International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 12, No. 3, 2025, pp. 394-404.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Hybrid Knowledge-Based and Data-Driven Explainable AI Architectures for Intelligent Fault Diagnosis in Cyber-Physical Systems

Dr.V.R. Sadasivam [a], Dr.H. Shaheen [b], Kakarla Hari Kishore [c]

[a] *Professor, Department of Information Technology, K.S. Rangasamy College of Technology, Tiruchengode, India. E-mail: sadasivam@ksrct.ac.in, Orcid: https://orcid.org/0000-0001-7443-8046*
[b] *Course Leader & Senior Lecturer, Department of Computing and Engineering, University of West London, Rak Branch Campus, UAE. Email: h.shaheen@uwl.ac.ae, Orcid: https://orcid.org/0000-0003-3544-5424*
[c] *Professor, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India. E-mail: kakarla.harikishore@kluniversity.in*

Abstract

The proposed study will design a hybrid, knowledge-based and data-driven, explainable artificial intelligence (XAI) architecture for intelligent fault diagnosis in cyber-physical systems (CPS), aiming to enhance diagnostic accuracy, interpretability, and reliability in complex industrial settings. The approach presented, combining machine learning models with rule-based reasoning and explainability mechanisms, is proposed in a hybrid framework. The sensor data measured on CPS components is preprocessed and features extracted, then fault classification is performed on data-driven models developed with CNN, LSTM, and Random Forest. Model predictions are authenticated and improved by a knowledge-based module that uses expert rules, and explainability methods, such as SHAP, are used to understand the effects of input features on diagnostic results. The experimental data show that the offered hybrid XAI model outperforms traditional methods across all evaluation metrics. The model has an accuracy of 96.3%, which is higher by 94.1% and 91.2%, respectively, compared to LSTM, CNN, and Random Forest. On the same note, the hybrid method yields a precision of 95.4%, a recall of 94.8%, and an F1-score of 95.1%, which indicates better classification balance and lower false positives and false negatives. These statistical gains confirm that the combination of domain knowledge and data-driven learning increases predictive performance and strength. To sum up, the suggested hybrid architecture is an effective solution in that it satisfies the requirements of the two fields: accuracy and understanding in CPS fault diagnosis. The framework not only enhances the performance of diagnostic processes but also provides clear, operational guidance to system operators through explainable AI and knowledge-based reasoning. This both qualifies it as a good solution for safety-critical industrial systems and offers strong prospects for future development in intelligent maintenance systems.

Keywords: Cyber-Physical Systems (CPS), Fault Diagnosis, Explainable Artificial Intelligence (XAI), Hybrid AI Architecture, Knowledge-Based Systems, Data-Driven Models, Predictive Maintenance.

DOI: 10.71086/IAJSE/V12I3/IAJSE1275
Received: 17 July 2025/Revised: 25 August 2025/Accepted: 12 September 2025/Published: 30 September 2025

Introduction

Cyber-Physical Systems (CPS) is the intersection of the physical processes, computational intelligence, communication networks, and control mechanisms, which may be used to facilitate intelligent automation in the contemporary industrial setting (Sanwar, 2024; Kasoju, 2024). The industry 4.0 paradigm includes CPS technologies as the core of such sophisticated manufacturing-related ecosystems as smart grids, intelligent transportation systems, and industrial automation. These systems are based on the monitoring of constant work with the sensors, in-built controllers, and integrated devices to ensure the efficiency of work and safety. Nevertheless, the growing complexity, heterogeneity, and interconnectivity of CPS infrastructures are also posing major problems when it comes to the reliability of the system and the management of faults. Malfunctions in CPS elements (sensors, actuators, communication modules, or control systems) may spread quickly to other interconnected subsystems, resulting in performance failure, unplanned downtime, and safety risks. Therefore, smart fault diagnosis is nowadays an urgent field of study, driven by the goal of detecting anomalies in the system, revealing faults at an early stage, and assisting maintenance decision-making.

The conventional methods of fault diagnosis were mostly either model-based or rule-based, based on physical system models or expert knowledge. However, it is the case that although the methods are interpretable and transparent, these approaches tend to be inflexible in dynamic environments and are prone to failing to observe nonlinear system behavior. In recent years, the application of artificial intelligence (AI) methods based on data, including machine learning and deep learning, has proved to have a high potential in the field of fault detection and classification (Moosavi et al., 2024; Lin et al., 2024). These methods are based on large masses of sensor data to automatically extract patterns related to normal and abnormal states of systems. Most AI models, especially deep learning models, are very predictive, but are black box in nature, providing low interpretability and transparency. This inability to explain makes deployment in safety-critical CPS settings a major challenge, and operators have to know how the automated decisions were made. Explainable Artificial Intelligence (XAI) is a potential solution to overcome the limitations of the transparency of traditional AI systems.

The purpose of the XAI techniques is to ensure that the AI decision-making can be explained to human experts by giving them interpretable explanations of the model predictors. Nevertheless, even pure data-driven XAI methods can still not capture useful domain knowledge that can be found in rules in engineering, physical models, and expert information (Wang et al., 2022). By combining the knowledge-based reasoning with data-driven learning mechanisms, it is thus possible to have a more reliable and interpretative framework of fault diagnosis. Knowledge-based systems together with data-driven models provide the future of hybrid architectures that can be used to improve the quality of the diagnoses and their interpretation. Using domain knowledge, rule-driven reasoning, and cause associations in addition to machine learning models, hybrid XAI engines are able to give significant explanations while also preserving excellent predictive capabilities. These architectures may be used to enable decision-making processes by system operators, increase the level of trust in AI systems, and make systems more reliable in complex CPS settings. The main goal of the study is to come up with an explanatory data-driven yet knowledge-based hybrid AI architecture to propel the intelligent fault diagnosis in cyber-physical systems (Liu et al., 2024; Azari et al., 2022). The suggested framework will combine domain knowledge representations, rule-based reasoning, and machine learning algorithms and explainability mechanisms to promote the interpretability, reliability, and accuracy of fault diagnosis processes.

The study aims to come up with an architecture that would be able to identify and categorize faults, besides offering human-readable explanations that would assist the maintenance engineers and system operators in their decision-making. Although the state of AI-based fault diagnostics has advanced considerably, there are some issues that are still present when it comes to the implementation of intelligent diagnostic systems in the cyber-physical setting (Ahmed et al., 2024). First, most of the current fault diagnosis models are purely based on data-driven machine learning methods that tend to need large sets of labeled data and might not be generalizable in the presence of varying system circumstances. Second, the majority of deep learning models are not transparent, and thus, engineers can hardly interpret predictions and rely on automated decisions in deep-learning-based safety-critical CPS applications (Mishra et al., 2024). Third, diagnostic methods based on knowledge, which include expert rules and physical models, offer an interpretable model, but are relatively often not scalable and adaptable in large-dimensional sensor data. Recent studies have examined the explainable AI methods, including the feature attribution methods, interpretable models, and visualization methods to enhance transparency (Hoenig et al., 2024; Nazir, 2022). Such techniques are, however, generally used once the model gets trained and can fail to incorporate the knowledge about domain information in the learning process. Moreover, there are few studies that concentrate on the development of unified structures that tightly integrate knowledge-based reasoning, data-driven learning, and explainability systems in a

single fault diagnosis system. Thus, one can say that there is an evident requirement for hybrid XAI architectures that can fill the gap in interpretability and predictive performance of cyber-physical system diagnostics.

The hypotheses of the research are that the combination of knowledge-based reasoning with data-driven machine learning models as a part of the explainable AI framework can contribute significantly to the reliability, interpretability, and diagnostic accuracy of fault detection systems in the cyber-physical environment. The hybrid architectures are able to present a deeper explanation of anomalies that occur in the system with the combination of domain knowledge and learned data patterns, improve fault detection performance, and make the users believe more in automated diagnostic systems.

1. It offers a hybrid explainable artificial intelligence framework that combines the reasoning based on knowledge with a machine learning based on data to allow intelligent and interpretable fault diagnosis in cyber-physical systems.
2. The research has the explainability mechanisms integrated into the diagnostic architecture to enable the system to have clear and human-interpretable information with regard to the model predictions, which are aimed at enhancing trust and decision-making among the operators and maintenance engineers of the systems.
3. To create an efficient mechanism of monitoring and maintenance of complex cyber-physical infrastructures, the proposed approach provides better accuracy, robustness, and reliability of fault detection and classification based on domain knowledge combined with sensor-based data analytics.

The structure of the paper is as follows: Section 1 is the background, gap in the research, objectives, and contribution of the study. Section 2 provides an overview of the literature on the field of fault diagnosis, data-driven, knowledge-based, and explainable AI methodologies. Section 3 introduces the proposed hybrid XAI methodology, such as the system architecture, data processing, machine learning models, and knowledge-based reasoning, as well as more explainability mechanisms. Section 4 outlines the setting of the experiment, data, and implementation. The results and a comparison of the performance are discussed in Section 5. Lastly, Section 6 summarizes the paper with the essential findings, limitations, and recommendations for further research.

Literature Review

Fault diagnosis is paramount to the reliability and safety of contemporary cyber-physical systems (CPS), which closely combine physical processes with computational and communication units (Mourtzis & Angelopoulos, 2024; Vaughn et al., 2024). Initial methodologies of fault detection in CPS primarily depended on more established and historically model-based methods in which mathematical models of system behavior were employed to identify deviations that had the potential to represent faults (Mhapsekar et al., 2024). Other approaches employed in industrial control systems were observer-based monitoring, parity space techniques, and residual analysis. Although these methods offered interpretable findings and well-founded theories, the success of the method in most cases relied on the presence of accurate system models. In complex CPS systems like smart manufacturing, energy systems, and automated production lines, it is more challenging to develop accurate mathematical models because of nonlinear dynamics, uncertainty in the environment, and the complexity of the system. As a result, model-based approaches tend to be inefficient in dealing with large-scale industrial CPS having heterogeneous sensors and a changing operation environment (Zhou et al., 2023).

As the industrial data produced by sensor networks and Internet of Things devices expands at a rapid pace, machine learning-based diagnostic methods have become of great importance to the CPS fault diagnostics. Patterns: Data-driven models can be used to automatically learn patterns using large datasets and identify anomalies without explicit system models. Other methods of traditional machine learning, like support vector machines, decision trees, and random forests, have been extensively utilized in condition monitoring and classification of system faults. More recently, convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks based on deep learning methods have performed well in deriving complex temporal and spatial information in industrial sensor data (Kidambi Raju et al., 2023). The methods contribute a lot to the accuracy of fault detection and the flexibility of complex CPS settings. Nevertheless, the majority of purely data-driven solutions are operated as black-box models and, as such, it is hard to tell how the engineers and system operators of the particular system generated the diagnosis decision.

Modern Industry 4.0 environments are also closely related to predictive maintenance strategies, which are also based on data-driven fault detection methods (Abdelillah et al., 2023). Predictive maintenance is designed to identify when the equipment is in the degradation process and anticipate its possible breakdown before it actually happens, hence

reducing maintenance and downtime (Joshi & Tiwari, 2024). Several different deep learning architectures have been used to solve predictive maintenance tasks, such as autoencoders to detect anomalies, deep belief networks to extract features, and hybrid CNN-LSTMs to predict temporal faults. These techniques make use of huge amounts of sensor data and determine unobtrusive patterns that can signal possible failures. Although these models have powerful predictive forces, they are usually opaque, with unclear explanations, and, therefore, may not be highly acceptable in safety-critical CPS environments where human operators need clear reasons why an automated decision should have been made.

Another way is the knowledge-based diagnostic systems, which include domain knowledge and explicit reasoning in fault diagnosis. These systems are normally represented by rule-based reasoning, expert systems, or ontology-based systems to represent system knowledge and deduce fault conditions. Rule-based systems represent the expert knowledge as a logical rule that relates symptoms to possible faults and allow for transparent and understandable decisions. Expert systems have managed to be effective in industrial monitoring, in which the diagnostic process is determined by preset rules depending on sensor values and system conditions. Nonetheless, knowledge-based systems are usually not able to cope with the high volume of real-time data and with the need to adapt to new or unfamiliar patterns of faults. Their functioning is strongly contingent on the fullness and the accuracy of expert knowledge that may not be easy to maintain in complicated CPS environments.

In the past few years, explainable artificial intelligence (XAI) has tried to resolve the lack of transparency of data-driven models in industrial systems (Almuqren et al., 2023; Capuano et al., 2022). SHAP (SHapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations) have also been extensively employed to explain, post hoc, the output of machine learning models by attributing the contribution of a single feature to the model output. Neural networks that are based on attention also give information on aspects of input data that affect the decisions made by the model. Although these techniques enhance interpretability, the majority of current research will use explainability techniques, which are introduced after the prediction phase and do not comprehensively incorporate the domain knowledge into the learning process. Consequently, it can be concluded that there is still a big research gap to create hybrid diagnostic architectures that are capable of successfully integrating knowledge-based reasoning with data-driven learning and include in-built explainability features. The solution to this gap will contribute to minimizing both the accuracy and interpretability of intelligent fault diagnosis systems in cyberspace physical settings.

Available approaches are either data-driven models that can be highly accurate or knowledge-based systems that can be interpreted. Explainable AI enhances transparency, yet it is largely post-hoc. This shows a blank in the hybrid architectures of combining information-driven learning, knowledge reasoning, and an in-built explainability in achieving precise and understandable fault examination in cyber-physical systems.

Methodology

System Architecture Overview

The suggested approach presents an explainable artificial intelligence (XAI) hybrid framework of intelligent fault diagnosis in cyber-physical systems (CPS). The architecture combines knowledge-based reasoning and data-driven machine learning models to improve the accuracy and explainability of the results of the diagnostic process. The framework comprises various modules that are connected and interact with each other, such as data acquisition, preprocessing, feature extraction, machine learning based fault classification, knowledge-based reasoning, and explainability analysis. First, sensor data taken out of CPS parts are processed to eliminate noise and extract features. Machine learning algorithms then use these attributes to detect failures in the system. Simultaneously, a knowledge-based module of expert rules and domain knowledge checks and refines the predictions made by the models that are data-driven. Lastly, an explainability module is used to decode the results of the model and offer explanations that can be understood to make diagnostic decisions.

Fig. 1 shows the design of the proposed hybrid explainable artificial intelligence framework that will be used in intelligent fault diagnosis in cyber-physical systems. The framework starts at the data acquisition phase, during which sensor data are obtained on components of CPS and are processed by preprocessing and feature extraction methods. A machine learning model based on the data is then used to identify and categorize system faults based on the processed information. Simultaneously, the knowledge-based reasoning module takes into consideration expert rules and domain knowledge to verify and interpret the diagnostic results. An explainability module further evaluates the model predictions using interpretability methods to determine the most important features that determine the decisions. Lastly, the system produces diagnostic outputs, understandable insights, and information to support decision-making by the system operators to be informed of faults and manage them effectively.

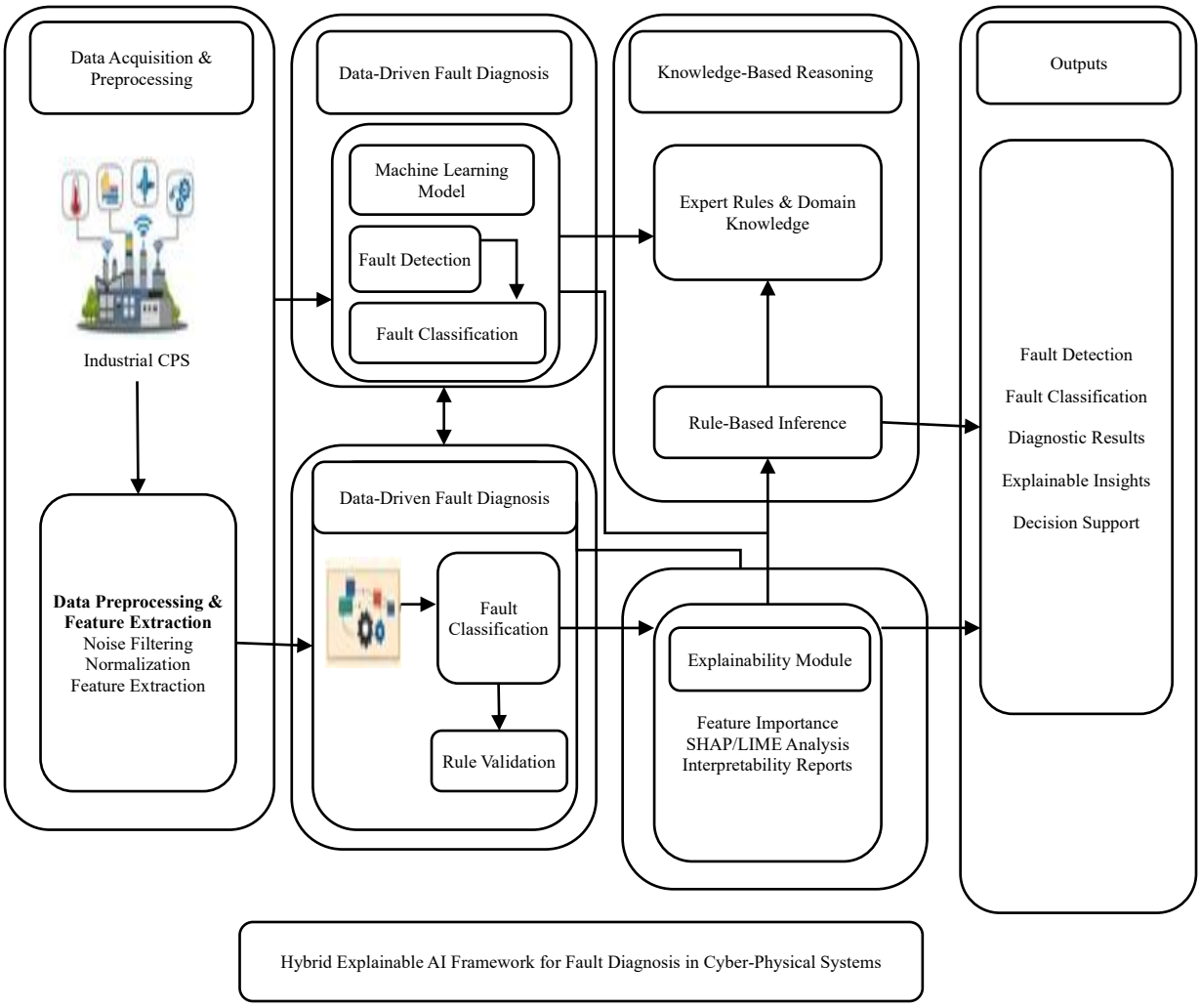


Fig. 1: Architecture of the Hybrid Knowledge-Based and Data-Driven Explainable AI Framework for Fault Diagnosis in Cyber-Physical Systems

Data Acquisition and Sensor Integration

Sensors, embedded devices, and communication networks are used to produce vast amounts of real-time running data in cyber-physical systems. The data gathered in this research is based on the various elements of CPS, including sensors, actuators, controllers, and network communication units. The data accumulated covers the operational parameters, which include temperature, vibration signal, the level of pressure, the level of voltage, current, and the system status indicators. These measurements reflect cases of normal operation and fault. The industrial IoT platforms ensure the continuous acquisition of data, which guarantees real-time monitoring and high-resolution data capture in the diagnostic analysis of information.

Data Preprocessing and Feature Extraction

Sensor data has numerous noises, missing values, and duplicate information, which have the possibility to influence the performance of machine learning algorithms. Thus, preprocessing methods are used to achieve the quality and reliability of data. This is followed by noise filtering, normalization, and elimination of outliers. Then, the feature extraction methods are used to obtain meaningful representations out of the raw sensor signals after preprocessing. Mean, variance, kurtosis, and skewness are expressed as statistical features, and time-domain and frequency-domain characteristics are provided where possible. The features will be used to give informative signs to differentiate between normal and malfunctioning system behavior.

Data-Driven Fault Diagnosis Model

The machine learning part of the proposed framework uses machine learning algorithms to identify and categorize defects in the CPS components. The training of supervised learning models is done on labeled datasets, which include the samples of normal operation as well as various fault conditions. Depending on the complexity of the dataset, it is possible to use algorithms like decision trees, random forests, support vector machines, or deep neural networks. The trained model is one that takes patterns and correlations of sensor features to find the correct classification of faults. This empirical methodology allows the system to retrieve nonlinear relationships and multifaceted system dynamics that are hard to model in the traditional method.

The machine learning model is trained to adopt a mapping between the feature set extracted and the fault class.

$$y = f(F; \theta) \quad (1)$$

In Equation (1), were

$F = \{f_1, f_2, \ldots, f_m\}$ is the set of features extracted from sensor data,θ is the learned model parameters, and y represents the fault type of the cyber-physical system that is predicted.

Knowledge-Based Reasoning Module

A knowledge-based reasoning module is added to the architecture to address the need to enhance interpretability and add expert knowledge. The module is composed of rule-based systems and knowledge of the domain based on the engineering guidelines, maintenance manuals, and expert knowledge. The knowledge base comprises fixed fault rules, causal relations, and operational thresholds of various CPS components. On identifying a possible fault by the machine learning model, the knowledge-based module is used to test the prediction by logical inference and rule validation. The hybrid method is beneficial to validate diagnostic findings, minimize false positives, and offer explainable justifications of the detected defects.

Hybrid Decision Fusion

The decision made by the machine learning prediction and the knowledge-based reasoning is the final diagnostic decision.

$$D_{final} = \alpha D_{ML} + (1 - \alpha) D_{KB} \quad (2)$$

In Equation (2), were

D_{ML} is the forecast of the machine learning model.

D_{KB} is the result of a knowledge-based rule system, and

$\alpha$is a weighting factor that balances between the effect of data-driven learning and professional knowledge.

Explainability and Interpretation Mechanism

The explainability module ensures greater transparency because it attempts to explain predictions made by the machine learning models. To identify the most important features that impact each diagnostic decision, explainable AI methods like feature importance analysis, SHAP values, or local interpretable explanations are employed. These descriptions permit the system operators and engineers to see why a certain fault was identified and which of the system parameters played the most important role in the prediction. Complex AI models and the way humans make decisions are therefore bridged by the explainability layer.

Model Evaluation and Performance Metrics

The usefulness of the suggested hybrid XAI architecture is measured with the help of the common performance measures of the classification work. The diagnostic performance of the model is evaluated using metrics like accuracy, precision, recall, F1-score, and confusion matrices. Besides predictive accuracy, interpretability and reliability are also taken into account during the evaluation process. It is possible to perform a comparative experiment with only data-driven models and the suggested hybrid architecture to prove the advantages of the fault diagnosis quality and explainability.

Implementation Framework

The suggested system can be deployed with the help of recent machine learning and data analytics systems. House Data: The programming environment to conduct data preprocessing and model training is Python, along with its libraries TensorFlow, Scikit-learn, and XAI toolkits. Knowledge-based reasoning module is implementable by rule

engines or expert system frameworks. A combination of these components creates a diagnosing platform that is holistic enough to facilitate smart monitoring, predictive maintenance, and failure analysis within the cyber-physical systems.

Experimental Setup

The experimental design will be aimed at examining the efficiency of the proposed hybrid explainable artificial intelligence structure for intelligent fault diagnosis in cyber-physical systems. The test is aimed at measuring the performance of the suggested framework with respect to the accuracy of classification, reliability, and interpretability. The hybrid model combines both the knowledge-based reasoning and data-based learning to identify faults in industrial machines. Testing of experiments is done with sensor data gathered in the cyber-physical environments, and the results of the proposed system are examined through standard machine learning evaluation measures.

Dataset

The Case Western Reserve University (CWRU) Bearing Dataset is the experiment dataset commonly used in cyber-physical systems research on fault diagnosis. The database has 10,000 samples of bearings that are in normal and faulty conditions in terms of inner race faults, outer race faults, and ball defects of different degrees of severity. The samples include the records of vibration signals at 12 kHz, which give abundant time-series information to be used in both data-driven learning and knowledge-based reasoning.

Implementation Environment

The hybrid XAI architecture proposed is developed on the Python programming language because it has a wide range of support for machine learning and data analysis libraries. The development of deep learning models is based on such frameworks as TensorFlow and PyTorch that offer effective means of building and training neural network models. The common Python libraries such as NumPy, Pandas, and Scikit-learn are used in the data preprocessing and feature engineering.

The experiments are also applied on a computer system with a high-performance processor and enough memory to work with large sensor data sets. The hardware setup usually consists of an Intel Core i7 processor, 16 GB RAM, and a dedicated graphics card to accelerate the design of deep learning. The environment of implementation facilitates effective hybrid architecture training and testing, which allows the assessment of the predictive performance and explainability features.

Training Details and Hyperparameters

Table 1: Training Configuration and Hyperparameters of the Models

Model	Architecture / Layers	Optimizer	Learning Rate	Batch Size	Training Iterations	Key Parameters
CNN	3 Convolutional layers with ReLU activation + Fully Connected layer	Adam	0.001	64	50 epochs	Kernel size = 3×3, Dropout = 0.5
LSTM	2 LSTM hidden layers with 128 units each	Adam	0.001	32	50 epochs	Sequence length = 100, Dropout = 0.3
Random Forest	Ensemble of Decision Trees	Gini-based splitting	0.01	128	200 trees	Max depth = 15, Min samples split = 2
Hybrid XAI Model	CNN + LSTM + Knowledge-based reasoning	Adam	0.001	64	50 epochs	Weighted voting = 0.5 (ML) + 0.5 (Rule-based)

The training configuration and hyperparameters used to train the models in the proposed framework are represented in table 1. The CNN and LSTM models are optimized with Adam, and the learning rate is 0.001 over 50 epochs; the Random Forest model is trained with 200 trees and a maximum depth of 15. The Hybrid XAI model suggested is a combination of the results of CNN and LSTM, and a knowledge-based reasoning unit based on a voting strategy with weighted voting to enhance the accuracy of fault diagnostics and interpretability.

Evaluation Metrics

In order to measure the performance of the proposed fault diagnosis framework, a number of standard classification measures are applied. These measurements are used to determine the model to predict fault conditions and differentiate them from the normal system conditions.

Accuracy is used to measure the total rightness of the classification model, to compute the ratio of the cases that are rightly predicted and the total number of cases.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

In Equation (3), TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

Precision measures the %age of correct cases of the faults predicted out of the total number of predicted cases of faults. A large value of precision shows that the model has fewer false alarms when detecting faults, as is shown in Equation (4).

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Recall is the capability of the model in the identification of the correct cases of fault in the data set. This measure shows how well the model is able to recognize the actual faults and minimize missing faults in the system as given in Equation (5).

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

F1-score offers a balanced statistic between precision and recall to use in the case of imbalanced data. The F1-score is obtained as below in Equation (6).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

Besides classification metrics, an explainability score is also viewed as a measure of how the model can give understandable explanations of its predictions. The score is obtained based on the consistency and relevance of feature importance explanation derived with XAI techniques, including SHAP and LIME. The experimental setup is effective in assessing predictive performance and interpretability by guaranteeing that the proposed hybrid architecture can be deployed in a realistic system in the cyber-physical space, where transparency and reliability are crucial values.

Results and Discussion

In this segment, experimental findings are provided, which were derived by assessing the suggested hybrid elucidating artificial intelligence framework in smart fault diagnosis within cyber-physical systems. The assessment of the proposed model is conducted on the basis of the evaluation metrics presented in the above section, such as accuracy, precision, recall, and F1-score. The findings are indicative of the efficiency of incorporating data-driven learning and knowledge-based reasoning and explainability programs to enhance the accuracy of fault detection and explainability. The experimental research results are the comparison of the proposed hybrid approach to the conventional machine learning and deep learning models to emphasize its benefits in CPS environments.

Fault Classification Performance

The effectiveness of the proposed hybrid architecture has been assessed in comparison with a number of baseline models that are most often applied in fault diagnosis systems. The same data is used in conventional machine learning algorithms and deep learning models training. The comparison is based on how each model is able to properly classify various fault conditions. The findings indicate that hybrid architecture has a high accuracy and a higher classification reliability in contrast to single models, which are based on data.

The machine-learning models like CNN, LSTM, and Random Forest exhibit excellent results in the process of identifying complicated patterns in the sensor data. The incorporation of the knowledge-based reasoning module, however, assists in refining the predictions by authenticating them by applying rules established by experts. This combination will minimize false alarms and enhance a generally better diagnostic consistency. Consequently, the hybrid proposed model is better in all the assessment indices. Table 1 indicates an example of the performance of the models.

Table 2: Performance Comparison of Fault Diagnosis Models

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	91.2%	89.8%	88.5%	89.1%
CNN	93.6%	92.1%	91.4%	91.7%
LSTM	94.1%	92.9%	92.2%	92.5%
Proposed Hybrid XAI Model	96.3%	95.4%	94.8%	95.1%

Table 2 includes the comparison of models of fault diagnosis in relation to the accuracy, precision, recall, and F1-score. Random Forest model has the highest accuracy with 91.2, and CNN and LSTM have 93.6% and 94.1% accuracy, respectively. The Hybrid XAI model delivers the strongest performance with 96.3% accuracy, 95.4% precision, 94.8% recall, and 95.1% F1-score, which proves the usefulness of integrating knowledge-based reasoning and data-driven learning to achieve the best fault diagnosis.

The findings show that the hybrid model that has been proposed is better than the models that are at the base because it combines machine learning predictions with expert knowledge reasoning.

Explainability Analysis

Besides enhancing accurate diagnosis, the offered architecture is also more interpretable as it incorporates explainable AI methods. The explainability module creates understanding of the manner in which the model makes its predictions by examining the value of individual input features. The feature importance scores achieved by using SHAP-based explanations provide insight into which sensor parameters can play the most significant role in making the decisions of classifying the faults.

As an illustration, bearing the fault detection case, the magnitude and temperature readings are determined as the most significant factors that impact the choice of the model. The contribution of feature visualization enables engineers to have a clear picture of how sensor measurements relate to the forecasted fault conditions. Interpretability is a prerequisite in industrial CPS applications that require human operators to confirm automated choices before acting appropriately.

Comparative Analysis with Existing Methods

In order to further prove the efficiency of the suggested hybrid architecture, the performance of the given solution is contrasted with the currently available methods of fault diagnosis published in the recent literature. Most of the conventional data-based models would have a high predictive accuracy, but not transparency, and knowledge-based systems are interpretable, but will not be able to handle complex data patterns. The hybrid method that is suggested in the current research is able to overcome these restrictions since it will combine the two methods into a single framework.

The comparative analysis indicates that the hybrid architecture is able to provide a balanced performance, whereby the classification accuracy is high, as well as an explanation given to every prediction performed by such an architecture. This feature will elevate user confidence in the system and will also increase its usefulness in the application of the system in industrial monitoring and predictive maintenance. These findings prove that data-driven learning, together with the techniques of knowledge-based reasoning and explainability, is a promising solution to smart fault diagnosis in the context of a cyber-physical system.

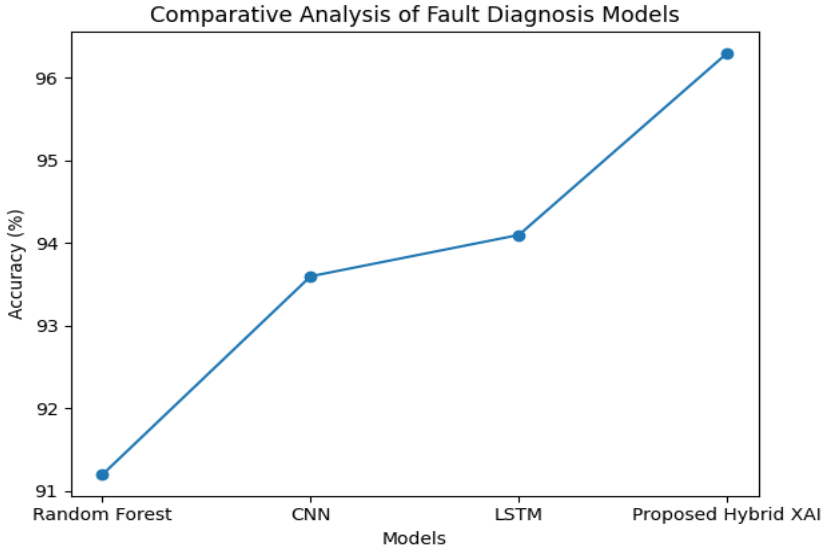


Fig. 2: Comparative Performance Analysis of Fault Diagnosis Models in Cyber-Physical Systems

Fig. 2 shows a comparative study of various machine learning models that can be applied in the diagnosis of faults in cyber-physical systems. The model of the random forest has an error rate of 91.2%, which implies that it has a strong capacity to classify fault conditions using the sensor features extracted. The Convolutional Neural Network (CNN) model results in an accuracy of 93.6, which proves the possibility of the model to detect sophisticated patterns in sensor signals. The accuracy is further enhanced with the Long Short-Term Memory (LSTM), which reaches 94.1 as it is able to learn time-dependent features that exist in time-series sensor data. The proposed Hybrid XAI model has the best accuracy of 96.3, which is better than the baseline models. This is done through better incorporation of data-driven learning alongside knowledge-based reasoning and explanatory mechanisms that improve the diagnostic accuracy as well as explainability of the model. This is evidently reflected in the results that show that the hybrid architecture offers better performance in the intelligent fault diagnosis of cyber-physical systems.

An ablation study is conducted to determine the contribution of each of the components of the proposed hybrid architecture. The accuracy of the data-driven models is 94.1% with an explainability score of 0.81, whereas the explainability score of the knowledge-based reasoning system is 0.95, and the accuracy is 88.7%. The full Hybrid XAI model offers the highest level of 96.3% accuracy and 0.92 explainability, with a clear indication that a combination of data-driven learning and knowledge-based reasoning is better in detecting faults and explaining them.

Conclusion

The findings show that the suggested hybrid explainable AI architecture is significantly more effective in terms of fault diagnosis in cyber-physical systems. The framework can provide a significant increase in the accuracy and reliability of the diagnostic outcomes by employing machine learning models with rule-based reasoning and explanatory mechanisms. Particularly, the hybrid model achieves an accuracy value of 96.3%, which is better than traditional methods, including Random Forest (91.2%), CNN (93.6%), and LSTM (94.1%). On the same note, other evaluation metrics are also improved, as precision stands at 95.4%, recall at 94.8%, and an F1-score of 95.1%, and the classification performance is well balanced and strong. These statistical improvements underline the usefulness of the data-driven learning combined with the domain knowledge to decrease the false positives and increase the consistency of fault detection. Besides, the explanation methods allow incorporating explainability capabilities into the system, which will deliver valuable insights on the model predictions to permit the engineers to discover their critical contributing features, including vibration and temperature in fault conditions. Not only does this increase the level of trust in automated decision-making, but it also informs the maintenance actions in safety-critical CPS settings. Although this is the case, there are some limitations. Even in dynamic industrial conditions, the hybrid system is still subject to the performance of both labeled datasets and expert-defined rules, which can still be in need of continuous updates. Also, the computational cost of linking multiple modules may affect real-time implementation in a resource-constrained environment. Future endeavors will be on creating adaptive knowledge bases that can use continual learning to update rules and also make the best use of model efficiency to deploy edges in real time. The combination of federated learning and digital twin technologies to improve the scalability and system-wide coordination can also be studied in the future. The extension of the analysis to multi-fault and cross-domain CPS cases will also enhance its usability and make it an all-inclusive solution to the next-generation, powerful, and smart fault diagnosing systems.

Reference

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. https://doi.org/10.3390/electronics13173497

Wang, J., Li, Y., Gao, R. X., & Zhang, F. (2022). Hybrid physics-based and data-driven models for smart manufacturing: Modelling, simulation, and explainability. *Journal of Manufacturing Systems*, *63*, 381-391. https://doi.org/10.1016/j.jmsy.2022.04.004

Lin, D., Celso Bation Co, Zeng, M., & Xu, T. (2024). Efficient Network-based Fault Detection in Elevator Vibration Signals: A Weighted Fusion Approach of Displacement and Acceleration Data. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 15*(3), 459-473. https://doi.org/10.58346/JOWUA.2024.I3.030

Sanwar, A. S. M. (2024). Explainable artificial intelligence into cyber-physical system architecture of smart cities: Technologies, challenges, and opportunities. *Journal of Electrical Systems, 20*(2), 2343–2362.

Mishra, N., Haval, A. M., Mishra, A., & Dash, S. S. (2024). Automobile Maintenance Prediction Using Integrated Deep Learning and Geographical Information System. *Indian Journal of Information Sources and Services*, *14*(2), 109-114. https://doi.org/10.51983/ijiss-2024.14.2.16

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. https://doi.org/10.1109/ACCESS.2024.3395444

Joshi, A., & Tiwari, S. (2024). Neural Network-Driven Predictive Maintenance for Gas Turbines. *International Academic Journal of Innovative Research, 11*(2), 52–57. https://doi.org/10.71086/IAJIR/V11I2/IAJIR1116

Zhou, X., Ahmed, B., Aylor, J. H., Asare, P., & Alemzadeh, H. (2023). Hybrid knowledge and data driven synthesis of runtime monitors for cyber-physical systems. *IEEE Transactions on Dependable and Secure Computing*, *21*(1), 12-30. 10.1109/TDSC.2023.3242653

Nazir, N. (2022). Intelligent Fault Diagnosis and Resilient Control of Safety-Critical Industrial Systems Using Explainable AI. *Journal of Software Engineering and Information Systems*, *7*,1-12.

Almuqren, L., Maashi, M. S., Alamgeer, M., Mohsen, H., Hamza, M. A., & Abdelmageed, A. A. (2023). Explainable artificial intelligence enabled intrusion detection technique for secure cyber-physical systems. *Applied Sciences*, *13*(5), 3081. https://doi.org/10.3390/app13053081

Kidambi Raju, S., Ramaswamy, S., Eid, M. M., Gopalan, S., Alhussan, A. A., Sukumar, A., & Khafaga, D. S. (2023). Enhanced dual convolutional neural network model using explainable artificial intelligence of fault prioritization for industrial 4.0. *Sensors*, *23*(15), 7011. https://doi.org/10.3390/s23157011

Mourtzis, D., & Angelopoulos, J. (2024). Artificial intelligence for human–cyber-physical production systems. In *Manufacturing from Industry 4.0 to Industry 5.0* (pp. 343-378). Elsevier. https://doi.org/10.1016/B978-0-443-13924-6.00012-0

Ahmed, U., Jiangbin, Z., Almogren, A., Khan, S., Sadiq, M. T., Altameem, A., & Rehman, A. U. (2024). Explainable AI-based innovative hybrid ensemble model for intrusion detection. *Journal of Cloud Computing*, *13*(1), 150. https://doi.org/10.1186/s13677-024-00712-x

Mhapsekar, R. U., Umrani, M. I., Faizan, M., Ali, O., & Abraham, L. (2024, September). Building trust in AI-driven decision making for cyber-physical Systems (CPS): A comprehensive review. In *2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA)* (pp. 1-8). IEEE. https://doi.org/10.1109/ETFA61755.2024.10710855

Abdelillah, F. M., Nora, H., Samir, O., & Sidi-Mohammed, B. (2023, November). Hybrid data-driven and knowledge-based predictive maintenance framework in the context of industry 4.0. In *International conference on model and data engineering* (pp. 319-337). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-49333-1_23

Kasoju, A. (2024). AI-Driven Anomaly Detection in Cyber-Physical Systems: A Technical Approach to Real-Time Threat Mitigation. *Iconic Research and Engineering Journals*, *8*(4), 804-817.

Liu, Y., Chen, S., Li, P., Wan, J., & Li, X. (2024). Status, challenges, and promises of data-driven battery lifetime prediction under cyber-physical system context. *IET Cyber-Physical Systems: Theory & Applications*, *9*(3), 207-217. https://doi.org/10.1049/cps2.12086

Vaughn, J., Acquaah, Y. T., & Roy, K. (2024, August). Anomaly detection in cyber-physical systems based on genetic algorithm with dynamic thresholding detection. In *2024 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)* (pp. 1-8). IEEE. https://doi.org/10.1109/icABCD62167.2024.10645284

Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access*, *10*, 93575-93600. https://doi.org/10.1109/ACCESS.2022.3204171

Azari, M. S., Flammini, F., & Santini, S. (2022, July). Improving resilience in cyber-physical systems based on transfer learning. In *2022 IEEE International Conference on Cyber Security and Resilience (CSR)* (pp. 203-208). IEEE. https://doi.org/10.1109/CSR54599.2022.9850282