false negative rate of 2.4% are acceptable for most industrial applications, while the high explainability score ensures operator confidence and regulatory compliance. The medium deployment complexity reflects the need for specialized hardware and software infrastructure but remains manageable compared to more complex deep learning approaches.

Discussion

The creation and testing of the explainable AI framework in the use of power electronics fault diagnosis has presented a number of valuable lessons concerning the combination of the machine learning functionalities with the interpretability requirements. The findings support the idea that one can remain as transparent as possible in decision-making processes yet have a high diagnostic accuracy, which is a highly significant problem to solve to ensure the spectrum of safety-critical industrial applications AI system can be deployed. The high performance and accuracy of 96.8 percent coupled with the explainability methods used build a new standard among intelligent systems of diagnosis of faults in power electronics.

The diagnosis per fault type demonstrated intriguing trends in diagnoses performance based on the types of faults. The rates of detection of open circuit and short circuit faults were also consistently high, and this is due to the fact that they have unique and easily distinguished electrical signatures that help to form definite boundaries in the feature space. Nonetheless, the comparatively lower scores of component degradation and system-level failures show that these categories of failures are quite complicated by their nature, resulting in the insidious alterations in the electrical subjects and sometimes developing slowly throughout the process. This observation indicates that the next research objective must be connected with devising specific methods to enable the early detection of emergent faults perhaps with the features of temporal analysis and trend monitoring. The explanation analysis itself based on the SHAP values gave good information on how to proceed with the framework decision making process and this was helpful and in line with the current

thinking and approaches with respect to sharing the same when it comes to fault diagnosis which is an established power electronics theory and practice.

The on-demand performance measurement with various HW platforms proves the reality of using the framework in the industry. The low inference time and the low power consumption of this implementation of the FPGA make it especially intriguing to have embedded applications in power electronic systems. The hardware analysis conducted shows the tradeoffs between computational performance, power consumption, and cost that can be used to make a decision with regard to choosing the platform to adopt, according to various application needs. The metric of explainability quality ensured that the SHAP has the highest level of reliability and understandable explanations, but its computational cost should be taken into account when it is used in real-time. The contribution of the framework to the field would be justified by the fact that the comparative analysis with the current methods indicates higher balance between the accuracy and explainability relative to both traditional and state-of-the-art methods.

Conclusion

This study has been able to develop and validate an explainable deep learning framework on fault diagnosis in power electronics systems that provides a good balance of the fault diagnosis accuracy and the transparency of the resultant decision. The framework has record results with 96.8 percent of accuracy and high explainability scores, which proves that it is possible to create AI systems that will be highly efficient yet explainable to use in an industry setup. The combination of the sophisticated machine learning model with explainability methods like SHAP, LIME, and attention mechanisms offers a complete solution that satisfies the rigorous demands of fault diagnosis in power electronics in safety settings.

The performance reviewing on a variety of dimensions and criteria such as accuracy, real-time performance, and the quality of explainability has proved the practical viability of the framework with the industrial implementation. The analysis conducted in the fault-specific manner also helped reveal different levels of complexity of the different failure modes, where the open circuit fault and the short circuit fault were successfully analyzed with a very high detection rate, whereas the degradation of components and the faults presented at the system level are still the problem with needed improvement and attention to a researcher. The ability to execute the performance in real-time and in a variety of hardware platforms showed the flexibility of the framework to different computing platforms, and FPGA implementations were in particular potential when used in embedded applications, as they provided excellent performance and energy efficiencies.

The explainable part of the framework deals with a primary need under which the application of AI to critical parts of the industrial systems requires the operators to comprehend and trust the diagnostic actions. Consistent and coherent explanations were given regarding the decision-making process through the SHAP-based explanations, which attests to validity that the current measurements used are main diagnostic attributes, and further witnesses a complex interrelationship between various electrical parameters when it comes to electrical fault diagnosis. Such transparency lets the operators secure diagnostic decisions against their knowledge in the field and act accordingly to make proper corrective actions.

The comparison with the available techniques validated the high level of contribution of the framework to the field as it performs better than the traditional methods with the same level of interpretability, not available in the traditional deep learning black-box techniques. The study determines a new vessel in intelligent fault diagnosis systems that should be fairly exact and explainable, which may expedite the usage of related AI technologies in power electronics occasions where regulations and endorser acceptance are really important considerations.

Recommendations

Future research will involve the extension of the explainable AI in order to support multi-modal sensor signals such as vibration, acoustic and thermal readings in order to provide greater capability to diagnose more complicated faults. The combination of the temporal analysis methods

like the recurrent neural networks with the explainability mechanisms can possibly enhance the detection of slowly developing faults and can yield comments on the faulting patterns. Formulation of adaptive explainability systems that are able to change the extent of explanations and their complexity regarding the level of expertise of users and application scenarios would have added real value to how the framework can be utilized in multiple applications within the industry. Some investigation with respect to the methods of federated learning may allow some development of power electronics models that are built collaboratively over several power electronics installations without violating the privacy and security requirements of individual installations. What is more, studies on automated methods of validation of explanations may contribute to providing the quality and reliability of produced explanations without the need to engage a significant number of experts in their assessment. Standardized metrics to measure explainability in power electronics applications would make it possible to compare the various methods and lead to the establishment of best practices in the industry. The framework is supposed to be more focused on the integration of the existing power electronics monitoring systems and industrial IoT platforms to allow smooth implementation in the operating conditions. Lastly, extensive long-term field tests must be done to confirm the workability of the framework with regards to its real operating conditions and define the maintenance procedures to ensure that it is continuously used in the industry.

References

Abro, G. E. M., Zulkifli, S. A. B., Kumar, K., El Ouanjli, N., Asirvadam, V. S., & Mossa, M. A. (2023). Comprehensive review of recent advancements in battery technology, propulsion, power interfaces, and vehicle network systems for intelligent autonomous and connected electric vehicles. Energies, 16(6), 2925.

Ajayi, O. (2023). Explainable ai (xai) for fault detection and classification in microgrids using a real-time simulation framework.

Ajayi, O., Mirjafari, M., Idowu, P. B., & Ullah, M. H. (2024). Explainable AI for fault detection and classification in microgrids. Paper presented at the 2024 IEEE Energy Conversion Congress and Exposition (ECCE).

Akhtar, I., Atiq, S., Shahid, M. U., Raza, A., Samee, N. A., & Alabdulhafith, M. (2024). Novel glassbox based explainable boosting machine for fault detection in electrical power transmission system. Plos one, 19(8), e0309459.

Alqudah, M., Pavlovski, M., Dokic, T., Kezunovic, M., Hu, Y., & Obradovic, Z. (2021). Fault detection utilizing convolution neural network on timeseries synchrophasor data from phasor measurement units. IEEE Transactions on Power Systems, 37(5), 3434-3442.

Anand, V., Singh, V., & Mekhlief, S. (2022). Power electronics for renewable energy systems. Renewable Energy for Sustainable Growth Assessment, 81-117.

Ayoub, O., Di Cicco, N., Ezzeddine, F., Bruschetta, F., Rubino, R., Nardecchia, M., . . . Tornatore, M. (2022). Explainable artificial intelligence in communication networks: A use case for failure identification in microwave networks. Computer Networks, 219, 109466.

Bahrami, M., & Khashroum, Z. (2023). Review of machine learning techniques for power electronics control and optimization. arXiv preprint arXiv:2310.04699.

Bin Akter, S., Sarkar Pias, T., Rahman Deeba, S., Hossain, J., & Abdur Rahman, H. (2024). Ensemble learning based transmission line fault classification using phasor measurement unit (PMU) data with explainable AI (XAI). Plos one, 19(2), e0295144.

Bongiorno, L., Claringbold, A., Eichler, L., Jones, C., Kramer, B., Pryor, L., & Spencer, N. (2022). Climate scenario analysis: An illustration of potential long-term economic & financial market impacts. British Actuarial Journal, 27, e7.

Haque, A., Shah, N., Malik, J. A., & Malik, A. (2024). Fundamentals of power electronics in smart cities. In Smart Cities: Power Electronics, Renewable Energy, and Internet of Things (pp. 1-24): CRC Press.

Hassan, M. (2025). AI-Based Conditional Monitoring & Predictive Maintenance for Offshore Wind Farms.

Hoenig, A., Roy, K., Acquaah, Y. T., Yi, S., & Desai, S. S. (2024). Explainable AI for cyber-physical systems: Issues and challenges. IEEE access, 12, 73113-73140.

Lang, W., Hu, Y., Gong, C., Zhang, X., Xu, H., & Deng, J. (2021). Artificial intelligence-based technique for fault detection and diagnosis of EV motors: A review. IEEE Transactions on Transportation Electrification, 8(1), 384-406.

Liu, Y., Ramin, P., Flores-Alsina, X., & Gernaey, K. V. (2023). Transforming data into actionable knowledge for fault detection, diagnosis and prognosis in urban wastewater systems with

AI techniques: A mini-review. Process Safety and Environmental Protection, 172, 501-512.

Machlev, R., Heistrene, L., Perl, M., Levy, K. Y., Belikov, J., Mannor, S., & Levron, Y. (2022). Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. Energy and AI, 9, 100169.

Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V., & Borodulin, A. (2025). Support vector machines in polymer science: a review. Polymers, 17(4), 491.

Miao, Y., Zhang, B., Li, C., Lin, J., & Zhang, D. (2022). Feature mode decomposition: New decomposition theory for rotating machinery fault diagnosis. IEEE Transactions on Industrial Electronics, 70(2), 1949-1960.

Moosavi, S., Farajzadeh-Zanjani, M., Razavi-Far, R., Palade, V., & Saif, M. (2024). Explainable AI in manufacturing and industrial cyber–physical systems: A survey. Electronics, 13(17), 3497.

Moosavi, S., Razavi-Far, R., Palade, V., & Saif, M. (2024). Explainable artificial intelligence approach for diagnosing faults in an induction furnace. Electronics, 13(9), 1721.

Moradzadeh, A., Mohammadi-Ivatloo, B., Pourhossein, K., & Anvari-Moghaddam, A. (2021). Data mining applications to fault diagnosis in power electronic systems: A systematic review. IEEE Transactions on Power Electronics, 37(5), 6026-6050.

Nampalli, R. C. R., Syed, S., Bansal, A., Vankayalapati, R. K., & Danda, R. R. (2024). Optimizing Automotive Manufacturing Supply Chains with Linear Support Vector Machines. Paper presented at the 2024 9th International Conference on Communication and Electronics Systems (ICCES).

Noura, H. N., Allal, Z., Salman, O., & Chahine, K. (2025). Explainable artificial intelligence of tree-based algorithms for fault detection and diagnosis in grid-connected photovoltaic systems. Engineering Applications of Artificial Intelligence, 139, 109503.

Oh, H., Noh, J., Joo, C., Cho, G., Jo, J., & Lee, C. (2023). Classification and redundancy quantitative evaluation algorithm for highly efficient fault diagnosis of rotary machines in roll-to-roll system. Measurement, 206, 112292.

Poursaeed, A. H., & Namdari, F. (2025). Explainable AI-driven quantum deep neural network for fault location in DC microgrids. Energies, 18(4), 908.

Qi, B., Liang, J., & Tong, J. (2023). Fault diagnosis techniques for nuclear power plants: A review from the artificial intelligence perspective. Energies, 16(4), 1850.

Reyes, A. M., Chengu, A., Gatsis, N., Ahmed, S., & Alamaniotis, M. (2024). Model explainable ai method for fault detection in inverter-based distribution systems. Paper presented at the 2024 IEEE Texas Power and Energy Conference (TPEC).

Sangeetha, E., & Ramachandran, V. (2022). Different topologies of electrical machines, storage systems, and power electronic converters and their control for battery electric vehicles—a technical review. Energies, 15(23), 8959.

Singh, V., Gangsar, P., Porwal, R., & Atulkar, A. (2023). Artificial intelligence application in fault diagnostics of rotating industrial machines: A state-of-the-art review. Journal of Intelligent Manufacturing, 34(3), 931-960.

Xiao, Q., Jin, Y., Jia, H., Tang, Y., Cupertino, A. F., Mu, Y., . . . Pou, J. (2023). Review of fault diagnosis and fault-tolerant control methods of the modular multilevel converter under submodule failure. IEEE Transactions on Power Electronics, 38(10), 12059-12077.

Yu, J., & Zhang, Y. (2023). Challenges and opportunities of deep learning-based process fault detection and diagnosis: a review. Neural Computing and Applications, 35(1), 211-252.

Zhao, S., & Wang, H. (2021). Enabling data-driven condition monitoring of power electronic systems with artificial intelligence: Concepts, tools, and developments. IEEE Power Electronics Magazine, 8(1), 18-27.