Journal of Intelligent Information and Communication

Review Paper

Big Data Analysis Techniques for Power System Fault Detection and Diagnosis: A Review

Lucie Kralova[1, *], Zuzana Horska[2]

[1] Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, Ostrava 70833, Moravian-Silesian, Czech Republic
[2] Faculty of Electrical Engineering, Czech Technical University in Prague, Prague 16627, Czech Republic
*Corresponding author: l.kralova@vsb.cz

Abstract. With the integration of renewable energy and the rapid development of smart grids, ensuring the reliability and security of modern power systems faces great challenges. This paper discusses how big data analytics and machine learning are transforming fault detection and diagnosis in dynamic power environments to address these complex issues. First, the main types and sources of fault data are introduced, and then the efficient data collection and preprocessing are emphasized. By adopting advanced feature engineering (including extraction, selection and dimensionality reduction), the diagnostic accuracy and system adaptability are significantly improved. Thru comparative analysis, the advantages and disadvantages of traditional machine learning, deep learning and hybrid algorithms in different fault scenarios are revealed. The paper then introduces the real-time implementation, focusing on the scalable data platform and how to combine it with the IoT-enabled smart grid. Finally, the paper also discusses important issues such as data quality, scalability, interpretability, and cyber-physical security. It also points out that the latest technologies such as edge artificial intelligence, federated learning, and digital twins represent directions with great potential for future development. These findings provide a comprehensive foundation for building a flexible, adaptive, and intelligent power grid management system, and provide valuable guidance for researchers and industry practitioners.

Keywords: *big data, fault detection, smart grids, machine learning, feature engineering*

Received on 12 September 2024, Accepted on 28 December 2024, Published on 11 February 2025

Copyright © 2025 Lucie Kralova and Zuzana Horska licensed to JIIC. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

Introduction

Modern power systems have become more complex and larger in scale, which poses great challenges to ensuring their reliability, security and efficiency. The increasing prevalence of renewable energy sources, the shift towards electric transportation, and the move to smart grids have collectively placed significant strain on conventional power system management and monitoring approaches [1]. Faults within power systems, which can vary from minor transient disturbances to severe catastrophic failures, have the potential to trigger cascading outages, substantial economic losses, and in extreme scenarios, extensive blackouts. As a result, ensuring the safe operation of contemporary power systems hinges critically on the ability to swiftly and precisely detect and diagnose faults [2]. Over the past decade, the rapid development of sensing, communication, and information infrastructure has driven a paradigm shift in power systems. Internet of Things (IoT) sensors, intelligent electronic devices, phasor measurement units (PMUs), and supervisory control and data acquisition (SCADA) systems have added a large amount of heterogeneous, high-frequency, and high-dimensional data [3]. This "big data" environment offers tremendous opportunities for deeper insights and automation, but it also poses significant challenges, including data management, processing, and interpretation. As the volume, velocity, and variety of data continue to increase, advanced analytics techniques are necessary [4]. These analytics can provide useful information for real-time decision making.

Undetected or misdiagnosed faults in power systems can have significant impacts. In addition to the greater impact on grid resilience, cyber-physical security, and the integration of distributed energy resources, they can also have a greater impact on direct technical and economic impacts such as equipment damage and service interruptions [5]. Traditional fault detection and diagnosis methods are mainly based on rule logic, expert systems, or simple thresholds [6] when facing non-stationary operating conditions, hidden faults, and complex interactions between grid components. Therefore, advanced data-driven and machine learning-based methods are being actively investigated to improve sensitivity, specificity, and adaptability. Compared with traditional systems [7], these methods automatically identify, locate and classify faults by using statistical learning, pattern recognition and optimization. However, translating these technological advances into scalable, reliable, and stable solutions remains a significant challenge. Data quality, privacy, model interpretability, and decision latency are hot topics of discussion [8].

This paper provides an overview of the latest research on big data-driven power system fault detection and diagnosis, including the analysis process of data acquisition and pre-processing, feature engineering, model development, and real-time implementation and system integration. The integration of hybrid intelligent framework, machine learning algorithms and big data platform is particularly emphasized, which will contribute to the realization of next-generation fault analysis. In addition, it critically examines the common problems of current methods, finds unresolved research problems, and proposes new directions for future research. There are three main objectives. First, to systematically classify and compare the latest techniques and methods for fault detection and diagnosis under the background of big data; second, to deeply study the challenges and practical factors faced by deploying these methods in the running power system; third, to propose a roadmap for future research to address the identified gaps and support the development of more resilient, adaptable and intelligent grid management systems. The rest of this paper details the current status and prospects of power system fault analysis based on big data. They include data sources, features and preprocessing, feature engineering techniques, fault detection and diagnosis algorithms, real-time problem solving and actual deployment scenarios, current challenges, emerging trends and future research directions.

Data Acquisition and Preprocessing in Power Systems

Types and Sources of Fault Data

Fault analysis of power system mainly depends on the comprehensiveness and fine granularity of the collected data flow. These data streams form the basis for the accuracy of fault detection and the interpretability of post-event diagnosis. Modern power grids have multi-layer sensing infrastructure. Supervisory control and data acquisition (SCADA) systems frequently measure system states, such as bus voltages, feeder currents, and circuit breaker statuses, with measurement intervals typically ranging from minutes to seconds [9][10]. Phasor measurement units (PMUs) can achieve system-wide synchronized phasor data acquisition to capture voltage and current phasors with high temporal resolution. PMUs typically sample 30 to 60 times per second, enabling real-time oscillation analysis and dynamic state estimation [11]. As an event driver, the protection relay can use waveform capture, event log and event sequence report to generate high-fidelity digital records [12]. With the popularity of Internet of Things (IoT) devices, the observability boundary is further expanded, and these devices add distributed environment and operational data, such as transformer temperature, humidity and partial discharge signals [13]. Databases for model training, validation, and forensic analysis, including maintenance logs, fault records, and the history of operator operations [14].

Table 1. Comparative Characteristics of Power System Fault Data Sources.

Data Source	Sampling Rate	Data Type	Main Advantages	Limitations
SCADA	2-10s	Steady-state	Wide coverage, operational status	Low granularity, latency
PMU	30-60 Hz	Synchrophasor	High temporal resolution, accuracy	Infrastructure cost, partial coverage
Relays	Event-driven	Waveforms; logs	Precise event capture	Sparse in normal operation
IoT Devices	Variable	Environmental	Distributed, context-aware	Data heterogeneity, cybersecurity
Historical DBs	N/A	Discrete records	Long-term trends, rare events	Data incompleteness, manual errors

Table 1. summarizes the main characteristics of the main data sources in power system fault analysis, highlighting the acquisition rate, data type and integration issues. The fusion of heterogeneous data streams remains a significant challenge [15] because time synchronization and inconsistent data formats often hinder the overall reconstruction of faults.

Data Quality Issues and Challenges

The accuracy of fault diagnosis depends largely on the quality of input data. SCADA telemetry data may lead to inaccurate state estimation due to missing values, quantization errors, and update delays, especially during fast transient processes [16]. Although PMU datasets are accurate, data loss, phase angle inconsistency, and GPS synchronization loss are all possible, especially in the case of communication congestion or hardware failure [17]. Relay protection records may have incompleteness or misalignment due to firmware errors or network delays, which makes accurate event sequencing more difficult. Node failures, data tampering, and clock drift are potential vulnerabilities of IoT sensor networks that can undermine the ability to augment measurement data. Historical databases often make supervised machine learning model training more difficult because they can lead to annotation errors, timestamp inconsistencies, and insufficient metadata [18]. To ensure the reliability of the analysis, robust preprocessing pipelines must be established, as these quality issues exist in both the temporal and spatial dimensions.

Data Cleaning, Imputation, Synchronization

Data preprocessing is a multi-stage process in power system fault analysis, including outlier detection, missing data filling and multi-source synchronization. Outlier detection uses statistical tests (Grubbs filter or Hampel filter [9]) and model-based residual analysis to identify and eliminate anomalous measurements. Missing data imputation can be performed using a variety of methods, including linear interpolation and model-based matrix completion, as well as expectation maximization and model-based matrix completion, which exploit system redundancy and spatial correlation [10]. Kalman filters or autoregressive models are often used to fill in PMU data gaps, estimating the propagation state within the missing interval [11]. To reconstruct accurate timelines of disturbances, time alignment algorithms such as cross-correlation maximization or dynamic time warping must be used to synchronize multiple data sources, especially between event-based data sources and continuous data streams [12]. These methods are not easy to use, as incorrect cleaning or attribution can lead to systematic biases that mask the anomalies one is trying to discover [13].

Big Data Platforms for Power Systems

With the rapid increase in data volume, speed and diversity, there is an urgent need for scalable big data architecture that can handle large amounts of data streams to achieve real-time data acquisition, storage and processing. The distributed storage and parallel computing capabilities based on the Hadoop platform support batch analysis for model retraining and historical event mining [14]. This paradigm has been extended by Spark Streaming to enable low-latency analysis of PMU and IoT data streams, which is critical for online fault location and predictive maintenance [15]. Commercial and research platforms such as Pioneer offer features such as event extraction, waveform analysis, and hybrid data fusion, specifically for power system applications [16]. These platforms typically support structured and unstructured data and support SQL and NoSQL engines [17]. Integrating these platforms with traditional SCADA systems and emerging Internet of Things (IoT) infrastructure remains a systems engineering challenge, particularly in terms of cybersecurity, data privacy, and regulatory compliance [18].

The integration and transformation of multi-source data into actionable intelligence is a technological achievement and a prerequisite for achieving smart, adaptive and resilient power systems. The future of power system fault analysis will be influenced by the ever-improving data collection and pre-processing techniques.

Feature Engineering for Fault Analytics

Feature Extraction Methods

The diagnostic effect of power system fault analysis mainly depends on the discrimination ability of features. Time-domain descriptors, including root mean square (RMS) value, average value and peak factor, can quantitatively describe the non-stationery and mutation characteristics of voltage and current signals, so as to capture transient and steady-state behaviors [19]. Frequency domain analysis uses Fourier and wavelet transforms to reveal changes in spectral content and oscillatory characteristics associated with arcing faults, line-to-ground events, and equipment resonance [20]. Higher-order statistical features such as kurtosis, skewness, and entropy can sensitively identify noise interference and waveform irregularity, which helps to isolate early anomalies [21]. Signal envelope, peak count, and zero-crossing rate are waveform shape parameters that can be used to distinguish fault categories by morphological distortion and encoding periodicity [22]. In recent studies, multi-domain feature fusion is of great significance for improving classification accuracy and robustness under highly variable operating conditions [23]. Table 2 presents a structured comparison of principal feature extraction methods and their diagnostic relevance across various fault types, demonstrating the need for tailored feature sets in multi-scenario analytics.

Table 2. Feature Extraction Methods in Power System Fault Analytics

Feature Category	Representative Features	Diagnostic Relevance	Typical Faults Detected
Time-domain	RMS mean crest factor rise time	Signal amplitude transition detection	Short-circuit transient overload
Frequency-domain	FFT coefficients spectral energy harmonics	Oscillation resonance frequency shifts	Arc resonance harmonic distortion
Statistical	Kurtosis skewness entropy variance	Irregularity impulsiveness	Intermittent stochastic faults
Waveform shape	Zero-crossing peak count envelope	Morphology periodicity waveform asymmetry	Grounding breaker fuse failures

Feature Selection Algorithms

For high-dimensional fault datasets, the curse of dimensionality requires strict feature selection to reduce redundancy, reduce noise, and improve the generalization ability of the model. Filter methods such as mutual information and correlation-based ranking can evaluate feature relevance independently of the classifier, thus improving the computational efficiency of large-scale screening [24]. Wrapper methods, which iteratively optimize feature subsets to fit a specific learning algorithm, including recursive feature elimination and sequential forward selection, but these methods increase the computational burden [25]. To directly exploit structure-induced sparsity, embedded techniques, such as LASSO regression and decision tree regularization, integrate feature selection into the model training process [26]. Rough set theory comes from granular computing, which uses approximate equivalence classes and indiscernibility to determine the minimum feature set that can maintain decision consistency. This approach is particularly suitable for erroneous or incomplete power system information [27]. Comparative studies have shown that the combined selection strategy using the filter and wrapper paradigm can effectively identify faults and data noise robustness [28]. Table 3 summarizes how the main feature selection and dimensionality reduction techniques work and their trade-offs, providing a reference for practitioners who want to optimize pipeline analysis.

Table 3. Comparison of Feature Selection and Dimensionality Reduction Techniques.

Method	Principle	Computational Cost	Interpretability	Robustness	Suitability for Fault Analytics
Filter	Statistical relevance	Low	High	Moderate	Initial screening large datasets
Wrapper	Classifier-based evaluation	High	Moderate	High	Model-specific optimization
Embedded	Regularization pruning	Medium	High	High	Sparse or structured data
Rough Set	Indiscernibility reducts	Medium	High	High	Incomplete imprecise datasets
PCA	Linear projection	Medium	Low	Moderate	Correlated features variance focus
LDA	Discriminant analysis	Medium	Moderate	Moderate	Multi-class well-separated classes

t-SNE	Manifold learning	High	Low	Moderate	Visualization complex nonlinearity

Dimensionality Reduction Techniques

Dimensionality reduction aims to collect key data while reducing computational inefficiency and the risk of overfitting. Principal component analysis (PCA) transforms the original variables into uncorrelated principal components thru orthogonal transformation. This maximizes variance retention and simplifies subsequent modeling. However, in highly nonlinear cases, PCA may weaken the physical interpretability [24]. Linear discriminant analysis (LDA) directly optimizes the class discrimination ability and supports multi-fault classification scenarios [25]. In addition, it projects the data onto the axs that best separate the class means. t-distributed stochastic neighbor embedding (t-SNE) preserves the local neighborhood structure in high-dimensional space and can reveal complex nonlinear relationships. However, it is computationally expensive and has limited scalability in real-time analysis [26]. Empirical studies have shown that combining dimensionality reduction with domain-driven feature selection can make the representation compact and interpretable while avoiding significant loss of discrimination ability [27,28].

Case Studies and Applications

Applied research indicates that feature extraction, feature selection, and dimension reduction are key to identifying high-fidelity faults in actual operating environments. The accuracy of support vector machine (SVM) and artificial neural network (ANN) models in transmission line fault classification can be improved by combining time-frequency features with a selection method based on rough set theory [19,20]. Statistical feature extraction and principal component analysis (PCA) techniques are used for fault location in distribution networks to maintain computational feasibility and reduce false alarm rates [21,22]. Real-world relay event logs processed by embedded feature selection and linear discriminant analysis (LDA) projection support automated restoration processes in large-scale power grids [23,24]. This enables rapid isolation of circuit breaker failures and ground faults. In a recently released hybrid analysis pipeline, the combination of package selection and manifold learning can identify rare and complex failure modes in high-dimensional IoT monitoring environments [25,26]. These findings highlight that feature engineering strategies must be matched to the data type and diagnostic goals [27-29].

In the field of power system fault analysis, feature engineering is shifting toward more automated, adaptive, and context-aware approaches. As the volume and complexity of operational data continue to increase, the future development of fault analysis will depend on the continuous integration of advanced feature engineering and scalable, explainable machine learning infrastructure.

Machine Learning and Hybrid Algorithms for Fault Detection and Diagnosis

Traditional Machine Learning Models

Algorithmic intelligence is changing the way power systems are detected and diagnosed. Support vector machine (SVM) achieves the maximum margin hyperplane in the high-dimensional feature space, and alleviates over-fitting by balancing empirical risk and regularization. Early comparative studies have shown that SVMs are more effective in dealing with small to medium-sized high-dimensional fault datasets [30]. Decision trees achieve interpretable rule-based diagnosis by recursively constructing space partitions parallel to the input axis; random forests, as an ensemble of decision trees, utilize bootstrap aggregation techniques to minimize variance and enhance generalization ability when redundant and nonlinear attributes exist [31,32]. K-means clustering partitions unlabeled data thru centroid-driven clustering. This helps with unsupervised anomaly detection and preliminary state differentiation, especially in sparse or unavailable labeled fault data [33]. Due to their computational efficiency and visibility, these models are widely used in the commercial field, especially in the traditional power grid where the amount of data is large [34].

Deep Learning and Advanced Models

With the increase of the number of sensors and the speed of data processing in modern power grids, deep learning architectures have been widely used. Artificial neural network (ANN) simulates the nonlinear mapping between multivariate inputs and diagnostic targets thru a hierarchical fully connected structure. This indicates its strong generalization ability in waveform-based fault analysis [35]. Convolutional neural networks (CNNs) are very effective for disturbance classification and sequence event recognition in varying operating environments because they extract hierarchical and translation-invariant features from raw time series and image-based oscillogram representations [36]. Recurrent neural networks (RNNs), including gated recurrent units (GRUs) and long short-term memory (LSTM), can achieve early warning and sequence-to-label mapping for initial fault prediction, which enables them to model the temporal dependencies and memory effects inherent in system dynamics [37]. By integrating convolutional and recurrent modules in a hybrid neural network, spatial and temporal feature hierarchies can be captured simultaneously. This has been demonstrated in recent benchmark studies on PMU and relay datasets [38].

Hybrid and Ensemble Learning Strategies

Hybrid models and ensemble models utilize complementary learning paradigms to overcome the limitations of single algorithm approaches. Before understanding the supervised boundary, the K-means-SVM pipeline performs unsupervised pre-segmentation. This enhances the ability to distinguish ambiguous or rare events [39]. Bayesian networks support reasoning in uncertain environments by explaining the probabilistic causal relationships between system variables and provide reliable fault inference capabilities for measurement data [40]. Genetic algorithms can be used for hyperparameter tuning and feature subset selection to improve the adaptability and performance of the base learner [41]. By using aggregated weak classifier prediction results, Boosting and Bagging ensembles reduce variance and bias, while improving stability in non-stationary fault conditions [42]. Experimental results show that the hybrid framework consistently outperforms the independent model in terms of accuracy, recall, and response time [43,44]. This is especially evident in multi-class and multi-source fault scenarios. Table 4 summarizes the comparative performance of major machine learning/deep learning models on representative benchmark datasets, highlighting diagnostic accuracy, computational cost, and applicability to multiple fault types.

Table 4. Performance Comparison of ML/DL Algorithms on Benchmark Datasets.

Model	Fault Type	Accuracy (%)	Precision (%)	Recall (%)	Inference Time (ms)
SVM	Line-Ground	96.5	95.9	96.8	12
Decision Tree	Line-Line	93.2	91.4	92.7	10
Random Forest	All Types	97.8	97.1	97.6	20
K-means	Anomaly Detection	89.4	84.7	87.9	8
ANN	Transient Faults	98.3	97.9	98.1	28
CNN	Oscillography	99.0	98.7	98.9	35
RNN (LSTM)	Sequence Events	97.6	97.4	97.2	40
Hybrid CNN-LSTM	Mixed Faults	99.5	99.3	99.4	55

Model Evaluation Metrics and Benchmarking

To evaluate the performance of the model, strict and multi-dimensional indicators are needed. The main indicators for diagnostic validity remain recall, accuracy, and precision. A common example of a discriminative classifier is the decision function of a support vector machine (SVM):

$$f(\mathbf{x}) = \mathrm{sign}(\textstyle\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b) \quad \text{Eq. (1)}$$

where α_i are the Lagrange multipliers, y_i the class labels, $K(\cdot,\cdot)$ the kernel function, and b the bias term[45].

Equation 2: Common Classification Metrics

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$
$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Eq. (2)}$$
$$\text{Recall} = \frac{TP}{TP+FN}$$

Where TP, TN, FP, FN denote true positives, true negatives, false positives, and false negatives, respectively [46]. To ensure the generalizability of reported metrics, benchmarking protocols must include cross-validation, stratified sampling, and statistical significance testing [47]. When comparing accuracy, robustness to noise, missing data, and class imbalance is often ignored.

Comparative Analysis and Discussion

Systematic evaluation shows that model selection is closely related to data type, application limitation and operational requirements. Random forest and support vector machine (SVM) have shown high accuracy and generalization ability on medium-sized [35], structured datasets with moderate non-linear features, but they sometimes have difficulty in dealing with high-dimensional, non-stationary time series data. Convolutional neural networks (CNNs) and hybrid CNN-LSTM architectures have shown excellent performance on datasets based on oscillograms and sequence events [36], thanks to their ability to automatically extract features and model time [38]. However, these benefits come with significant computational costs and reduced interpretability, which increase the real-time deployment and operator trust.

Ensemble and hybrid models have successfully balanced the system interpretability and diagnostic accuracy. The diagnostic methods based on K-means-SVM and Bayesian network show robustness to data incompleteness and label ambiguity, which helps to make robust decisions in practical application deployment [39,40]. Combining genetic algorithm optimization with feature selection and ensemble learning can reduce overfitting problems and improve adaptability to the ever-changing power grid topology [41,42]. Table 5 compares the strengths, weaknesses, and notable applications of leading hybrid and ensemble approaches.

Table 5. Hybrid and Ensemble Fault Diagnosis Approaches.

Approach	Core Principle	Main Strengths	Limitations	Typical Applications
K-means-SVM	Unsupervised plus supervised fusion	Rare event discrimination	Parameter sensitivity	Multi-stage fault detection
Bayesian Network	Probabilistic causal inference	Uncertainty handling	Scalability	Root cause analysis
Genetic Algorithm	Evolutionary global optimization	Feature selection	Convergence speed	Model tuning feature ranking
Bagging/Boosting	Aggregated weak learners	Stability under noise	Model complexity	Large-scale multi-class tasks

Following the research trajectory, building a context-aware, scalable, and explainable failure analysis infrastructure is essential. Domain knowledge, advanced learning algorithms, and rigorous evaluation protocols need to be combined to bridge the gap between laboratory prototypes and actual deployment. Future research needs to consider how to balance scalability, interpretability, and performance, as well as how to seamlessly integrate with existing supervisory control and protection systems.

Real-Time Implementation and Applications

Big Data Processing Frameworks

With the increase of high-frequency and heterogeneous data sources in modern power systems, there is an urgent need for a powerful big data processing framework to perform real-time fault analysis. Due to its distributed storage and batch-based MapReduce paradigm, Hadoop has laid the foundation for large-scale offline data processing in power systems and supports asset management and historical trend analysis [48]. Nevertheless, the inherent latency issues of Hadoop and its limited support for low-latency event processing make it unsuitable for real-time fault detection scenarios. Spark improves the speed of iterative data processing and millisecond-level analysis of power quality and protection signals thru in-memory computing and micro-batch stream processing [49]. Pdminer is a domain-specific platform dedicated to optimizing data from high-frequency sensors and event streams in power systems. It uses streaming feature extraction and multi-dimensional spatio-temporal modeling to achieve this goal [50]. These frameworks differ in terms of parallel granularity, fault tolerance, and scalability. According to industrial deployment, Spark finds a balance between real-time response and ease of use, while Pdminer is more suitable for the data types required in the power grid field.

Real-Time Event Stream Processing and Fault Classification

The operation of the next generation power system depends on continuous event stream processing. Spark Streaming technology has recently been deployed in transmission and distribution networks. It supports dynamic situational awareness and rapid anomaly identification, and allows real-time acquisition and analysis of PMU and SCADA event streams [51]. Pdminer's stream processing engine is designed to handle multi-source power signals. It can simultaneously monitor voltage sags, current surges, and frequency anomalies in thousands of substations, and provide interactive dashboards for operator intervention [52]. High-throughput classification models are integrated into these platforms to distribute various fault types (transformer faults, line-to-ground faults, and line-to-line faults) online in sub-seconds, thereby immediately initiating protection actions and grid reconfiguration commands [53]. Empirical results from a provincial power company show that the event-driven architecture can achieve a fault detection accuracy of over 98% and an average fault isolation time reduced by over 40% compared with the traditional polling system. Table 6 summarizes the key performance indicators and actual application results of leading real-time analysis systems in different operating scenarios. Table 6 details the main performance indicators and case study outcomes of typical real-time analytics platforms for rapid fault detection and classification.

Table 6. Real-Time System Performance Metrics and Case Study Results.

System Platform	Application Scenario	Response Time (ms)	Localization Error (km)	Detection Accuracy (%)	Isolation Time Reduction (%)
Hadoop	Offline Historical Analysis	6500	1.5	94.2	5
Spark	Real-Time Stream Detection	10	0.3	98.9	42
Pdminer	Multi-Node Live Monitoring	15	0.2	98.5	45

Integration with Smart Grid and IoT Systems

The combination of big data analytics and IoT-enabled smart grids is changing the way real-time fault diagnosis is performed. Cloud analytics platforms, edge computing nodes, and distributed sensor arrays form a cyber-physical infrastructure with the ability to perform rapid data aggregation and distributed intelligent processing [54]. The real-time analytics platform interoperates with grid automation and SCADA systems thru standardized APIs. It supports advanced functions such as adaptive load management, fault self-healing, and resilient microgrid operation. Field devices collect various data, including voltage, current, vibration, and environmental parameters, while edge gateways perform preliminary feature extraction and noise filtering. Millisecond-level closed-loop response can be achieved by using centralized clusters of Spark or Pdminer for deep learning-based trend analysis and fault identification [55]. International standards (IEC 61850, DL/T 860) solve the interoperability problem and realize seamless data exchange and system integration. According to empirical research, the integration of the Internet of Things and big data analysis can improve the self-healing rate of distribution networks by 12% and shorten the fault isolation time by 35%.

Challenges in Real-Time Deployment

Despite the great progress, there are still many technical and operational issues to be addressed to realize the real-time deployment of fault analysis in critical power grid environments. Network congestion and processing latency become more serious due to the exponential growth of streaming data. There is an urgent need to adopt a hierarchical edge-cloud architecture and adaptive buffering strategy. Model generalization is limited to rare or extreme operating conditions. However, transfer learning and multi-source data fusion are emerging as promising solutions to enhance unstable dynamic environments [48][52]. To ensure high reliability and system availability, both hardware and software layers require redundancy and are equipped with advanced fault detection and recovery systems. According to actual deployment, single point failure and node failure are still the main obstacles faced by the current platform [53,54]. End-to-end encryption, secure data isolation, and real-time audit logs at each stage of the data lifecycle are widely adopted because data security and privacy issues are becoming increasingly serious. Standardization and compatibility issues have hindered multi-vendor deployment and cross-domain integration. This indicates that it is necessary to establish unified protocols and open interfaces in industry governance.

Challenges, Trends, and Future Directions

Data Quality, Security, and Privacy

The reliability of power system fault analysis depends mainly on the quality and integrity of data. Diagnostic accuracy and model robustness are affected by data incompleteness, noise, missing values, and synchronization errors, which are particularly common in wide-area measurement systems [56]. Sensor drift, calibration inconsistency, and communication delays increase the risk of false positives and false negatives in fault events, especially in large-scale, heterogeneous power grids. Interconnected devices and remote sensing units are facing increasingly urgent challenges in data security. With each point of integration increasing the attack surface, power infrastructure is at risk of eavesdropping, spoofing, and data tampering [57]. The aggregation of operational and customer-side data as utilities adopts advanced metering infrastructure raises privacy concerns. Regulations such as the General Data Protection Regulation (GDPR) impose strict limitations on data storage, sharing, and processing, forcing the use of strong encryption, anonymization, and access control methods [58].

Scalability and Adaptability of Algorithms

With the increase of data volume and the complexity of power grid structure, higher requirements are put forward for scalable analysis solutions. Traditional machine learning models perform well on medium-sized datasets, but they need to handle the computational and memory pressure brought by high-frequency, multi-source streaming data. Distributed frameworks such as Hadoop and Spark can process in parallel, but they face other problems such as model synchronization, latency management, and fault tolerance in real-time pipelines [59]. Algorithm adaptability remains a problem. Changes in grid topology, operating mode, and equipment population can cause static models trained on historical data to quickly lose relevance. Concept drift, i.e., the statistical properties of the target variable change over time, requires adaptive threshold adjustment, online learning, and continuous retraining, but these methods are rarely effective in industrial deployment. The lack of a unified benchmark dataset to evaluate the scalability and adaptability of algorithms makes comparability across studies more difficult [60].

Model Interpretability and Trust

As the analysis process becomes more complex, operator trust and interpretability have become important obstacles to the adoption of technology. Black-box models such as deep neural networks can capture complex non-linear relationships, but they fail to provide insight into decision-making mechanisms or causal relationships. This opacity can cause grid operators and regulators to lose trust in them, especially in mission-critical protection and control applications. Rule-based systems and decision tree ensembles can provide greater transparency, but their prediction accuracy may decrease in highly dynamic or noisy environments [61]. Recent advances in model-agnostic interpretability, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), enable post-hoc explanations of prediction results. However, they are still not ready for integration with real-time, safety-critical workflows. Finding a balance between computational feasibility, interpretability, and accuracy remains a major obstacle to research [62].

Emerging Technologies: Edge AI, Federated Learning, Digital Twins

Many new paradigms have recently emerged in the field of power system analysis. Edge AI brings computing power closer to the data source, thereby reducing latency and bandwidth consumption, and reducing dependence on centralized failures. Due to the embedding of lightweight machine learning models into hardware, local anomaly detection and preliminary diagnosis of substations and field equipment have become easier [63]. Federated learning enables distributed nodes to collaborate in training global models without exchanging raw data. This helps address data sovereignty and privacy issues, particularly in cross-jurisdictional or multi-utility environments. Digital twins, which are virtual replicas of physical grid assets and systems, can be used for real-time simulation, scenario analysis, and predictive maintenance. These technologies lay the foundation for resilient and adaptive grid management. Nevertheless, the actual integration of these technologies has been limited to pilot-scale demonstrations, and there are still unresolved issues in interoperability, standardization, and life-cycle management [64].

Recommendations for Future Research

To fully leverage technological advancements and address ongoing challenges, the following research directions are recommended. By improving data validation, cleaning, and missing value imputation techniques, noise and incompleteness in measurement streams can be reduced. The application of privacy-preserving analysis techniques (such as secure multi-party computation, homomorphic encryption, and differential privacy) in power systems should be carefully studied and evaluated. Online learning frameworks suitable for changes in grid topology and concept drift need to be developed. To enhance model interpretability, inherently transparent algorithms should be adopted, along with post-hoc explanation tools developed for grid operators. Further research is needed on federated learning architectures and edge AI in terms of security, resource optimization, and real-time coordination. Digital twin platforms need to mature to enable scalable modeling, bidirectional data integration, and integration with existing supervisory control and data acquisition (SCADA) systems.

Table 7 summarizes the main issues and research gaps identified in the current literature, providing a structured reference for future research; it outlines the main technical and methodological obstacles facing real-time fault analysis, as well as key research gaps that require further investigation.

Table 7. Current Challenges and Research Gaps in Real-Time Fault Analytics.

Challenge Area	Key Issues	Research Gaps
Data Quality	Incomplete noisy unsynchronized data	Advanced validation and robust imputation
Security & Privacy	Cyber threats data leakage regulatory compliance	Privacy-preserving analytics secure computation
Scalability	Big data volumes real-time processing constraints	Benchmark datasets adaptive distributed algorithms
Adaptability	Concept drift evolving grid topology	Lifelong and transfer learning approaches
Interpretability	Black-box models operator distrust	Real-time actionable explanations
Integration	Heterogeneous systems lack of standards	Unified protocols interoperable frameworks
Emerging Tech	Edge Cloud IoT integration federated training twins	Practical pilot deployments lifecycle management

Conclusion

This paper presents a detailed review of the development process, state-of-the-art techniques, practical applications, and future prospects of data-driven fault analysis in modern power systems. The analysis shows that the combination of advanced feature engineering and hybrid algorithm framework and machine learning is essential to improve the accuracy, robustness and speed of fault detection and diagnosis. Time-domain, frequency-domain, and multi-domain feature extraction techniques can be used to meticulously characterize complex fault patterns, while rigorous feature selection and dimensionality reduction can reduce overfitting and improve interpretability. Traditional machine learning models, such as support vector machine (SVM), decision tree ensemble, and clustering, have been proven to be well-structured for medium-dimensional datasets. However, their scalability and adaptability are often limited when facing high-frequency multi-source streaming data. The paradigm shift was triggered by deep learning architectures, especially the emergence of convolutional neural networks and recurrent neural networks. These networks excel at handling high-dimensional, non-stationary, and sequential data. Hybrid and ensemble models consistently outperform single-algorithm solutions in multi-category, multi-source, and ever-changing power grid scenarios. It makes full use of the complementary advantages of unsupervised, supervised and optimization-driven methods. The balance between prediction accuracy, computational cost, noise resistance, and interpretability is determined thru rigorous benchmarking and multi-metric evaluation. With the integration of scalable big data processing frameworks, the transition from research to practical application has become faster.

References

[1] Ahmad, T., Chen, H., & Wang, J. (2018). A review on applications of ANN and SVM for building electrical lo ad forecasting. Energy and Buildings, 158, 1397-1410. https://doi.org/10.1016/j.enbuild.2017.11.058

[2] Alhamadi, H. M., & Soliman, S. A. (2018). Long-term/mid-term electric load forecasting based on short-ter m correlation and annual growth. Electric Power Systems Research, 80(3), 277-284. https://doi.org/10.10 16/j.epsr.2009.11.005

[3] Amin, S. U., Biswas, S., & Kim, H. (2018). Power system event classification via convolutional neural netwo rk and stacked autoencoder. Applied Sciences, 8(4), 625. https://doi.org/10.3390/app8040625

[4] Anwar, S., & Mahmood, A. (2018). Data analytics for smart grid fault diagnosis: State-of-the-art, challenge s, and future directions. Electric Power Systems Research, 163, 116-125. https://doi.org/10.1016/j.epsr.2 018.06.003

[5] Aslam, S., & Javaid, N. (2018). Towards efficient anomaly-based intrusion detection in smart grids: Recent developments and challenges. Energies, 11(4), 861. https://doi.org/10.3390/en11040861

[6] Barik, R. K., Dubey, H., & Rodrigues, J. J. P. C. (2018). Cloud-based big data analytics for smart future grids. Journal of Cleaner Production, 197, 1232-1243. https://doi.org/10.1016/j.jclepro.2018.06.245

[7] Bedi, J., & Toshniwal, D. (2018). Deep learning framework to forecast electricity demand. Applied Energy, 238, 1312-1326. https://doi.org/10.1016/j.apenergy.2019.01.065

[8] Mohamad, M., Selamat, A., Krejcar, O., Crespo, R. G., Herrera-Viedma, E., & Fujita, H. (2021). Enhancing b ig data feature selection using a hybrid correlation-based feature selection. Electronics, 10(23), 2984. http s: //doi.org/10.3390/electronics10232984

[9] A. Vosughi, S. Pannala and A. K. Srivastava, "Event Detection, Classification and Localization in an Active D istribution Grid Using Data-Driven System Identification, Weighted Voting and Graph," in IEEE Transaction s on Smart Grid, vol. 14, no. 3, pp. 1843-1854, May 2023, doi: 10.1109/TSG.2022.3213255.

[10] Chen, Y., Liu, W., & Liu, J. (2018). Big data-based fault diagnosis for large-scale wind turbines using feature selection and machine learning. Renewable Energy, 116, 587-599. https://doi.org/10.1016/j.renene.201 7.09.019

[11] Cui, F., Wang, X., & Zhang, J. (2018). Big data analytics in smart grids: A review. Energy Reports, 4, 376-38 3. https://doi.org/10.1016/j.egyr.2018.09.009

[12] Ding, Y., Han, Y., & Wang, K. (2018). A review on data mining for failure diagnosis and prediction in smart grid. Renewable and Sustainable Energy Reviews, 81, 590-602. https://doi.org/10.1016/j.rser.2017.08.04 3

[13] El-Sappagh, S., & Ali, F. (2018). A comprehensive survey on big data analytics in smart grids. Sustainable C ities and Society, 38, 498-512. https://doi.org/10.1016/j.scs.2018.01.024

[14] Fan, C., Sun, H., & Li, Y. (2018). Data-driven short-term load forecasting using a hybrid model based on im proved grey relational analysis and machine learning. Applied Energy, 228, 1987-1999. https://doi.org/10. 1016/j.apenergy.2018.06.124

[15] Feng, C., Wang, L., & Liu, J. (2018). Online fault detection for power systems based on streaming data. Ele ctric Power Systems Research, 158, 98-107. https://doi.org/10.1016/j.epsr.2018.01.026

[16] Syed, D., Zainab, A., Ghrayeb, A., Refaat, S. S., Abu-Rub, H., & Bouhali, O. (2020). Smart grid big data analy tics: Survey of technologies, techniques, and applications. IEEE Access, 9, 59564-59585. doi: 10.1109/ACC ESS.2020.3041178.

[17] Ghosh, S., Chowdhury, S., & Chowdhury, S. P. (2018). Big data analytics for smart grid: A survey. Electric P ower Systems Research, 167, 181-190. https://doi.org/10.1016/j.epsr.2018.10.009

[18] Guo, Y., & Li, J. (2018). Data-driven anomaly detection for smart grid using feature selection and machine learning. International Journal of Electrical Power & Energy Systems, 97, 84-93. https://doi.org/10.1016/j. ijepes.2017.10.011

[19] Hossain, M. S., & Muhammad, G. (2018). Cloud-assisted industrial internet of things (IIoT)–enabled frame work for health monitoring. Computer Networks, 136, 118-126. https://doi.org/10.1016/j.comnet.2018.0 2.019

[20] Huang, B., & Li, H. (2018). Multi-source data fusion for fault diagnosis in smart grids. IEEE Transactions on Industrial Informatics, 14(6), 2577-2586. https://doi.org/10.1109/TII.2018.2794982

[21] Jindal, A., Dua, A., & Sofat, S. (2018). Internet of Things (IoT) and cloud computing for smart grid: A surve y. Energy Reports, 4, 401-409. https://doi.org/10.1016/j.egyr.2018.09.009

[22] Jordehi, A. R. (2018). Optimisation of electric distribution systems: A review. Renewable and Sustainable E nergy Reviews, 82, 3120-3131. https://doi.org/10.1016/j.rser.2017.10.024

[23] Kang, S., & Lee, J. (2018). Big data analytics for smart grid operation and monitoring: Methodologies, chall enges, and opportunities. Renewable and Sustainable Energy Reviews, 82, 1001-1018. https://doi.org/10. 1016/j.rser.2017.09.050

[24] Avancini, D. B., Rodrigues, J. J., Rabêlo, R. A., Das, A. K., Kozlov, S., & Solic, P. (2021). A new IoT-based sma rt energy meter for smart grids. International Journal of Energy Research, 45(1), 189-202.

[25] https://doi.org/10.1002/er.5177

[26] Li, J., & Li, Y. (2018). Data-driven methods for fault diagnosis in smart grids: A review. Energies, 11(7), 183 6. https://doi.org/10.3390/en11071836

[27] Hou, J., Wu, Y., Ahmad, A. S., Gong, H., & Liu, L. (2021). A novel rolling bearing fault diagnosis method bas ed on adaptive feature selection and clustering. Ieee Access, 9, 99756-99767. DOI: 10.1109/ACCESS.2021. 3096723

[28] Huang, K., Wu, S., Li, F., Yang, C., & Gui, W. (2021). Fault diagnosis of hydraulic systems based on deep lea rning model with multirate data samples. IEEE Transactions on neural networks and learning systems, 33 (11), 6789-6801. DOI: 10.1109/TNNLS.2021.3083401

[29] Ma, X., & Liu, J. (2018). Review of fault diagnosis methods for smart grid. Journal of Intelligent & Fuzzy Sys tems, 34(4), 2451-2459. https://doi.org/10.3233/JIFS-169197

[30] Yang, H., Liu, X., Zhang, D., Chen, T., Li, C., & Huang, W. (2021). Machine learning for power system protec tion and control. The Electricity Journal, 34(1), 106881. https://doi.org/10.1016/j.tej.2020.106881

[31] Min, Y., & Kim, J. (2018). Real-time fault detection in smart grids using machine learning. Sensors, 18(7), 2 277. https://doi.org/10.3390/s18072277

[32] Mohammadi, M., Al-Fuqaha, A., Sorour, S., & Guizani, M. (2018). Deep learning for IoT big data and strea ming analytics: A survey. IEEE Communications Surveys & Tutorials, 20(4), 2923-2960. https://doi.org/10. 1109/COMST.2018.2844341

[33] Nair, A. G., & George, S. (2018). Data-driven approaches for fault detection in smart grids: A survey. Inter national Journal of Electrical Power & Energy Systems, 99, 594-601. https://doi.org/10.1016/j.ijepes.2018. 01.031

[34] Nejad, M., & Gharavian, D. (2018). Machine learning applications in smart grid: A review. Renewable and Sustainable Energy Reviews, 82, 1679-1695. https://doi.org/10.1016/j.rser.2017.09.020

[35] Siniosoglou, I., Radoglou-Grammatikis, P., Efstathopoulos, G., Fouliras, P., & Sarigiannidis, P. (2021). A unif ied deep learning anomaly detection and classification approach for smart grid environments. IEEE Transa ctions on Network and Service Management, 18(2), 1137-1151. DOI: 10.1109/TNSM.2021.3078381

[36] Pei, S., & Wang, G. (2018). Big data analytics for power system event detection using machine learning. A pplied Sciences, 8(11), 2197. https://doi.org/10.3390/app8112197

[37] Dawood, B. A., Al-Turjman, F., Hussain, A. A., & Deebak, B. D. (2022). Data protection and privacy preserv ation mechanisms for applications of IoT in smart grids using AI. In Sustainable Networks in Smart Grid (p p. 207-231). Academic Press. https://doi.org/10.1016/B978-0-323-85626-3.00004-1

[38] Shi, Y., & Wang, K. (2018). Big data analytics for smart grid: A survey. IEEE Access, 6, 49279-49289. https://doi.org/10.1109/ACCESS.2018.2869398

[39] Sun, Y., & Zhang, H. (2018). Data-driven methods for smart grid fault detection and diagnosis. Journal of Electrical Engineering & Technology, 13(6), 2284-2291. https://doi.org/10.5370/JEET.2018.13.6.2284

[40] Tang, Y., & Wang, W. (2018). Fault detection in smart grid using machine learning. Applied Sciences, 8(5), 836. https://doi.org/10.3390/app8050836

[41] Wang, J., & Yang, J. (2018). Data-driven approaches for fault detection and diagnosis in power systems. Energies, 11(6), 1515. https://doi.org/10.3390/en11061515

[42] Wu, Y., & Li, X. (2018). Big data analytics for smart grid: A review. Energy Reports, 4, 376-383. https://doi.org/10.1016/j.egyr.2018.09.009

[43] Iqbal, A. (2019). Intrusion Detection in Smart Grid Using Machine Learning Approach. Journal of Computational and Theoretical Nanoscience, 16(9), 3808-3816. DOI: https://doi.org/10.1166/jctn.2019.8254

[44] Yan, Y., Qian, Y., Sharif, H., & Tipper, D. (2018). A survey on smart grid communication infrastructures: Motivations, requirements and challenges. IEEE Communications Surveys & Tutorials, 15(1), 5-20. https://doi.org/10.1109/SURV.2012.021312.00034

[45] Kaplan, H., Tehrani, K., & Jamshidi, M. (2021, August). Fault diagnosis of smart grids based on deep learning approach. In 2021 World Automation Congress (WAC) (pp. 164-169). IEEE. DOI: 10.23919/WAC50355.2021.9559474

[46] Yao, L., & Wang, X. (2018). Feature selection for smart grid fault detection. Applied Energy, 228, 1987-1999. https://doi.org/10.1016/j.apenergy.2018.06.124

[47] Gasparin, A., Lukovic, S., & Alippi, C. (2021). Deep learning for time series forecasting: The electric load case. CAAI Transactions on Intelligence Technology, 7(1), 1–25. https://doi.org/10.1049/cit2.12060

[48] Tao, F., Zhang, H., Liu, A., & Nee, A. Y. C. (2019). Digital twin in industry: State-of-the-Art. IEEE Transactions on Industrial Informatics, 15(4), 2405–2415. https://doi.org/10.1109/tii.2018.2873186

[49] Zhou, B., & Wang, Y. (2018). Data-driven fault analysis for smart grid using machine learning. International Journal of Electrical Power & Energy Systems, 99, 594-601. https://doi.org/10.1016/j.ijepes.2018.01.031

[50] Zhou, Y., & Li, Y. (2018). Data mining for smart grid fault detection. Journal of Electrical Engineering & Technology, 13(6), 2284-2291. https://doi.org/10.5370/JEET.2018.13.6.2284

[51] Zideh, M. J., Chatterjee, P., & Srivastava, A. K. (2023). Physics-Informed Machine Learning for data anomaly Detection, Classification, Localization, and Mitigation: A Review, Challenges, and Path forward. IEEE Access, 12, 4597–4617. https://doi.org/10.1109/access.2023.3347989

[52] Rasheed, A., San, O., & Kvamsdal, T. (2020c). Digital Twin: values, challenges and enablers from a modeling perspective. IEEE Access, 8, 21980–22012. https://doi.org/10.1109/access.2020.2970143

[53] CTao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. Journal of Manufacturing Systems, 48, 157–169. https://doi.org/10.1016/j.jmsy.2018.01.006

[54] Dong, Q., & Yang, Z. (2019). Feature selection for smart grid fault diagnosis. Applied Sciences, 9(12), 2462. https://doi.org/10.3390/app9122462

[55] Fang, X., Misra, S., Xue, G., & Yang, D. (2019). Smart grid–The new and improved power grid: A survey. IEEE Communications Surveys & Tutorials, 14(4), 944-980. https://doi.org/10.1109/SURV.2011.101911.00087

[56] Wang, Y., Chen, Q., Hong, T., & Kang, C. (2018). Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. IEEE Transactions on Smart Grid, 10(3), 3125–3148. https://doi.org/10.1109/tsg.2018.2818167

[57] Colmenares-Quintero, R. F., Quiroga-Parra, D. J., Rojas, N., Stansfield, K. E., & Colmenares-Quintero, J. C. (2021). Big Data analytics in Smart Grids for renewable energy networks: Systematic review of information and communication technology tools. Cogent Engineering, 8(1). https://doi.org/10.1080/23311916.2021.1935410

[58] Zhang, C., Patras, P., & Haddadi, H. (2019d). Deep learning in mobile and wireless Networking: a survey. IEEE Communications Surveys & Tutorials, 21(3), 2224–2287. https://doi.org/10.1109/comst.2019.2904897

[59] Szczepaniuk, H., & Szczepaniuk, E. K. (2022). Applications of artificial intelligence algorithms in the energy sector. Energies, 16(1), 347. https://doi.org/10.3390/en16010347

[60] Liu, L., Wang, B., Ma, F., Zheng, Q., Yao, L., Zhang, C., & Mohamed, M. A. (2022). A concurrent fault diagnosis method of transformer based on graph convolutional network and knowledge graph. Frontiers in Energy Research, 10. https://doi.org/10.3389/fenrg.2022.837553

[61] Luo, Y., & Wu, J. (2019). Data-driven approaches for smart grid fault detection and diagnosis: A review. IEEE Access, 7, 74607-74617. https://doi.org/10.1109/ACCESS.2019.2939201

[62] Zhang, Y., & Li, Y. (2019). Hybrid machine learning for smart grid fault detection. Energies, 12(10), 1945. https://doi.org/10.3390/en12101945

[63] Zhou, T., & Wang, Y. (2019). Big data analytics for smart grid event detection. IEEE Access, 7, 74607-74617. https://doi.org/10.1109/ACCESS.2019.2939201

[64] Zhang, C., Patras, P., & Haddadi, H. (2019c). Deep learning in mobile and wireless Networking: a survey. IEEE Communications Surveys & Tutorials, 21(3), 2224–2287. https://doi.org/10.1109/comst.2019.2904897