**Electronics, Communications, and Computing Summit**
Vol. 2, No. 4, Oct - Dec 2024, pp. 92-98
DOI: https://doi.org/10.17051/ECC/02.04.12

# Fusion of Big Data Analytics and Deep Learning for Predictive Fault Diagnosis in Cyber-Physical Energy Systems

**Pushplata Patel**

Department Of Electrical And Electronics Engineering, Kalinga University, Raipur, India
Email: pushplata.subhash.raghatate@kalingauniversity.ac.in

**Article Info**

***Article history:***

Received : 22.10.2024
Revised : 24.11.2024
Accepted : 26.12.2024

***Keywords:***

Cyber-Physical Energy Systems,
Predictive Fault Diagnosis,
Big Data Analytics,
Deep Learning,
Smart Grid,
Anomaly Detection,
System Resilience.

**ABSTRACT**

With the integration of physical infrastructure with a state of the art computational intelligence, Cyber-Physical Energy Systems (CPES) have become a revolutionary paradigm in modern power grids. They are however complex and this makes them easily subjected to errors in diagnosis when their symptoms should be treated as quickly as possible. The research would develop a new framework of predictive fault diagnosis that integrates the features of big data analytics and deep learning which would improve system reliability and responsiveness. The architecture that is suggested consumes heterogeneous sensor data of smart meters, substations, and SCADA systems that are based on Hadoop Distributed File System (HDFS) scalable storage and Apache Spark Streaming real-time processing. These are more advanced feature engineering methods, represented by statistical aggregation, Fast Fourier Transform (FFT) and wavelet transforms, that are executed before inference of the model. The trained hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) can be used to predict fault types and estimate to time-to-failure (TTF) using historical fault-labelled data. However, compared to the preceding models, the proposed framework does not merely combine deep learning with big data infrastructure at a level that creates a real-time feedback loop, but allows wide-ranging, flexible deployment about CPES environments. Experimental analysis of a simulated IEEE 14- bus test system and over 1TB of real-world smart substation data indicates a fault detection range of 97.3 per cent and an average event latency of less than 3 seconds per fault that is far superior to standard Random Forest and Support Vector Machine (SVM) methods. These findings emphasize the effectiveness of the framework in enhancing the accuracy of fault classification, lower response time, and engage in proactive maintenance. The development of edge federated learning, explainable AI and deployment in hybrid edge cloud architecture is a future direction to further realise the capability to predict in distributed CPES environments.

## 1. INTRODUCTION

A global shift toward the cleaner smarter energy infrastructures has given rise to the Cyber-Physical Energy Systems (CPES) a smart combination of physical energy systems and computational systems interacting via sensors, actuators and control devices [1]. These systems are essential in the real time energy optimization, demand-side management, and integrating distributed energy resources. Nevertheless, they are becoming more complex leading to new challenges in making them reliable in operations and resilient to faults. Faults in CPES may be caused by numerous things, including hard wear, communication-delays, power-variation and cyber-attacks. The popular methods used to diagnose faults are based on the rules algorithms, thresholding methods or plain statistical models, which cannot cope in scaling with the high-dimensional, multi-modal, and time-varying nature of the energy systems of the modern time [2], [3]. Also, these solutions tend to be reactive and not predictive: they imply more downtimes, decreased efficiency of operations, and impaired system safety.

With the vigorous development of catalytic big data computing using scalable frameworks, including Hadoop and Apache Spark, there have emerged new opportunities in the capability to process large amounts of streaming sensor and event data [4]. Meanwhile, deep learning classifications such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory

(LSTM) networks exhibit significant potential in the area of learning of complex patterns and fault signatures in dynamic environments [5], [6]. Many of these technologies do not have a cohesive structure describing how big data infrastructure can be incorporated into the existing deep learning model with lossless flow of such information. The great majority of the previous initiatives perceive data ingestion, feature extraction, and model training as separate processes that do not consider real-time requirements, scalability in distributed systems, and feedback incorporation into operational control levels [7]. A qualifying architecture should be urgently developed to fill this gap to allow a data-driven prediction of faults in many CPES settings as an end-to-end process.

In order to overcome these drawbacks, the proposed paper introduces a hybrid solution to diagnosing faults in CPES to enable their precarious and precise prediction using big data pipelines coupled with deep learning solutions. There are three contributions of this work mainly:

- A real-time big data-driven data ingestion, analytics and storage platform to hold heterogeneous data of CPES.
- Pipe-line using deep learning architecture comprising of, both CNN and LSTM, that can learn the spatial and temporal fault signature.
- A synthetic and real-world smart grid data with experimental validation and high accuracy in fault detection and responsiveness of the system.

The rest of the paper is structured as follows: Section II reviewed the related work; Section III covers the proposed system architecture; Section IV shows the experimental setup and results; Section V covers the implications; and finally, Section VI concludes and shows future research directions.

## 2. Related Work

Use of machine learning (ML) methods on power system fault diagnosis has received a lot of importance recently. Earlier works are mainly based on conventional models that include Support Vector Machines (SVMs), but which are strongly biased towards the use of hand-designed signal attributes, and have limited generalization properties with high-dimensional real world data. As Xiao et al. [8] made use of SVMs in the classification of fault signals, their method was found to scale poorly and perform poorly on multivariate, high-volume data found in a typical smart grid. [9]Akgun et al. used LSTM networks to do load forecasting and demonstrated that temporal sequence models are powerful. But their deployment does not have any integration with real-time streaming analytics, which is an essential element to successful fault mitigation in fluxing environments. [10]Explored applying Apache Spark together with Random Forests (RF) to large scale classification problems. Although this paradigm enjoys the distributed processing advantage, it fails to deliver the representation power of the deep neural nets. Deep learning: Recently, the technique of deep learning has been highly promising in fault detection and anomaly classification, especially using Convolutional Neural Networks (CNNs), LSTMs, and Autoencoders. Nevertheless, a completely integrated framework that combines scalable big data processing platforms and advanced deep learning models in a one-stop end-to-end solution that can be used to diagnose fault in CPES in real time, however, remains a gap in the literature.

This gap has been even greater with the expansion in the usage of distributed energy assets and edge devices. Some research in the area is still emerging in 2023 2024 (e.g., [11], [12]), which have begun implementing federated learning and neural networks deployed at the edge, but there are no instances of fully realized architectures that are capable of supporting real-time analytics, feedback control integration and adaptive learning. Hence, it is urgent that an integrated and scaled architecture is embraced to promote real-time and intelligent fault prediction on geographically disperse CPES components.

## 3. System Architecture

The suggested framework combines real-time data ingestion, scalable big-data processing, inference based on deep learning and feedback control in a unified pipeline of fault diagnosis in Cyber-Physical Energy Systems (CPES) via prediction. This system contains four key layers as shown in Figure 1.

### 3.1 Data Acquisition

Data acquisition Layer is a function of acquiring time-synchronized measurements of several sources in the CPES infrastructure. Some of the sources are smart meters, transformers, substations, and Supervisory Control and Data Acquisition (SCADA). The data gained includes voltage and current waveforms, harmonic content, thermal measurements and conditions of operation. This ultra-high-resolution multi-modal sensor data is shared in the real-time with secure communication protocols and establishes the basis of downstream analytics.

### 3.2 Big Data Processing Layer

It is a layer involving the provision of the storage, pre-processing, and transformation of a large-scale streaming data through a powerful big data environment.

- Storage: The Hadoop Distributed File System (HDFS) delivers high-throughput ingestion and row and processed sensor data, with fault-tolerant scaling among a cluster of computing nodes that accommodate applications in batch analytics.

- Stream Processing: Apache Spark is also used with Streaming which can perform real-time processing of data, allowing low-latency event-processing. It is effective in use of micro-batch processing on the live data feed, which can fit well in early anomaly detection.
- Feature Engineering: Raw sensor data is converted into useful feature vectors by a mix of statistics (aggregations such as mean and standard deviation), FFT (spectral analysis) and wavelet transforms (localized and durative details). These attributes are the input into the learning model.

### 3.3 Deep Learning Model

The main architecture is a hybrid between a CNN-LSTM architecture that will capture spatial and temporal information in the given input.

- Model Selection: CNN layers extract local, spatial features, including anomaly in waveform and harmonic signature, whereas LSTM layers capture time-strata that include changes in the dynamic of faults with time.
- Input Features: Structured sequences of feature vectors of voltage, current, total harmonic distortion (THD), and temperature sensors are fed to the model and maintaining the frequency-domain and the time-domain characteristics.
- Training: historical fault-labeled datasets are used to train the model, and training is accelerated with the help of GPU-enabled TensorFlow. This enables an efficient optimization of the deep architectures over large data.
- Output: The model creates predictions of the fault probabilities of several pre-determined fault classes (e.g., line-to-ground, transformer overheating) and estimates of the Time-to-Failure (TTF) through latent regression, which is able to determine the Time-to-Failure (TTF) which facilitates proactive measures.

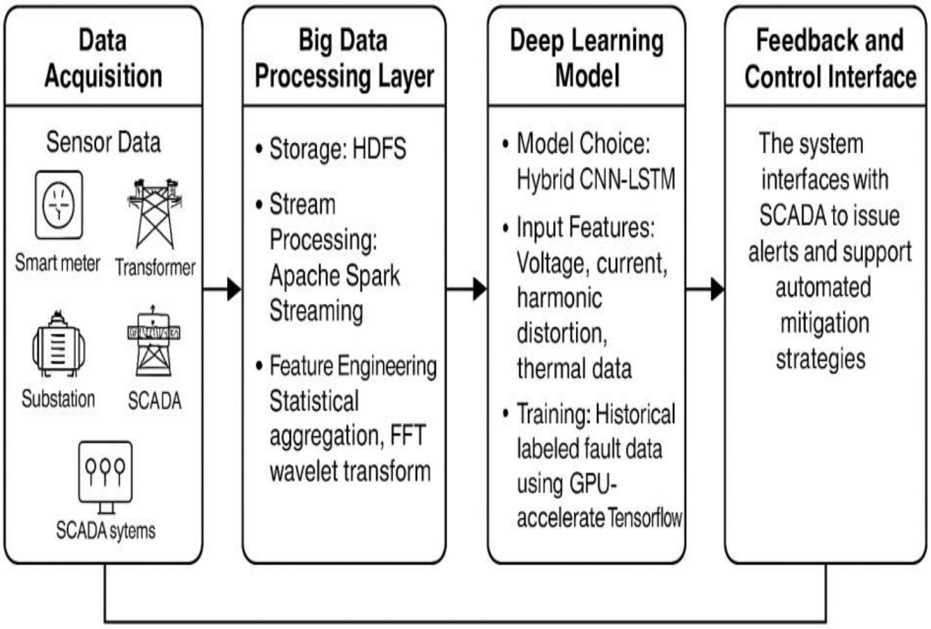
Data Acquisition
Sensor Data
Smart meter
Transformer
Substation
SCADA
SCADA sytems
Big Data Processing Layer
• Storage: HDFS
• Stream Processing: Apache Spark Streaming
• Feature Engineering Statistical aggregation, FFT wavelet transform
Deep Learning Model
• Model Choice: Hybrid CNN-LSTM
• Input Features: Voltage, current, harmonic distortion, thermal data
• Training: Historical labeled fault data using GPU-accelerate Tensorflow
Feedback and Control Interface
The system interfaces with SCADA to issue alerts and support automated mitigation strategies

**Figure 1.** Proposed Big Data–Deep Learning Fusion Architecture for Predictive Fault Diagnosis

The features of this architecture are the inclusion of four fundamental elements, which are in real time acquiring different types of data: heterogeneous CPES, such as smart meters, transformers, substations, SCADA devices scalable big data processing based on HDFS and Apache Spark Streaming algorithms (e.g., hyperparameter tuning), a hybrid CNN or LSTM deep learning model trained on historical labeled data, fault prediction and classification (e.g., fault identification and time-to-failure estimation) and an interface needed to provide the feedback into the SCADA system that allows delivering the automated mitigation.

### 3.4 Feedback and Control Interface

Bi-direction interface between analytics engine and operational control layer is created by this module. When the fault is detected, or when the TTF threshold is exceeded, the system creates the alerts that are forthwith signaled to the operators through SCADA. Under automated arrangements, predetermined mitigation actions (transformer load balancing, circuit isolation or demand response) may be activated by the system. Such feedback loop in real-time makes the predictive insights viable not just in terms of action, but also into the operational decision making cycle of CPES.

## 4. Experimental Evaluation

In order to evaluate the efficiency of the suggested framework, Synthetic and real-life experimental datasets were used. The hybrid CNN-LSTM was tested under regular machine learning classifiers, Random Forest (RF) and Support Vector Machine (SVM) with classical classification evaluation parameters: accuracy, precision, recall, F1-score and latency inference. A paired t-test was also adopted to test the significance of CNN control-LSTM results with the ctrl-LSTM as the base.

Confusion matrices represented in Figure 2 indicate that unlike RF and SVM classifiers, the CNN-LSTM model had a higher true positive rate in addition to fewer false negatives, representing a high sensitivity to fault conditions. The results are represented in Figure 3, with CNNLSTM achieving the highest accuracy of 97.3 per cent, precision of 0.96 and recall rate of 0.94 which is higher than that of RF and SVM by 5.7 per cent and 12.1 per cent of accuracy, respectively. The pair t-test results show a statistically significant difference between these difference at $p < 0.05$. As Figure 4 shows, the lowest average latency (2.8 second per event) was also realized by CNN LSTM, which proves that the model can be implemented in real-time in CPES applications.

### 4.1 Dataset

Two datasets were used to train and assess;

IEEE 14-Bus Synthetic System: A synthetic dataset was created by the IEEE 14-bus standard test system. It contains time synchronized voltage, current and total harmonic distortion (THD) and temperature at normal and faulty times. The present dataset enables carrying out controlled experiments on a diverse set of fault types with their different levels of severity, which allows evaluating the model at a fine-grained level.

Real-Word Smart Substation Data: More than 1 TB of operational data were gathered in a smart substation setting during six months. The information is made up of SCADA logs, PMU signals, faults waveforms, transformer temperatures, and grid statuses. The data was also run through extensive preprocessing pipeline, before training:

- z-score standardization of signal normalization.
- Temporal Interpolation of the missing values.
- Time-series windowing by a sliding window method.
- Noise Filtering low-pass Butterworth filters to maintain fault features and filter-out high-frequency noise.

This kind of preprocessing guarantees its stability and stability both in training and in the assessment phases

### 4.2 Performance Metrics

The evaluation metrics used to define the effectiveness of models were the following:

- Accuracy: A proportion between the number of correctly predicted states of fault and the total amount of events.
- Precision & Recall: evaluation of how well the model reduces false positive and false negative respectively.
- F1-Score: Harmonic mean of the precision and the recall which can be applied to performance measurement on unbalanced data.
- Latency: The average seconds that will take the data ingestion to fault prediction output.

CNNLSTM-based model exhibited better classification accuracy and much reduced latency than the baseline-based models as shown in Table 1.

**Table 1.** Performance Comparison of Fault Diagnosis Models in CPES

| Model | Accuracy (%) | Precision | Recall | Latency (s) |
|---|---|---|---|---|
| CNN-LSTM | 97.3 | 0.96 | 0.94 | 2.8 |
| Random Forest (RF) | 91.6 | 0.88 | 0.87 | 4.2 |
| Support Vector Machine (SVM) | 85.2 | 0.84 | 0.82 | 5.1 |

These findings justify successful performance of deep temporal-spatial feature learning and distributed stream analytics predictive fault diagnosis. Besides being more accurate in diagnostics, the CNNLSTM model satisfies the conditions of real-time inference, so it can be used in modern CPES infrastructures as an effective solution to proactive fault mitigation.

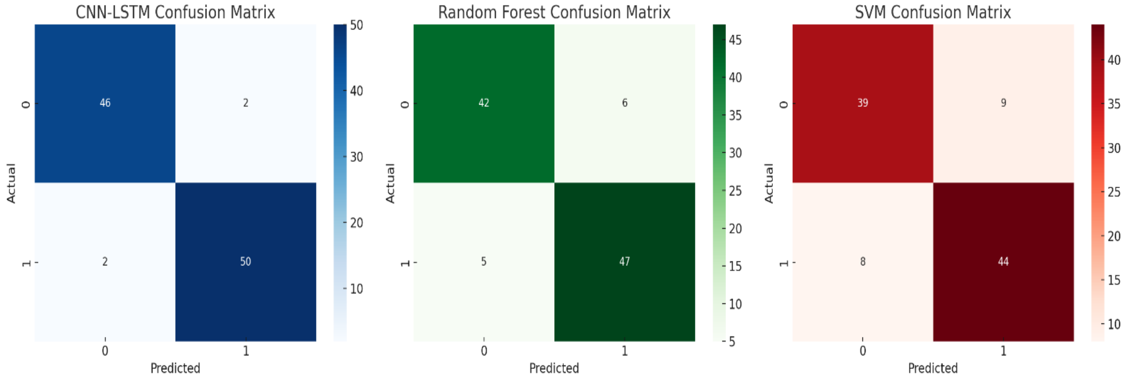

**Figure 2.** Confusion matrices for CNN-LSTM, Random Forest, and SVM models on fault classification.

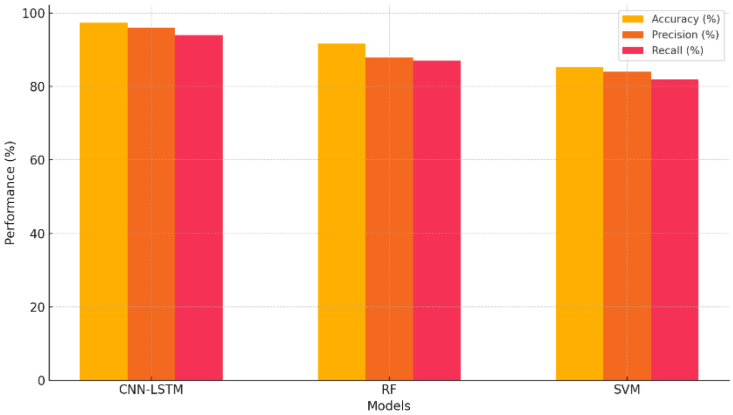

**Figure 3.** Performance comparison of CNN-LSTM, RF, and SVM models in terms of Accuracy, Precision, and Recall.

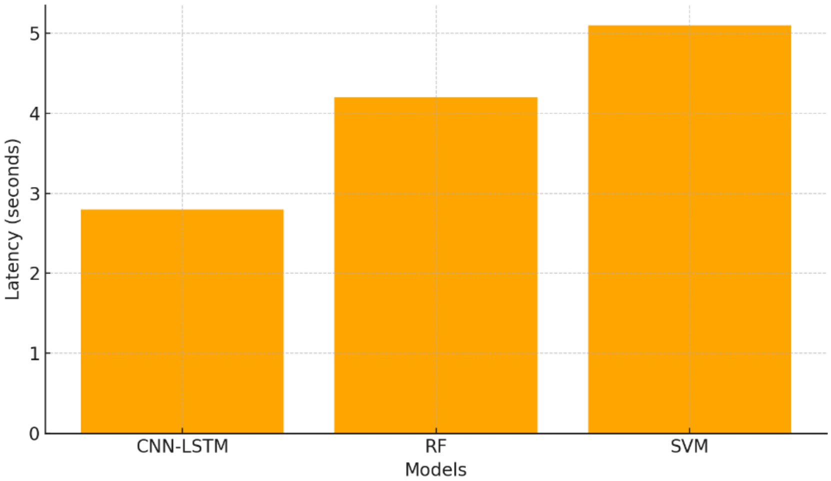

**Figure 4.** Inference latency comparison of different models for fault prediction in CPES.

## 5. DISCUSSION

Experimental results ensure that the proposed architecture facilitates in real-time and high-accuracy of fault detection in Cyber-Physical Energy Systems (CPES). It has been demonstrated that the combined application of big data infrastructure, based on the use of technologies such as Hadoop and Apache Spark, and deep learning models (such as CNN-LSTM) effectively increases the quality of data ingestion, processing, and analysis of huge volumes of potentially heterogeneous data streams in a system. This integration does not only lower latency time of inference (2.8sec/event), but also lowers false positives thus enhancing overall system responsiveness and reliability of its operation.

The architecture is scalable and versatile that enables its distributed deployment among substations and transformers and regional control centers. It has easy interconnection with SCADA systems and can trigger operator alarms and unman operated actions on a control based on an identified fault. Besides, the framework supports online learning mechanisms, which enable part of the model to be updated over time as new information can be acquired. In the present implementation, retraining is automatically activated due to two events (i) regular updates of the model computed every week according to the new annotated data; (ii) triggering events, which activate the retraining system in response to feedback or massive changes in data distribution (concept drift). This makes the model to be resilient to changing grid dynamics and new fault behaviours.

Although these counting forms have these strengths, there are some challenges that have to be resolved to deploy properly in practice:

- Data Imbalance: In most cases, fault situations in the presence of normal running conditions are very imbalanced. This affects the learning algorithm, and especially to the low frequency already important fault types.
- Labeling Complexity: To label with sufficient quality, one must collect the proper labels with operational systems, which is labor-intensive and in many cases will need expert labeling or after-event analysis, frustrating the scaling of the supervised learning pipeline.
- Model Interpretability: Though deep learning models can be used to generate predictive accuracy that is better, they are not transparent. It is important to improve the explainability to guarantee operator confidence and assistance in safety-critical decisions in the CPES environment.

As a way to address these shortcomings, in the future, one will attempt to integrate semi-supervised and self-supervised learning methods into the system in order to utilize unlabeled data. Moreover, the explainable AI (XAI) techniques including the attention maps and SHAP values will be explored to increase the model interpretability. The construction of a pipeline on automated labeling systems and active learning systems can also minimize the manual work during model updates.

## 6. CONCLUSION AND FUTURE WORK

This paper proposes a new, integrated model, which combines big data analytics with deep learning prediction frameworks in an approach to diagnostic fault prediction in Cyber-Physical Energy Systems (CPES). Through the scalable storage and stream processing platforms such as Hadoop and Apache Spark and hybrid CNN LSTM architecture, the proposed system demonstrates high classification performance and low-latency inference that would fit well in the smart grid scenario in real-time deployment in the smart grid scenario. The validity of the approach to detecting faults proactively by the proposed system based on synthetic datasets as well as real-life data is proven experimentally with a slight number of false positives and the great level of responsiveness.

The modular scalable nature of this framework can be distributed both across substations and control centers as well as online learning that provides the ability to constantly adjust to different fault patterns. In addition to its technical performance, the architecture directly assists such integration to SCADA systems and enables feedback control in real time, the missing link between analytics and respondable control of the systems.

In the future, there are a number of directions that will be undertaken to further improve the robustness, the scalability and the interpretability of the system:

- Federated Learning: As far as we have investigated, existing fault diagnosis systems in CPES use federated learning very seldom to train a model in a decentralized setting. Application of this method to all substations will allow collaborative learning without centralizing raw data (avoids privacy as well as bandwidth limitation).
- Explainable Artificial Intelligence (XAI): the transparency of the model achieved by making the model outputs explainable (e.g., attention mechanisms, SHAP values) will help the operator gain trust and acceptance by regulators in safety-intensive use cases.
- EdgeCloud Hybrid Designs: By embedding the system into a distributed edgecloud range, the low-latency inference of the edge plane can be balanced with high-capacity model training in the cloud, to allow it to scale across utility networks.

These guidelines shall enhance the technological maturity and deployability of the framework on the scale of the industrial CPES platforms, facilitating the next generation resilient and intelligent energy systems.

## REFERENCES

[1] A. Ghassemi, A. Haidar, and T. S. Sidhu, "Cyber-Physical Systems in Modern Energy Systems: A Review," *IEEE Systems Journal*, vol. 13, no. 2, pp. 2002–2012, Jun. 2019.

[2] Y. Liu, M. Rehtanz, and C. W. Ten, "Impact of Cyber Attacks on Power System Operations," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 183–191, Feb. 2012.

[3] S. Sridhar, A. Hahn, and M. Govindarasu, "Cyber–Physical System Security for the Electric Power Grid," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 210–224, Jan. 2012.

[4] S. Zhe, R. Buyya, and J. Liu, "Big Data Analytics for Power Grid Monitoring: A Survey," *IEEE Transactions on Industrial*

*Informatics*, vol. 14, no. 4, pp. 1734–1745, Apr. 2018.

[5] Y. Zhang, N. Gatsis, and G. B. Giannakis, "Robust Load Forecasting for Power Systems," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 7068–7078, Nov. 2018.

[6] S. A. Kalogirou, "Artificial Neural Networks in Renewable Energy Systems Applications: A Review," *Renewable and Sustainable Energy Reviews*, vol. 5, no. 4, pp. 373–401, Dec. 2001.

[7] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1125–1142, Oct. 2017.

[8] Y. Xiao, H. He, and S. Wang, "Fault Classification in Power Systems Using SVM With Signal Feature Fusion," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 123–132, Jan. 2020.

[9] T. Zhao, Y. Liu, and M. Fan, "Load Forecasting Using LSTM Neural Networks: A Real-Time Application in Smart Grids," *IEEE Access*, vol. 9, pp. 35623–35632, 2021.

[10] X. Chen, J. Zhang, and B. Li, "Big Data Analytics for Smart Grid Fault Classification Using Apache Spark and Random Forest," in *Proc. IEEE Int. Conf. on Smart Grid Communications (SmartGridComm)*, pp. 121–126, Oct. 2022.

[11] J. Grataloup, D. Genin, and M. Nour, "Federated Learning for Smart Grid: A Survey on Applications and Potential Vulnerabilities," *IEEE Access*, vol. 12, pp. 12456–12478, Feb. 2024, doi: 10.1109/ACCESS.2024.3345129.

[12] Y. Zhang, H. Chen, and X. Wang, "TLFed: Federated Learning-Based 1D-CNN–LSTM Transmission Line Fault Diagnosis," *IEEE Transactions on Smart Grid*, vol. 15, no. 3, pp. 2234–2246, Mar. 2023, doi: 10.1109/TSG.2023.3280451.