Explainable AI for harmonic fingerprinting and voltage sag diagnosis in decentralized power grids: trends, challenges and future directions

Mohd Hatta Jopri[1], Tole Sutikno[2], Yacine Djeghader[3], Mohd Riduan Mohd Shariff[4], Wan Azlan Wan Zainal Abidin[5]

[1]Faculty of Electrical Technology and Engineering, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[2]Department of Electrical Engineering, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[3]Laboratory of Electrical Engineering Electronic and Renewable Energy, Souk-Ahras University, Souk Ahras, Algeria
[4]Plant 1, Maintenance Department, Malaysian Refining Company Sdn. Bhd., Melaka, Malaysia
[5]Department of Electrical and Electronics Engineering, Faculty of Engineering, Universiti Malaysia Sarawak, Sarawak, Malaysia

Article Info

Article history:

Received Jun 3, 2026
Revised Aug 1, 2026
Accepted Aug 15, 2026

Keywords:

Decentralized grids
Explainable artificial intelligence
Harmonic fingerprinting
Physics-informed neural networks
Voltage sag diagnosis

ABSTRACT

The evolution of decentralized power grids has increased the complexity of power-quality monitoring, particularly harmonic fingerprinting and voltage sag diagnosis. Artificial intelligence improves disturbance detection and classification, yet black-box models limit transparency, engineering validation, and operator trust. This review synthesizes 108 selected studies on explainable artificial intelligence (XAI) for power-system diagnostics, focusing on SHapley Additive exPlanations (SHAP), local interpretable model-agnostic explanations (LIME), attention-based interpretability, visual analytics, and physics-informed learning. The review integrates harmonic fingerprinting with voltage sag diagnosis through their shared requirements for source attribution, temporal interpretation, physical consistency, and operator-oriented explanation. Four major deployment gaps are identified: data quality, computational latency, physical grounding, and trustworthiness. Future priorities include real-time embedded XAI, physics-informed neural networks, federated learning, standardized trustworthiness metrics, and adaptive model lifecycle management. The findings indicate that reliable autonomous diagnosis requires explainability to be integrated with predictive performance, electrical-system physics, computational efficiency, and field validation. This integration provides a stronger foundation for transparent, resilient, and trustworthy diagnostic systems in decentralized power grids.

Corresponding Author:

Mohd Hatta Jopri
Faculty of Electrical Technology and Engineering, Universiti Teknikal Malaysia Melaka
Melaka, Malaysia
Email: hatta@utem.edu.my

1. INTRODUCTION

The increasing penetration of renewable energy sources, inverter-based resources, electric vehicles, and distributed generation has changed the behaviour of decentralized power grids. These developments introduce bidirectional power flow, nonlinear operating conditions, variable grid topology, and complex power-quality disturbances [1], [2]. Harmonic distortion and voltage sag remain important concerns because they affect equipment performance, protection reliability, grid stability, and operational continuity [3], [4]. Artificial intelligence and machine learning have improved disturbance detection, classification, and source localization [5], [6]. Yet many advanced models remain difficult to interpret, limiting operator trust, technical validation, and practical deployment in critical electrical infrastructure [7], [8]. Explainable artificial intelligence (XAI) addresses this limitation by linking model predictions with interpretable evidence such as feature importance, temporal relevance, spectral contribution, and physical operating conditions [9], [10].

Harmonic fingerprinting and voltage sag diagnosis share common requirements for source attribution, temporal interpretation, physical consistency, and operator-oriented explanation [11], [12]. Existing studies report strong diagnostic performance, yet persistent limitations remain in dataset quality, validation consistency, computational burden, reproducibility, and physical grounding [13]–[15]. Existing reviews also tend to examine XAI, harmonic analysis, and voltage sag diagnosis as separate research streams, leaving limited synthesis of their shared diagnostic requirements in decentralized grids. This review addresses this gap through three contributions. First, it integrates XAI developments across harmonic fingerprinting and voltage sag diagnosis within a common power-system context. Second, it critically compares explainability methods according to interpretability, computational requirements, physical consistency, and deployment relevance. Third, it identifies technical gaps and future priorities involving real-time XAI, physics-informed learning, federated intelligence, trustworthiness metrics, and adaptive operation.

To improve the transparency of the review process, relevant literature was identified from Scopus, Web of Science, IEEE Xplore, and ScienceDirect. The primary search covered publications from January 2018 to June 2026, while earlier seminal studies were retained where they provided fundamental concepts or established diagnostic methods. Search terms combined “explainable artificial intelligence”, “XAI”, “harmonic fingerprinting”, “harmonic source identification”, “voltage sag diagnosis”, “power quality disturbance”, “decentralized grid”, “microgrid”, “SHAP”, “LIME”, “physics-informed neural network”, and “interpretable machine learning”. Boolean combinations of these terms were applied to identify studies related to explainable power-system diagnostics. Articles were screened according to relevance to harmonic or voltage-sag analysis, methodological clarity, diagnostic contribution, explainability, physical interpretation, applicability to decentralized power systems. Peer-reviewed journal articles and conference papers were prioritized. Duplicate records, unrelated AI applications, studies without sufficient technical relevance, non-substantive publications were excluded. Earlier established works were retained where necessary to preserve the technical development of the field. This structured approach provides a focused basis for examining current XAI trends, technical limitations, comparative research evidence, and future research priorities for harmonic fingerprinting and voltage sag diagnosis in decentralized power grids.

2. CURRENT TREND IN XAI POWER SYSTEM

Current trends in explainable artificial intelligence for power systems show a transition from opaque predictive models toward diagnostic architectures that expose how individual electrical variables influence model decisions. This development is important in decentralized grids where renewable generation, inverter-based resources, bidirectional power flow, and changing operating conditions increase the complexity of harmonic fingerprinting and voltage sag diagnosis [16]–[22]. Figure 1 organizes current XAI techniques according to explanation scope, model dependence, and application level. At the highest level, the taxonomy separates global interpretability, which explains overall model behaviour, from local interpretability, which explains individual predictions. It then distinguishes model-agnostic post-hoc approaches such as partial dependence, accumulated local effects, surrogate models, SHAP, LIME, and counterfactual explanations from model-specific intrinsic approaches such as tree-based models, attention mechanisms, gradient methods, DeepLIFT, and layer-wise relevance propagation.

The taxonomy also links these methods to power-system tasks such as visualizing network attention, identifying dominant features, decomposing voltage-sag signatures, sensitivity analysis, and classifying disturbances [23]–[37]. This structure shows that no single XAI method satisfies every engineering requirement. Global methods support model-level understanding, while local methods are more suitable for event-specific diagnosis. Model-agnostic techniques offer flexibility across architectures, while intrinsic methods provide closer integration with the predictive model. Figure 1 therefore provides the conceptual basis for evaluating XAI approaches according to interpretability, diagnostic usefulness, computational demand, and physical relevance in decentralized power-system applications.

2.1. The paradigm shift from black-box models to glass-box transparency

The transition from black-box models to glass-box transparency reflects a major change in power-system AI. Deep learning models have achieved high diagnostic accuracy, including reported fault-classification performance of 98.7% [17]. Yet prediction accuracy alone is insufficient for decentralized grids where operators need to understand why a model identifies a disturbance, which electrical variables influence the result, whether the explanation remains consistent with physical system behaviour [16], [18]. This requirement has increased interest in explainable and hybrid architectures that expose model reasoning instead of presenting only final classifications [19].

Current research therefore places greater emphasis on epistemic trust, explanation fidelity, and physical consistency [20]–[22]. A reliable XAI model should show that its prediction is driven by relevant voltage, current, harmonic, and temporal features rather than incidental correlations or noise. Explainability scores and trustworthiness metrics have emerged to assess whether model decisions remain stable, interpretable, and technically meaningful [22]. For harmonic fingerprinting and voltage sag diagnosis, this shift is important because the diagnostic explanation must support source identification, event interpretation, and operator action under changing grid conditions.

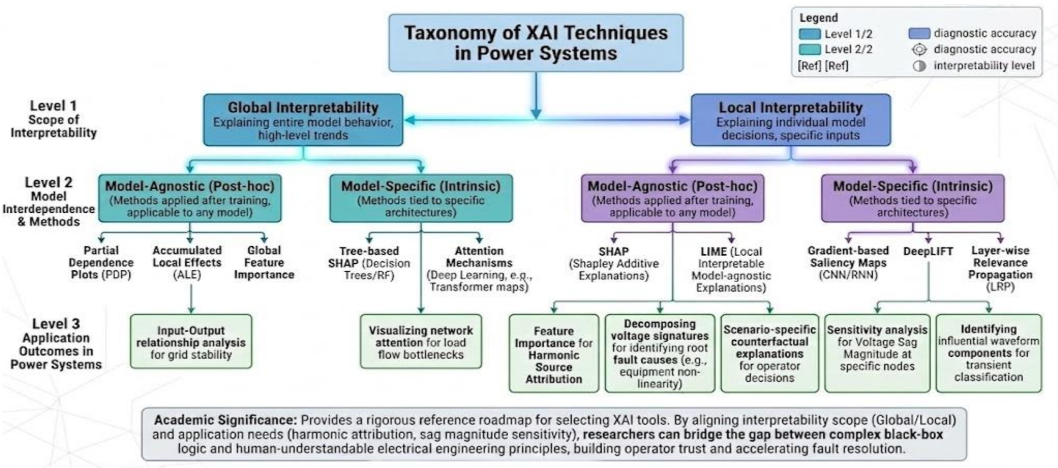

Figure 1. Taxonomy of XAI techniques in power systems

2.2. Dominance of feature-level explanations: SHAP and LIME

Feature-level explanation methods remain prominent in XAI-based power-system diagnostics because they translate complex model outputs into identifiable electrical variables. SHapley Additive exPlanations (SHAP) provides global and local feature attribution, while local interpretable model-agnostic explanations (LIME) focuses on local behaviour around individual predictions [23], [24]. In fault and power-quality applications, these methods identify influential variables such as voltage magnitude, current derivative, harmonic components, and disturbance duration [25]. Reported studies also demonstrate strong classification performance across several disturbance classes, including F1-scores approaching 0.99 in microgrid applications [26]. These values require interpretation within their respective datasets, operating conditions, and validation protocols.

Figure 2 illustrates an XAI-enhanced voltage sag diagnosis pipeline from raw three-phase voltage-current measurements to operator-oriented decision support. The sampled waveforms are normalized before a CNN-LSTM architecture extracts spatial-temporal features. SHAP or LIME then evaluates the model output to quantify the contribution of diagnostic variables. The illustrated local explanation identifies df/dt as the strongest trigger, followed by voltage magnitude and harmonic distortion. This explanation is subsequently translated into operator-understandable information for manual review or automated protection action. SHAP offers stronger consistency for feature attribution, while LIME provides computationally lighter local explanations [27], [28]. The pipeline therefore demonstrates how feature-level XAI links deep-learning predictions with engineering interpretation, root-cause confirmation, and protection decisions.

2.3. Attention-based interpretability and visual analytics

Attention-based models offer a more intrinsic form of interpretability by identifying the signal regions that contribute most strongly to a prediction. Transformer and vision transformer architectures apply self-attention to emphasize relevant temporal, spectral or spatial features within power-quality data [29], [30]. Reported studies show classification accuracy reaching 99.26% with optimized relevance maps, while spectrogram-based vision transformer approaches have achieved values up to 99.38% [31]–[33]. These results indicate strong diagnostic capability, although direct comparison remains inappropriate because datasets, disturbance classes, sampling conditions and validation protocols differ.

Visual analytics further supports engineering interpretation through Grad-CAM, saliency maps, relevance maps and latent-space visualization [34]–[37]. For harmonic fingerprinting, these techniques reveal spectral regions associated with specific harmonic signatures. For voltage sag diagnosis, they highlight temporal regions linked to event initiation, duration, and recovery. Their main advantage is the ability to present model reasoning in a form suitable for expert inspection. Limitations remain in noisy attention patterns, unstable visual explanations, and possible emphasis on features without physical significance. Reliable deployment therefore requires visual explanations to be validated against electrical-system behaviour rather than interpreted as direct evidence of causality.

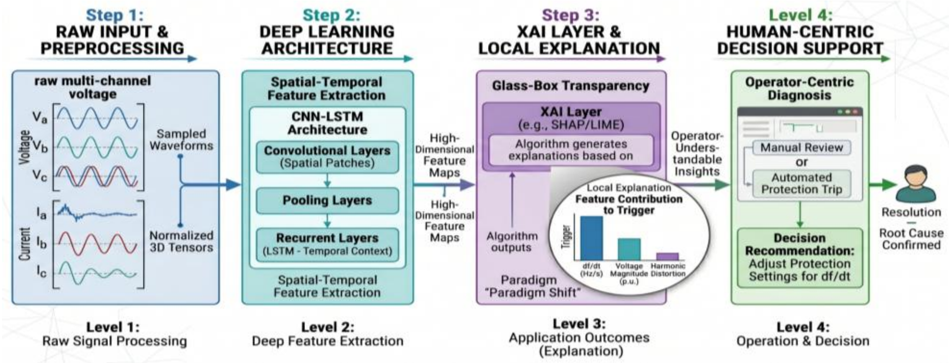

Figure 2. XAI-enhanced voltage SAG diagnosis pipeline

2.4. Convergence with decentralized grids and physics-informed AI

The convergence of XAI with decentralized power grids increasingly focuses on physically constrained learning, especially under bidirectional power flow, inverter-dominated generation, and changing network topology [38], [39]. Harmonic fingerprinting in such systems requires more than identifying total harmonic distortion or spectral patterns. The diagnostic model must relate harmonic signatures to plausible electrical sources while respecting network laws and operating constraints [40], [41]. Physics-informed approaches address this issue by embedding Kirchhoff's Current Law, Kirchhoff's Voltage Law, line impedance, and related physical relationships into the learning process. This reduces the risk of statistically accurate outputs that remain physically inadmissible.

Figure 3 illustrates this distinction between a standard data-driven neural network and a physics-informed neural network framework. The conventional model processes voltage-current harmonic inputs without explicit physical constraints, allowing predictions to fall outside the physically admissible solution space. In contrast, the PINN combines harmonic data with grid-model constraints through a physics-informed loss function. Related studies demonstrate the value of interpretable analysis for examining power-system stability and identifying influential operating factors [42], [43]. In broader engineering applications, structured optimization through Taguchi analysis has also been applied to determine influential parameters affecting system stability [44]. These approaches reinforce the importance of systematic parameter assessment when developing technically reliable models.

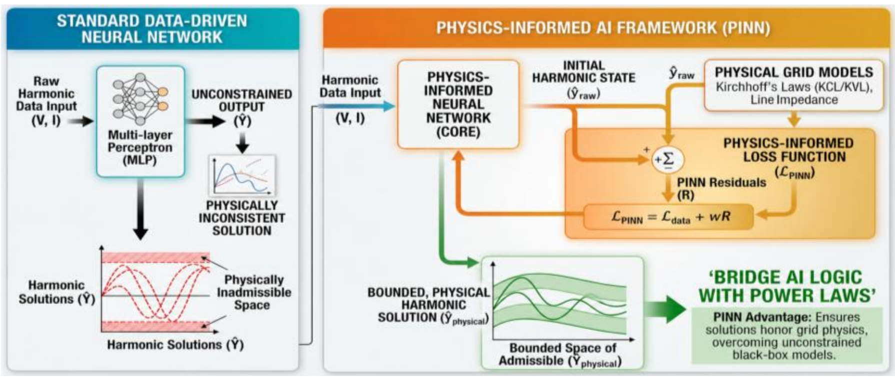

Figure 3. Physics-informed neural network (PINN) framework for harmonics

The resulting physics-informed framework penalizes violations of governing electrical laws and directs predictions toward physically consistent operating conditions. Interpretable learning has also been extended to active corrective control in power systems, demonstrating its relevance to operational decision support [45]. The integration of data-driven intelligence, physical constraints, and systematic parameter assessment therefore strengthens diagnostic credibility, robustness, and operational trust for harmonic fingerprinting and voltage sag diagnosis in decentralized power grids.

3. CHALLENGES AND LIMITATIONS

The deployment of explainable artificial intelligence for harmonic fingerprinting and voltage sag diagnosis faces interconnected technical, physical, data-related, and human-centred limitations. High-performance deep-learning models often require complex architectures, extensive data processing, and post-hoc explanation layers. These requirements increase computational latency, reduce transparency, and complicate real-time implementation in decentralized grids. Other persistent issues include limited real-world datasets, inconsistent disturbance labels, changing grid topology, weak reproducibility, and explanations that reflect statistical correlation rather than electrical causality. These limitations become critical when diagnostic outputs support protection, control, and service-continuity decisions.

Figure 4 illustrates the trust-accuracy trade-off across different AI architectures. Conventional models occupy different positions between predictive accuracy and human-centric interpretability, while deeper CNN-LSTM architectures increase model complexity but create a wider trust gap. The XAI-augmented pathway introduces an interpretability layer between predictive models and engineering users, aiming to preserve diagnostic accuracy while increasing explanation quality. The proposed XAI-CNN-LSTM position represents this balance by combining deep-learning performance with stronger interpretability. The figure therefore emphasizes a central challenge for XAI deployment: increasing transparency without degrading diagnostic performance, computational responsiveness, and engineering usefulness. The following subsections examine this challenge through data quality, real-time processing, physical consistency, trust, and standardization.

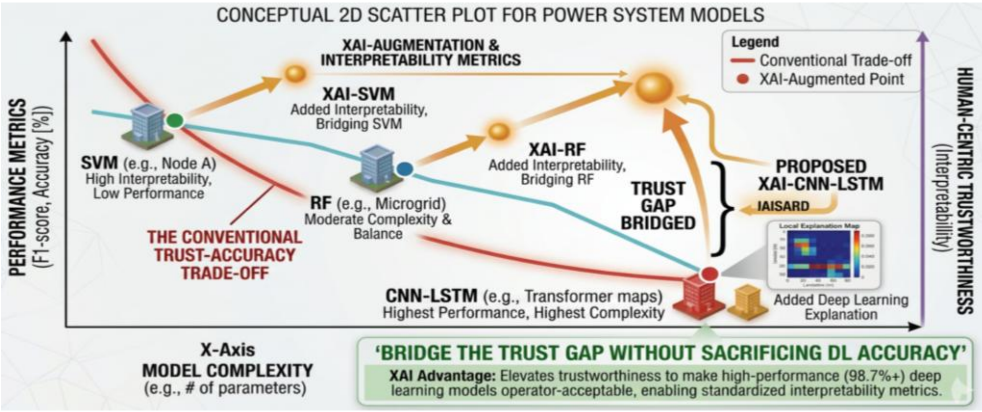

Figure 4. The trust-accuracy trade-off and interpretability metrics

3.1. Data scarcity, quality, and labeling complexities

A major limitation in XAI-based harmonic fingerprinting is the scarcity of high-quality, diverse, consistently labelled datasets [46]. Simulated data remain widely used because real measurements from decentralized grids are noisy, site-dependent, and difficult to generalize [47]. In photovoltaic installations, harmonic spectral patterns often require manual post-processing to separate background distortion from inverter-related components. High-resolution acquisition also creates practical constraints. Harmonic source localization across multiple feeders may require sampling rates above 28.8 kHz, increasing communication bandwidth, storage demand, and processing burden [48]. These conditions encourage localized data silos, limiting model transferability across different microgrids [49], [50].

Ground-truth labelling presents a second challenge. Harmonic distortion often results from multiple simultaneous sources, including nonlinear industrial loads, converters, and EV chargers [51]. This makes source attribution more difficult than discrete fault diagnosis. Overlapping signatures increase labelling uncertainty, while rare voltage sag events and resonance conditions create severe class imbalance [52]–[54]. These limitations affect both predictive accuracy and explanation reliability. An XAI model trained on incomplete or weakly labelled data may produce convincing feature attributions without reflecting the actual physical disturbance mechanism. Dataset diversity, transparent labelling procedures, and field validation therefore remain essential for reproducible and trustworthy diagnosis.

3.2. Technical limitations of XAI and real-time processing constraints

The main technical limitation of XAI lies in the trade-off between predictive performance, explanation fidelity, and computational demand. Deep architectures such as Transformers and TCN-iTransformers capture nonlinear harmonic behaviour effectively, yet their complexity often requires post-hoc methods such as SHAP or LIME to explain individual predictions [55], [56]. These explanation layers introduce additional processing time. This becomes critical in voltage sag diagnosis because protection and control functions operate within short response windows, often within 3 to 10 cycles [57]. A model that produces an accurate diagnosis but generates its explanation too slowly offers limited value for real-time autonomous operation.

Explanation fidelity also remains inconsistent. LIME is generally faster than SHAP, but local explanations may not represent the broader behaviour of the power system [58]. Both methods also face difficulty when applied to multivariate time-series containing coupled temporal and frequency-domain characteristics. Harmonic fingerprinting requires identification of when a disturbance emerges, which spectral components dominate, and how long the condition persists. Existing XAI methods often provide feature importance without sufficient temporal resolution [59], [60]. Future implementation therefore requires lighter explanation architectures, hardware-aware processing, and temporal attribution methods that preserve both diagnostic speed and engineering relevance.

3.3. The gap between AI logic and physical power system laws

A critical limitation of current XAI frameworks is the gap between statistical model reasoning and the physical laws governing electrical networks. Most XAI methods explain correlations learned from data, but correlation does not establish electrical causality [61]. A model may associate a voltage deviation at one bus with a harmonic event while the actual source originates from a remote inverter-based resource or another nonlinear load [62]. Without physical constraints, explanations may appear mathematically consistent within the model while remaining inconsistent with network behaviour, power-flow relationships, circuit laws [63], [64]. More broadly, engineering studies have shown that measurable performance characteristics often exhibit identifiable correlations with underlying physical properties [65]. Physics-informed approaches seek to reduce this gap by embedding electrical constraints into model development and interpretation [66]. This alignment is essential when diagnostic outputs are expected to support technically defensible decisions under complex and changing operating conditions.

The challenge becomes greater in decentralized grids because bidirectional power flow, distributed generation, islanding, and network reconfiguration continuously alter system topology [67], [68]. Explanations developed under a fixed topology may lose validity after reconfiguration, reducing operator trust and diagnostic reliability [69], [70]. Reliable XAI therefore requires closer alignment between learned model behaviour and physically consistent power-system conditions, supported by validation across multiple operating scenarios and network configurations.

3.4. Human-centric challenges: trust, bias, and standardization

Human-centred limitations remain a major barrier to the practical adoption of XAI in power-system diagnostics. There is no widely accepted standard defining what constitutes a useful explanation for power engineers [71]. A numerical feature attribution or SHAP value may satisfy a data-science requirement, yet it offers limited value if the operator cannot relate it to a protection action, disturbance source, or operating risk [72]. This gap complicates comparison between XAI methods because interpretability, stability, robustness, and engineering relevance are assessed using different criteria [73].

Trust is also affected by cognitive bias. Automation bias may lead operators to accept plausible but incorrect explanations, while skepticism bias may cause rejection of valid recommendations when explanations appear too complex or unfamiliar [74], [75]. Visual and feature-attribution methods may emphasize unstable or less relevant signal regions when input measurements are noisy, incomplete, or weakly informative [76], [77]. These risks become more serious in autonomous systems where explanations influence switching, protection, and corrective control. The absence of mature accountability frameworks further limits deployment because responsibility for incorrect AI-supported decisions remains unclear [78], [79]. Standardized explainability metrics, operator-oriented interfaces, and validation against engineering judgement are therefore required to support trustworthy XAI implementation.

3.5. Summary of key limitations

The reviewed literature identifies four persistent barriers to trustworthy XAI deployment in decentralized power grids. The data gap arises from limited real-world measurements, inconsistent labels, and class imbalance [79], [80]. The computational gap reflects latency from post-hoc explanations in time-critical voltage sag diagnosis [81], [82]. The physical gap concerns explanations that lack clear consistency with electrical causality, network topology, and governing laws [83], [84]. The trust gap results from inconsistent interpretability metrics, operator bias, and limited standardization [85], [86]. Addressing these barriers requires efficient, physics-informed, field-validated XAI frameworks that support reliable harmonic fingerprinting and voltage sag diagnosis.

4. COMPARISON OF PAST RESEARCH

Past research shows a clear progression from high-accuracy diagnostic models toward explainable, physics-aware approaches. Mohammadi *et al.* [52] reported strong voltage sag source detection performance, while Balouji *et al.* [48] applied high-resolution sampling with attention-LSTM for harmonic localization. Yao *et al.* [68] extended this direction to bidirectional microgrid conditions using a two-stage spatio-temporal approach. Explainability was later strengthened through CNN-LSTM with SHAP [17], formal explainability metrics [8], transformer relevance maps [22], and LSTM-SHAP disturbance classification [87]. More recent studies have placed greater emphasis on physical consistency and operational interpretation through epistemically informed harmonic analysis [20], TCN-iTransformer modelling [56], and LIME-based voltage stability analysis [28]. Reported performance values should not be interpreted as direct comparisons because the studies employ different datasets, disturbance classes, sampling conditions, grid configurations, and validation procedures.

Conventional signal-processing and hybrid approaches remain relevant despite the increasing use of XAI. Time-frequency analysis, wavelet-based methods, and engineered electrical features provide deterministic representations with lower computational complexity for disturbance detection and embedded implementation [13], [26]. Hybrid architectures that combine physically meaningful signal features with machine learning therefore remain valuable when interpretability, computational efficiency, and real-time operation are prioritized.

Table 1 (see Appendix) [8], [12], [17], [20], [22], [28], [48], [52], [56], [68], [87] presents ten representative studies selected to illustrate the major methodological trajectories in XAI-based harmonic fingerprinting and voltage sag diagnosis. These studies do not represent the entire evidence base but provide a focused comparison of dominant techniques, key innovations, current limitations, and future research directions across the field. The comparison shows that predictive accuracy alone does not resolve the main barriers to practical deployment. Computational complexity remains significant for advanced architectures [17], [56], while generalization across grid configurations remains insufficiently demonstrated [68], [87]. Explanation reliability and robustness also require stronger validation [8], [22]. High-resolution harmonic monitoring introduces substantial data acquisition, storage, and communication requirements [48], while physics-aware approaches remain difficult to scale across complex decentralized networks [20]. The comparative evidence therefore indicates a shift from accuracy-centred evaluation toward interpretability, reproducibility, physical consistency, real-time suitability, and deployment readiness [28], [52].

The reviewed studies use simulated, laboratory and field datasets with varying sampling rates, features, disturbance classes, validation protocols and grid configurations. These differences limit reproducibility and direct comparison. Practical deployment therefore requires cross-site validation, independent replication, hardware-in-the-loop testing and field evaluation. Trustworthy diagnosis ultimately depends on balancing predictive performance, explainability, physical consistency, computational efficiency, and real-world validation.

5. FUTURE PERSPECTIVE AND RESEARCH DIRECTIONS

Future research should prioritize real-time embedded XAI with lower computational latency [88]–[91], stronger physics-informed constraints [92]–[96], federated learning for decentralized data [97], [98], and standardized trustworthiness metrics [99], [100]. Adaptive updating, recalibration, topology changes and concept drift also require attention throughout the model lifecycle. Human-AI interfaces should translate explanations into operator-relevant information [101], [102]. Future systems should also support resilience and service continuity during islanding, abnormal operating conditions, network reconfiguration and degraded communication environments.

5.1. Development of real-time and embedded XAI architectures

Real-time deployment remains constrained by the computational overhead of post-hoc methods such as SHAP and LIME [88], [89]. Protection-oriented applications require explanations within short operating windows, making latency a critical design requirement [90]. Future work should therefore focus on intrinsically interpretable models, lightweight XAI modules, hardware acceleration, edge deployment through micro-PMUs or smart inverters [91]. Knowledge distillation also offers a practical route for transferring complex model behaviour into smaller architectures suitable for real-time harmonic fingerprinting and voltage sag diagnosis.

5.2. Advanced fusion of physics-informed neural networks with XAI

Future XAI frameworks should integrate physics-informed neural networks to reduce the gap between statistical correlation and electrical causality [92]. Embedding Kirchhoff's laws, energy conservation, and grid-topology constraints within the learning process helps ensure model explanations remain physically admissible

[93], [94]. This approach is especially relevant for harmonic source identification in PV-rich decentralized grids, where overlapping distortion sources complicate diagnosis [95], [96]. The main research priority is therefore to combine explainability with scalable physics-based constraints without imposing excessive computational complexity.

5.3. Decentralized federated learning and data privacy

Federated learning supports collaborative model development across decentralized grids without sharing raw operational data [50], [97]. Figure 5 illustrates local harmonic data processed through edge PINN models, where physical constraints and XAI explanations are retained at each site. Only masked model updates are exchanged with the central aggregator, supporting data sovereignty while improving cross-site learning [98]. Future work should strengthen reproducibility through common aggregation protocols, independent multi-site validation, and consistent explainability assessment. The model lifecycle should also include periodic updating, recalibration, and drift monitoring as grid topology or operating conditions change.

5.4. Standardization of trustworthiness metrics and human-AI interfaces

Standardized trustworthiness metrics are required before XAI becomes suitable for utility-grade deployment [99]. Future evaluation should measure explanation clarity, stability, fidelity, robustness, and engineering relevance rather than predictive accuracy alone [100]. Human-AI interfaces should also translate complex outputs such as SHAP values, attention maps, and latent-space features into actionable information for grid operators [101], [102]. This is especially important for harmonic fingerprinting and voltage sag diagnosis, where explanations must support rapid engineering judgement, reduce automation bias, and remain meaningful under changing operating conditions.

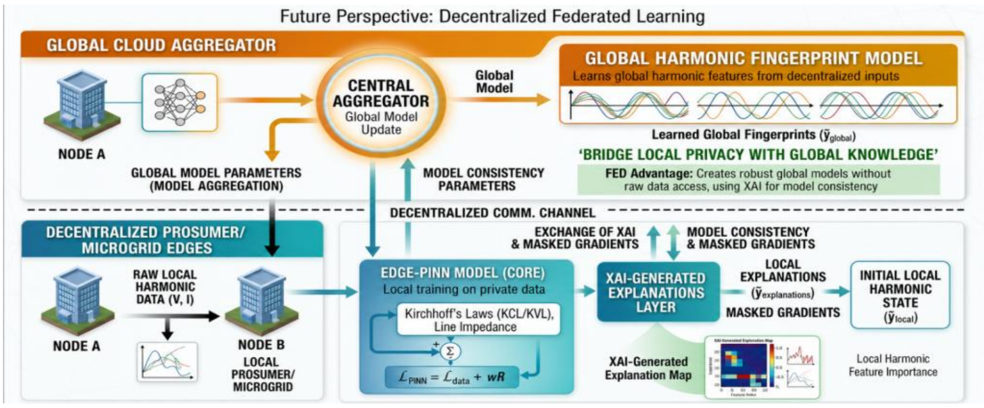

Figure 5. Decentralized federated XAI architecture

6. RESULTS AND DISCUSSION

The reviewed literature demonstrates a clear transition from accuracy-driven black-box models toward explainable and physics-aware diagnostic systems for decentralized power grids [17], [20], [22]. CNN-LSTM models support high diagnostic performance with feature-level explanation through SHAP [17], while Transformer-based approaches use relevance mapping and attention mechanisms to expose influential signal regions [22]. Harmonic localization studies demonstrate the value of high-resolution measurements for feeder-level diagnosis [48], whereas machine-learning methods support voltage sag source detection under different operating conditions [52]. Reported performance values should be interpreted within each study's dataset, disturbance classes, sampling conditions, and validation procedures rather than as direct numerical comparisons [22], [48], [52].

The evidence also shows that diagnostic performance alone is insufficient for engineering use. Epistemically informed harmonic analysis emphasizes the need to relate model outputs to physically meaningful electrical behaviour [20], while explainable voltage-stability analysis links interpretable features with preventive control decisions [28]. Advanced architectures such as TCN-iTransformer improve modelling capability but introduce higher computational complexity [56]. LSTM-SHAP approaches further illustrate the trade-off between classification performance and explanation overhead [87]. These findings support evaluation frameworks that consider prediction quality, explanation fidelity, and operational relevance together [8], [28], [87].

Harmonic fingerprinting and voltage sag diagnosis share several diagnostic requirements. Harmonic source localization depends on reliable attribution when multiple sources overlap or network conditions change [48], [70], while voltage sag diagnosis requires robust identification of event origin and propagation [52], [67], [68]. Both applications also require temporal interpretation, physical consistency, and explanations that support engineering judgement [20], [28], [71]–[73]. These common requirements support their treatment within a unified XAI diagnostic perspective rather than as isolated application domains.

Physics-informed XAI provides a direct route for reducing the gap between statistical association and electrical causality. Explainable graph-based methods incorporate power-system structure into diagnostic reasoning [92], while physics-informed machine-learning frameworks embed system constraints and trustworthiness principles into power-system analysis [93]. Studies on physics-informed neural networks identify both technical potential and present limitations for smart-grid applications [94]. Data-mechanism fusion has been applied to harmonic source localization [95] While hybrid deep-learning frameworks have been proposed for photovoltaic harmonic source identification [96]. Collectively, this evidence supports closer alignment between model explanations, network topology, and governing electrical relationships [92]–[96].

Federated learning provides a complementary pathway for decentralized implementation. Survey evidence shows that distributed smart-grid sites can support collaborative model development without centralized collection of raw operational data [50]. Federated learning combined with attention-based deep learning and XAI has been investigated for secure and interpretable smart-grid load forecasting [97]. Edge-based federated learning also demonstrates distributed implementation potential for energy applications such as non-intrusive load monitoring [98]. For power-quality diagnosis, practical value will depend on consistent aggregation protocols, cross-site validation, privacy protection, and stable explanation quality under changing grid conditions [50], [97], [98].

Four deployment gaps remain prominent across the reviewed literature. The data gap reflects limited real-world measurements, inconsistent labels, and class imbalance [79], [80]. The computational gap arises from the processing burden associated with advanced models and explanation methods [81], [82]. The physical gap concerns weak alignment between learned model reasoning and actual electrical-system behaviour [83], [84]. The trust gap involves inconsistent explainability metrics, operator confidence, and limited standardization [85], [86]. Deployment readiness should therefore be assessed through predictive performance, explanation quality, physical validity, computational efficiency, and field validation [79]–[86].

Broader smart-grid studies reinforce these findings. Predictive electricity-management platforms show the expanding role of AI in operational energy decision support [103], while XAI-based smart-grid stability assessment demonstrates the value of interpretable models for identifying factors associated with system stability [104]. Reviews of AI techniques for photovoltaic systems stress systematic evaluation and benchmarking across operating conditions [105]. Research on explainability in knowledge-based and machine-learning smart-grid models further highlights transparent reasoning as an important requirement for dependable decision support [106].

Digital-twin research extends the discussion from isolated diagnostic models toward adaptive grid representations. Reinforcement-learning-based digital twins have been investigated for extracting network structures and load patterns in distribution-system planning and operation [107]. Explainable, interpretable, and trustworthy AI has also been examined within digital-twin environments to improve transparency and lifecycle reliability [108]. These developments indicate that future XAI-based power-quality diagnosis should interact with wider digital-grid architectures capable of updating system representations as operating conditions evolve [107], [108].

Future research should prioritize lightweight and embedded XAI for time-critical operation [88]–[91], stronger integration between explainability and physics-informed learning [92]–[96], federated intelligence for distributed learning and data privacy [97], [98], standardized trustworthiness metrics [99], [100], and operator-oriented interfaces [101], [102]. These priorities address complementary requirements for speed, physical credibility, distributed operation, and human interpretation. Despite substantial progress, much of the current evidence still comes from controlled, simulated, or limited validation settings [88], [94]. Stronger cross-site validation, real-time evaluation, topology adaptation, operator interaction, and long-term reliability assessment are therefore required before XAI-based power-quality diagnosis can be regarded as deployment-ready for real decentralized grids [47], [83], [94].

7. CONCLUSION

This review examined the development of explainable artificial intelligence for harmonic fingerprinting and voltage sag diagnosis in decentralized power grids. The findings show a clear transition from black-box prediction toward diagnostic approaches that place greater emphasis on transparency, engineering interpretation, physical consistency, and operator trust. A central finding is that harmonic fingerprinting and voltage sag diagnosis share several common diagnostic requirements. Both require reliable source attribution, temporal interpretation, physically meaningful reasoning, and clear operator-oriented explanation. This commonality supports their integration within a unified XAI perspective rather than treating them as isolated diagnostic problems.

The review also identifies four major barriers to practical deployment: data quality, computational latency, physical grounding, and trustworthiness. Despite strong progress at the model-development level, evidence of deployment readiness in real decentralized grids remains limited. Current studies still require stronger cross-site validation, real-time evaluation, topology adaptation, operator interaction, and long-term reliability assessment. Physics-informed XAI provides a promising direction because it strengthens the relationship between model reasoning and electrical laws. Federated learning also supports decentralized intelligence by enabling collaborative model development without direct exchange of raw operational data. Standardized trustworthiness metrics and operator-centred interfaces remain equally important for consistent engineering use. The main contribution of this review is the integration of harmonic fingerprinting and voltage sag diagnosis within a common explainable diagnostic framework. It also identifies the technical and operational priorities required for future XAI systems, including real-time capability, physical consistency, adaptive model management, distributed learning and field validation. These priorities provide a clearer pathway toward transparent, reliable and deployment-ready power-quality diagnostic systems for decentralized grids.

ACKNOWLEDGMENTS

Authors gratefully acknowledge Universiti Teknikal Malaysia Melaka (UTeM) and the Ministry of Higher Education Malaysia (MOHE) for their support. Authors also express their sincere appreciation to the Faculty of Electrical Technology and Engineering (FTKE) and the Socio-Technical Competencies for Community Empowerment (SoTeCH) Research Group, both under UTeM, for their institutional and research support.

FUNDING INFORMATION

This work was supported by Universiti Teknikal Malaysia Melaka (UTeM) and the Ministry of Higher Education Malaysia (MOHE). The authors gratefully acknowledge the institutional and research support provided for the preparation and completion of this study.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mohd Hatta Jopri	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Tole Sutikno		✓				✓		✓	✓	✓	✓			
Yacine Djeghader				✓		✓	✓			✓	✓		✓	
Mohd Riduan Mohd Shariff			✓	✓		✓	✓			✓	✓		✓	
Wan Azlan Wan Zainal Abidin		✓		✓	✓		✓			✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] S. Kul, B. Arslan, and S. Sagiroglu, "Explainable AI in smart grid applications," in *2025 IEEE 19th International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG)*, May 2025, pp. 1–6, doi: 10.1109/CPE-POWERENG63314.2025.11027305.

[2] R. Alsaigh, R. Mehmood, and I. Katib, "AI explainability and governance in smart energy systems: a review," *Frontiers in Energy Research*, vol. 11, Jan. 2023, doi: 10.3389/fenrg.2023.1071291.

[3] B. E. Altun, F. Alpsalaz, H. Uzel, and Y. Türkay, "Explainable DL based classification for power quality disturbances in renewable-energy-integrated distribution networks," *IET Renewable Power Generation*, vol. 20, no. 1, 2026, doi: 10.1049/rpg2.70269.

[4] M. N. Dehaghani, T. Korõtko, and A. Rosin, "AI applications for power quality issues in distribution systems: a systematic review," *IEEE Access*, vol. 13, pp. 18346–18365, 2025, doi: 10.1109/ACCESS.2025.3533702.

[5] Y. Liu, D. Yuan, H. Fan, T. Jin, and M. A. Mohamed, "A multidimensional feature-driven ensemble model for accurate classification of complex power quality disturbance," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023, doi: 10.1109/TIM.2023.3265756.

[6] I. S. Samanta *et al.*, "Artificial intelligence and machine learning techniques for power quality event classification: a focused review and future insights," *Results in Engineering*, vol. 25, 2025, doi: 10.1016/j.rineng.2024.103873.

[7] J. Yang, W. Chen, X. Xiao, and Z. Zhang, "Explainable deep learning method for power system stability evaluation with incomplete voltage data based on transfer learning," *Measurement: Journal of the International Measurement Confederation*, vol. 247, 2025, doi: 10.1016/j.measurement.2025.116781.

[8] R. Machlev, M. Perl, J. Belikov, K. Y. Levy, and Y. Levron, "Measuring explainability and trustworthiness of power quality disturbances classifiers using XAI—explainable artificial intelligence," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5127–5137, Aug. 2022, doi: 10.1109/TII.2021.3126111.

[9] S. Yang, T. Shan, and X. Yang, "Interpretable DWT-1DCNN-LSTM Network for power quality disturbance classification," *Energies*, vol. 18, no. 2, 2025, doi: 10.3390/en18020231.

[10] I. Hassan *et al.*, "Explainable deep learning model for grid-connected photovoltaic system performance assessment for improving system reliability," *IEEE Access*, vol. 12, pp. 120729–120746, 2024, doi: 10.1109/access.2024.3452778.

[11] G. Anis *et al.*, "Smart and transparent grid stability prediction for efficient energy management using explainable AI," *Energy Strategy Reviews*, vol. 64, p. 102083, Mar. 2026, doi: 10.1016/j.esr.2026.102083.

[12] M. R. Shadi, H. Mirshekali, and H. R. Shaker, "Explainable artificial intelligence for energy systems maintenance: a review on concepts, current techniques, challenges, and prospects," *Renewable and Sustainable Energy Reviews*, vol. 216, 2025, doi: 10.1016/j.rser.2025.115668.

[13] N. M. Rodrigues, F. M. Janeiro, and P. M. Ramos, "Deep learning for power quality event detection and classification based on measured grid data," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, 2023, doi: 10.1109/TIM.2023.3293555.

[14] Y. Li, J. Cao, Y. Xu, L. Zhu, and Z. Y. Dong, "Deep learning based on transformer architecture for power system short-term voltage stability assessment with class imbalance," *Renewable and Sustainable Energy Reviews*, vol. 189, p. 113913, Jan. 2024, doi: 10.1016/j.rser.2023.113913.

[15] S. Gawde, S. Patil, S. Kumar, P. Kamat, K. Kotecha, and A. Abraham, "Multi-fault diagnosis of Industrial rotating machines using data-driven approach : a review of two decades of research," *Engineering Applications of Artificial Intelligence*, vol. 123, 2023, doi: 10.1016/j.engappai.2023.106139.

[16] F. Henao, R. Edgell, A. Sharma, and J. Olney, "AI in power systems: a systematic review of key matters of concern," *Energy Informatics*, vol. 8, no. 1, p. 76, Jun. 2025, doi: 10.1186/s42162-025-00529-1.

[17] D. Shah and M. Gupta, "Explainable artificial intelligence (XAI) for fault detection in smart grid distribution systems: a SHAP-based CNN-LSTM approach," *International Journal For Multidisciplinary Research*, vol. 8, 2026, doi: 10.36948/ijfmr.2026.v08i02.71195.

[18] M. Massaoudi, H. Abu-Rub, and A. Ghrayeb, "Building trust by design through explainable AI for resilient and cognitive smart grids," *AI and Digitalization in Energy Management*, pp. 329–355, 2025, doi: 10.1049/PBPO263E_ch16.

[19] M. Alfayan and H. Hagras, "Towards an explainable artificial intelligence approach for smart grid systems," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 40, Apr. 2025, doi: 10.1007/s44163-025-00261-5.

[20] A. N. Achadiyah, A. N. Afandi, and S. Patmanthara, "From THD to causality: epistemology of artificial intelligence-based harmonic analysis in hybrid microgrids," *Energy: Jurnal Ilmiah Ilmu-Ilmu Teknik*, pp. 362–370, Dec. 2025, doi: 10.51747/energy.si2025.252.

[21] C. Rudin *et al.*, "Machine learning for the New York City power grid," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 328–345, 2012, doi: 10.1109/TPAMI.2011.108.

[22] I. Kapuza, E. Ginzburg-Ganz, R. Machlev, and Y. Levron, "Improving robustness of transformers for power quality disturbance classification via optimized relevance maps," *Engineering Applications of Artificial Intelligence*, vol. 161, 2025, doi: 10.1016/j.engappai.2025.112138.

[23] R. Machlev, M. Perl, A. Caciularu, J. Belikov, K. Y. Levy, and Y. Levron, "Explaining the decisions of power quality disturbance classifiers using latent space features," *International Journal of Electrical Power and Energy Systems*, vol. 148, 2023, doi: 10.1016/j.ijepes.2023.108949.

[24] L. Baur, K. Ditschuneit, M. Schambach, C. Kaymakci, T. Wollmann, and A. Sauer, "Explainability and interpretability in electric load forecasting using machine learning techniques – a review," *Energy and AI*, vol. 16, 2024, doi: 10.1016/j.egyai.2024.100358.

[25] T. Giri, B. B. Thapa, B. Paneru, and B. Paneru, "An XAI-driven CNN-Transformer model for transmission line fault detection and classification," *Energy Reports*, vol. 15, 2026, doi: 10.1016/j.egyr.2025.108929.

[26] U. Sipai, R. Jadeja, N. Kothari, T. Trivedi, R. Mahadeva, and S. P. Patole, "Performance evaluation of discrete wavelet transform and machine learning based techniques for classifying power quality disturbances," *IEEE Access*, vol. 12, pp. 95472–95486, 2024, doi: 10.1109/ACCESS.2024.3426039.

[27] J. T. Dandamudi, R. Kandula, R. D. A. Raj, R. M. R. Yanamala, and K. K. Prakasha, "Explainable AI for wind energy systems: state-of-the-art techniques, challenges, and future directions," *Energy Conversion and Management: X*, vol. 28, 2025, doi: 10.1016/j.ecmx.2025.101277.

[28] B. Shan, A. Borghetti, W. Zheng, and Q. Guo, "Explainable AI-based short-term voltage stability mechanism analysis: explainability measure and stability-oriented preventive control," *CSEE Journal of Power and Energy Systems*, vol. 11, no. 6, pp. 2673–2683, 2025, doi: 10.17775/CSEEJPES.2025.02850.

[29] A. B. A. Ferreira, J. B. Leite, and D. H. P. Salvadeo, "Power substation load forecasting using interpretable transformer-based temporal fusion neural networks," *Electric Power Systems Research*, vol. 238, 2025, doi: 10.1016/j.epsr.2024.111169.

[30] S. Akkaya and S. Dümen, "A novel vision transformer-based power quality disturbance classification method," *Ain Shams Engineering Journal*, vol. 16, no. 11, 2025, doi: 10.1016/j.asej.2025.103718.

[31] T. Wang, J. Zhuo, Y. Hou, Z. Lu, and Y. Li, "Power quality disturbance classification via a time-frequency feature-fused transformer model with cross-attention mechanism," *Electric Power Systems Research*, vol. 251, 2026, doi: 10.1016/j.epsr.2025.112330.

[32] Z. Gao, X. Liao, Z. Xiong, and X. Liu, "Power quality disturbance classification model combining two-layer one-dimensional convolutional neural network and transformer," *2025 6th International Symposium on New Energy and Electrical Technology, ISNEET 2025*, pp. 392–396, 2025, doi: 10.1109/ISNEET68549.2025.11407180.

[33] G. S. K, R. Satheesh, S. Rajan, and H. H. Alhelou, "Comparative analysis of vision transformer models in power quality disturbance classification," *IEEE Transactions on Industry Applications*, pp. 1–12, 2026, doi: 10.1109/TIA.2026.3675198.

[34] A. Sharma, H. Devarajan, and T. Bhavani Shanker, "Improved saliency maps for interpretability of acoustic signal classification in condition monitoring of power transformers," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, 2025, doi: 10.1109/TIM.2025.3614908.

[35] K. G. Sreshtamol, R. Satheesh, S. Rajan, and H. H. Alhelou, "Classification of power quality disturbance with synchrosqueezing and vision transformer," *2025 IEEE Kansas Power and Energy Conference, KPEC 2025*, 2025, doi: 10.1109/KPEC65465.2025.11045008.
[36] H. Xu, A. Boyaci, J. Lian, and A. J. Wilson, "Explainable AI for multivariate time series pattern exploration: latent space visual analytics with temporal fusion transformer and variational autoencoders in power grid event diagnosis," *IEEE Access*, vol. 14, pp. 11744–11767, 2026, doi: 10.1109/ACCESS.2025.3602635.
[37] Y. S. U. Vishwanath, S. Esakkirajan, B. Keerthiveena, and R. B. Pachori, "A generalized classification framework for power quality disturbances based on synchrosqueezed wavelet transform and convolutional neural networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, 2023, doi: 10.1109/TIM.2023.3308235.
[38] R. Machlev *et al.*, "Explainable artificial intelligence (XAI) techniques for energy and power systems: review, challenges and opportunities," *Energy and AI*, vol. 9, 2022, doi: 10.1016/j.egyai.2022.100169.
[39] A. Cifci, "Interpretable prediction of a decentralized smart grid based on machine learning and explainable artificial intelligence," *IEEE Access*, vol. 13, pp. 36285–36305, 2025, doi: 10.1109/ACCESS.2025.3543759.
[40] S. Chung and Y. Zhang, "Intelligence applications in electric distribution systems: post-pandemic progress and prospect," *Applied Sciences (Switzerland)*, vol. 13, no. 12, 2023, doi: 10.3390/app13126937.
[41] X. Zhou *et al.*, "A unified transformer-based harmonic detection network for distorted power systems," *Energies*, vol. 19, no. 3, 2026, doi: 10.3390/en19030650.
[42] S. Pütz, B. Schäfer, D. Witthaut, and J. Kruse, "Revealing interactions between HVDC cross-area flows and frequency stability with explainable AI," *Energy Informatics*, vol. 5, 2022, doi: 10.1186/s42162-022-00241-4.
[43] J. Kruse, B. Schäfer, and D. Witthaut, "Revealing drivers and risks for power grid frequency stability with explainable AI," *Patterns*, vol. 2, no. 11, 2021, doi: 10.1016/j.patter.2021.100365.
[44] M. I. H. C. Abdullah, A. Othman, R. Abdullah, and M. F. Abdollah, "Optimization on the nanoparticles stability in liquid phased condition by using Taguchi analysis," *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences*, vol. 61, no. 1, pp. 129–139, 2019.
[45] B. Li *et al.*, "Data-driven active corrective control in power systems: an interpretable deep reinforcement learning approach," *Frontiers in Energy Research*, vol. 12, 2024, doi: 10.3389/fenrg.2024.1389196.
[46] R. A. De Oliveira, V. Ravindran, S. K. Ronnberg, and M. H. J. Bollen, "Deep learning method with manual post-processing for identification of spectral patterns of waveform distortion in PV installations," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5444–5456, 2021, doi: 10.1109/TSG.2021.3107908.
[47] J. E. Caicedo, D. Agudelo-Martínez, E. Rivas-Trujillo, and J. Meyer, "A systematic review of real-time detection and classification of power quality disturbances," *Protection and Control of Modern Power Systems*, vol. 8, no. 1, 2023, doi: 10.1186/s41601-023-00277-y.
[48] E. Balouji, J. I. Al-Ashhab, S. M. Khatib, and L. Tulldahl, "On deep machine learning based harmonic localization using high resolution merging units," *8th IEEE International Conference on Energy Internet, ICEI 2024*, pp. 586–592, 2024, doi: 10.1109/ICEI63732.2024.10917246.
[49] M. Khalid, "Smart grids and renewable energy systems: perspectives and grid integration challenges," *Energy Strategy Reviews*, vol. 51, 2024, doi: 10.1016/j.esr.2024.101299.
[50] Z. Zhang, S. Rath, J. Xu, and T. Xiao, "Federated learning for smart grid: a survey on applications and potential vulnerabilities," *ACM Transactions on Cyber-Physical Systems*, vol. 10, no. 1, 2026, doi: 10.1145/3760788.
[51] L. Sun, H. Wang, L. Qi, J. Yan, and M. Jiang, "Composite harmonic source detection with multi-label approach using advanced fusion method," *Electronics (Switzerland)*, vol. 13, no. 7, 2024, doi: 10.3390/electronics13071275.
[52] Y. Mohammadi, S. M. Miraftabzadeh, M. H. J. Bollen, and M. Longo, "Voltage-sag source detection: developing supervised methods and proposing a new unsupervised learning," *Sustainable Energy, Grids and Networks*, vol. 32, p. 100855, Dec. 2022, doi: 10.1016/j.segan.2022.100855.
[53] G. Porawagamage, K. Dharmapala, J. S. Chaves, D. Villegas, and A. Rajapakse, "A review of machine learning applications in power system protection and emergency control: opportunities, challenges, and future directions," *Frontiers in Smart Grids*, vol. 3, 2024, doi: 10.3389/frsgr.2024.1371153.
[54] J. Qin, D. Yang, B. Zhou, and Y. Sun, "Semi-supervised variational bi-directional sampling on multi-class imbalanced electric power data for fault diagnosis," *International Journal of Electrical Power & Energy Systems*, vol. 155, p. 109512, Jan. 2024, doi: 10.1016/j.ijepes.2023.109512.
[55] Y. Jiang, Y. Li, and Y. Chen, "Interpretable short-term load forecasting via multi-scale temporal decomposition," *Electric Power Systems Research*, vol. 235, p. 110781, Oct. 2024, doi: 10.1016/j.epsr.2024.110781.
[56] Y. Li, Z. Wang, H. Wang, and S. Tao, "Harmonic source parameter identification method based on TCN-iTransformer," in *2025 International Conference on New Power System Technology (PowerCon)*, 2025, pp. 1–6, doi: 10.1109/PowerCon66300.2025.11294976.
[57] J. Caicedo Navarro and E. Rivas Trujillo, "Comparative assessment of machine learning techniques for fault-induced voltage sag classification," *Investigación e Innovación en Ingenierías*, vol. 13, no. 1, pp. 60–81, 2025, doi: 10.17081/invinno.13.1.7565.
[58] D. Albahdal, D. M. Ibrahim, T. Moulahi, and A. Omri, "Renewable energy management using explainable artificial intelligence," *Frontiers in Energy Research*, vol. 14, Apr. 2026, doi: 10.3389/fenrg.2026.1753200.
[59] A. Mino Calero, A. Rasheed, and A. M. Lekkas, "ChronoSHAP: Revealing temporal patterns from transformers and LTSF-linear models in time series forecasting," *Knowledge-Based Systems*, vol. 331, 2025, doi: 10.1016/j.knosys.2025.114802.
[60] A. Theissler, F. Spinnato, U. Schlegel, and R. Guidotti, "Explainable AI for time series classification: a review, taxonomy and research directions," *IEEE Access*, vol. 10, pp. 100700–100724, 2022, doi: 10.1109/ACCESS.2022.3207765.
[61] B. Arslan and S. Sagiroglu, "Toward real-time explainable AI in smart grids: an adaptive framework for operational transparency," in *2025 14th International Conference on Renewable Energy Research and Applications (ICRERA)*, Oct. 2025, pp. 1072–1077, doi: 10.1109/ICRERA66237.2025.11283679.
[62] S. Wang, S. Zhang, W. Zhang, R. Luo, and Z. Yu, "A cross-scenario robust verification framework for transmission systems via knowledge graph and causal intervention," *Results in Engineering*, vol. 30, 2026, doi: 10.1016/j.rineng.2026.109733.
[63] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, 2021, doi: 10.1002/widm.1424.
[64] M. S. Eydenberg *et al.*, "Physics-informed machine learning with optimization-based guarantees: applications to AC power flow," *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4245628.
[65] A. A. Latiff *et al.*, "Correlation of wear characteristics with hardness of recycled carbon fiber prepreg reinforced polypropylene composites," *Journal of Materials Research*, vol. 31, no. 13, pp. 1908–1913, 2016, doi: 10.1557/jmr.2016.39.
[66] B. Lusch, J. N. Kutz, and S. L. Brunton, "Deep learning for universal linear embeddings of nonlinear dynamics," *Nature Communications*, vol. 9, no. 1, 2018, doi: 10.1038/s41467-018-07210-0.
[67] A. Saadat, R. A. Hooshmand, A. Kiyoumarsi, and M. Tadayon, "Voltage sag source location in distribution networks with DGs using cosine similarity," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, 2022, doi: 10.1109/TIM.2022.3218516.

[68] R. Yao, H. Bai, S. Jiang, T. Liu, Y. Lei, and Y. Zheng, "A two-stage voltage sag source localization method in microgrids," *Energies*, vol. 19, no. 1, 2026, doi: 10.3390/en19010258.
[69] H. Mirshekali, R. Dashti, A. Keshavarz, and H. R. Shaker, "Machine learning-based fault location for smart distribution networks equipped with micro-PMU," *Sensors*, vol. 22, no. 3, 2022, doi: 10.3390/s22030945.
[70] W. Shuai, J. Sun, H. Bai, Y. Zheng, Y. Peng, and Z. Shuai, "Harmonic source localization for distribution networks with network parameter uncertainty," *2026 IEEE PES International Meeting, PES IM 2026*, 2026, doi: 10.1109/PESIM67009.2026.11438906.
[71] S. Mehrotra, C. C. Jorge, C. M. Jonker, and M. L. Tielman, "Integrity-based explanations for fostering appropriate trust in AI agents," *ACM Transactions on Interactive Intelligent Systems*, vol. 14, no. 1, 2024, doi: 10.1145/3610578.
[72] W. Yang *et al.*, "Survey on explainable AI: from approaches, limitations and applications aspects," *Human-Centric Intelligent Systems*, vol. 3, no. 3, pp. 161–188, 2023, doi: 10.1007/s44230-023-00038-y.
[73] C. Utama, C. Meske, J. Schneider, R. Schlatmann, and C. Ulbrich, "Explainable artificial intelligence for photovoltaic fault detection: A comparison of instruments," *Solar Energy*, vol. 249, pp. 139–151, 2023, doi: 10.1016/j.solener.2022.11.018.
[74] O. Lammert, B. Richter, C. Schütze, K. Thommes, and B. Wrede, "Humans in XAI: increased reliance in decision-making under uncertainty by using explanation strategies," *Frontiers in Behavioral Economics*, vol. 3, 2024, doi: 10.3389/frbhe.2024.1377075.
[75] A. Bertrand, R. Belloum, J. R. Eagan, and W. Maxwell, "How cognitive biases affect XAI-assisted decision-making," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, Jul. 2022, pp. 78–91, doi: 10.1145/3514094.3534164.
[76] G. Gürses-Tran, T. A. Körner, and A. Monti, "Introducing explainability in sequence-to-sequence learning for short-term load forecasting," *Electric Power Systems Research*, vol. 212, 2022, doi: 10.1016/j.epsr.2022.108366.
[77] M. Perl, Z. Sun, R. Machlev, J. Belikov, K. Y. Levy, and Y. Levron, "PMU placement for fault line location using neural additive models—A global XAI technique," *International Journal of Electrical Power and Energy Systems*, vol. 155, 2024, doi: 10.1016/j.ijepes.2023.109573.
[78] V. C. Sharma and K. Babber, "Explainable AI for decision-making in decentralized energy trading," in *Advancing the Clean Energy Frontier Through AI-Powered Green Innovation*, 2025, pp. 135–174, doi: 10.4018/979-8-3373-5540-5.ch005.
[79] V. Hassija *et al.*, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, Jan. 2024, doi: 10.1007/s12559-023-10179-8.
[80] X. Li, Y. Xi, and F. Zhou, "Power quality disturbances classification with imbalanced/insufficient samples based on WGAN-GP-SA and DCNN," *Digital Signal Processing: A Review Journal*, vol. 151, 2024, doi: 10.1016/j.dsp.2024.104518.
[81] H. F. Hamann *et al.*, "Foundation models for the electric power grid," *Joule*, vol. 8, no. 12, pp. 3245–3258, 2024, doi: 10.1016/j.joule.2024.11.002.
[82] H. U. Manzoor, A. Jafri, and A. Zoha, "Adaptive single-layer aggregation framework for energy-efficient and privacy-preserving load forecasting in heterogeneous federated smart grids," *Internet of Things (The Netherlands)*, vol. 28, 2024, doi: 10.1016/j.iot.2024.101376.
[83] A. El Rhatrif, B. Bouihi, and M. Mestari, "Challenges and limitations of artificial intelligence implementation in modern power grid," *Procedia Computer Science*, vol. 236, pp. 83–92, 2024, doi: 10.1016/j.procs.2024.05.008.
[84] M. J. Santofimia, X. Del Toro Garcia, F. J. Villanueva, D. Villa, S. Escolar, and J. C. Lopez, "A comprehensive common-sense-based architecture for understanding voltage-sag events in electrical grids," *Integrated Computer-Aided Engineering*, vol. 25, no. 4, pp. 397–416, 2018, doi: 10.3233/ICA-180576.
[85] D. Wang, T.-C. Chen, and K. Mahapatra, "Interpretability, explainability and trust metrics in anomaly detection method for power grid sensors," in *2025 IEEE PES Grid Edge Technologies Conference & Exposition (Grid Edge)*, Jan. 2025, pp. 1–5, doi: 10.1109/GridEdge61154.2025.10887434.
[86] U. Cali, F. O. Catak, and U. Halden, "Trustworthy cyber-physical power systems using AI: dueling algorithms for PMU anomaly detection and cybersecurity," *Artificial Intelligence Review*, vol. 57, no. 7, 2024, doi: 10.1007/s10462-024-10827-x.
[87] I. Jrhilifa, H. Ouadi, H. Rafia, A. Jilbab, N. Mounir, and Y. Lahrarti, "Explainable deep learning for voltage disturbance classification in microgrid AC bus using LSTM and SHAP," in *2025 13th International Conference on Systems and Control (ICSC)*, Oct. 2025, pp. 61–66, doi: 10.1109/ICSC67755.2025.11334735.
[88] R. Kandula, J. T. Dandamudi, R. D. A. Raj, R. M. R. Yanamala, and K. K. Prakasha, "Explainable AI for smart grid intelligence: a comprehensive review of techniques, applications and future directions," *Energy and AI*, vol. 25, 2026, doi: 10.1016/j.egyai.2026.100820.
[89] O. Amin, B. Brown, B. Stephen, and S. McArthur, "Case-study led investigation of explainable AI (XAI) to support deployment of prognostics in the industry," *PHM Society European Conference*, vol. 7, no. 1, pp. 9–20, 2022, doi: 10.36001/phme.2022.v7i1.3336.
[90] M. Kiasari and H. Aly, "Agentic artificial intelligence for smart grids: a comprehensive review of autonomous, safe, and explainable control frameworks," *Energies*, vol. 19, no. 3, 2026, doi: 10.3390/en19030617.
[91] S. Bin Akter, T. S. Pias, S. R. Deeba, J. Hossain, and H. A. Rahman, "Ensemble learning based transmission line fault classification using phasor measurement unit (PMU) data with explainable AI (XAI)," *PLoS ONE*, vol. 19, no. 2, 2024, doi: 10.1371/journal.pone.0295144.
[92] R. Bosso, C. Chang, M. Zarif, and Y. Tang, "Explainable graph neural networks for power grid fault detection," *IEEE Access*, vol. 13, pp. 129520–129533, 2025, doi: 10.1109/ACCESS.2025.3591604.
[93] P. Ellinas, I. Karampinis, I. Ventura Nadal, R. Nellikkath, J. Vorwerk, and S. Chatzivasileiadis, "Physics-informed machine learning for power system dynamics: A framework incorporating trustworthiness," *Sustainable Energy, Grids and Networks*, vol. 43, 2025, doi: 10.1016/j.segan.2025.101818.
[94] J. C. Portu, C. Delle Femine, K. S. Muro, M. Quartulli, and M. Restelli, "Limitations of physics-informed neural networks: a study on smart grid surrogation," *2025 IEEE Kiel PowerTech, PowerTech 2025*, 2025, doi: 10.1109/PowerTech59965.2025.11180429.
[95] F. Chen *et al.*, "Harmonic source localization in distribution networks based on data–mechanism fusion and spatiotemporal graph modeling," in *2025 4th Asia Power and Electrical Technology Conference (APET)*, 2025, pp. 886–890, doi: 10.1109/APET68390.2025.11428859.
[96] C. C. Dang, V. Q. Huynh, and K. P. Nguyen, "A proposed hybrid conditional multi-modal deep learning framework for harmonic source identification at PCC regarding photovoltaics," *2025 International Symposium on Electrical and Electronics Engineering (ISEE)*, 2025, pp. 196–201, doi: 10.1109/ISEE68370.2025.11223430.
[97] M. A. A. Sarker, B. Shanmugam, S. Azam, and S. Thennadil, "Enhancing smart grid load forecasting: an attention-based deep learning model integrated with federated learning and XAI for security and interpretability," *Intelligent Systems with Applications*, vol. 23, 2024, doi: 10.1016/j.iswa.2024.200422.
[98] Y. Zhang *et al.*, "FedNILM: applying federated learning to NILM applications at the edge," *IEEE Transactions on Green Communications and Networking*, vol. 7, no. 2, pp. 857–868, 2023, doi: 10.1109/TGCN.2022.3167392.
[99] H. Liu *et al.*, "Trustworthy AI: a computational perspective," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 1, 2022, doi: 10.1145/3546872.
[100] M. Pawlicki, A. Pawlicka, F. Uccello, S. Szelest, S. D'Antonio, and M. K. R. Choraś, "Evaluating the necessity of the multiple metrics for assessing explainable AI: a critical examination," *Neurocomputing*, vol. 602, 2024, doi: 10.1016/j.neucom.2024.128282.

[101] A. T. Nguyen and Y. Ahn, “Human-in-the-loop artificial intelligence in the energy sector: a systematic review of paradigm shifts toward Industry 5.0,” *Advances in Applied Energy*, vol. 22, p. 100267, Jun. 2026, doi: 10.1016/j.adapen.2026.100267.
[102] K. Zhang, Y. Wang, and P. Xu, “Towards enhanced decision-making: a human-machine collaborative approach with XAI for power system corrective control,” *Electric Power Systems Research*, vol. 254, p. 112619, May 2026, doi: 10.1016/j.epsr.2025.112619.
[103] P. Velmurugan, “AI-enabled smart electricity management system with predictive analytics,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 13, no. 7, pp. 1201–1207, Jul. 2025, doi: 10.22214/ijraset.2025.73163.
[104] F. Ucar, “A comprehensive analysis of smart grid stability prediction along with explainable artificial intelligence,” *Symmetry*, vol. 15, no. 2, p. 289, Jan. 2023, doi: 10.3390/sym15020289.
[105] A. Kumar, A. K. Dubey, I. Segovia Ramírez, A. Muñoz del Río, and F. P. García Márquez, “Artificial intelligence techniques for the photovoltaic system: a systematic review and analysis for evaluation and benchmarking,” *Archives of Computational Methods in Engineering*, vol. 31, no. 8, pp. 4429–4453, 2024, doi: 10.1007/s11831-024-10125-3.
[106] G. Santos, T. Pinto, C. Ramos, and J. M. Corchado, “Editorial: Explainability in knowledge-based systems and machine learning models for smart grids,” *Frontiers in Energy Research*, vol. 11, 2023, doi: 10.3389/fenrg.2023.1269397.
[107] W. Hua, B. Stephen, and D. C. H. Wallom, “Digital twin based reinforcement learning for extracting network structures and load patterns in planning and operation of distribution systems,” *Applied Energy*, vol. 342, 2023, doi: 10.1016/j.apenergy.2023.121128.
[108] K. Kobayashi and S. B. Alam, “Explainable, interpretable, and trustworthy AI for an intelligent digital twin: a case study on remaining useful life,” *Engineering Applications of Artificial Intelligence*, vol. 129, 2024, doi: 10.1016/j.engappai.2023.107620.

APPENDIX

Table 1. Comparison table: key research features, innovations, and future trajectories

Study	Core technique	Best feature/innovation	Current limitations	Future research directions
[56]	TCN-iTransformer	Dynamic global attention for harmonic flows in power-electronic grids [56].	High computational complexity of the iTransformer architecture for very long time-series sequences [56].	Optimization for real-time edge computing and low-latency parameter identification.
[17]	CNN-LSTM + SHAP	High-precision fault detection (98.7%) with feature-level transparency [17].	SHAP's post-hoc nature introduces significant computational overhead during real-time fault events [12], [17].	Integration with embedded real-time monitoring systems and hardware acceleration.
[8]	XAI Metrics	Introduction of a formal "explainability score" to measure model trust [8].	Metric validation is primarily theoretical and lacks testing on large-scale utility datasets [8].	Developing standardized, industry-wide benchmarks for quantifying explanation robustness.
[20]	Epistemic Fusion	Enforces KCL/KVL constraints to align AI logic with physical laws [20].	High mathematical complexity when defining constraints for large, non-linear hybrid microgrids [20].	Scalable physics-informed solvers for complex, multi-agent decentralized systems.
[22]	Relevance Maps	Uses Transformer attention to increase accuracy to 99.26% [22].	Attention maps can be noisy or misleading if the input signal contains high background distortion [22].	Improving the robustness of visual maps against signal noise and sensor inaccuracies.
[48]	Attention-LSTM	High-resolution sampling (28.8kHz) for precise feeder-level localization [48].	Extremely high data storage and transmission requirements for utility-wide deployment [48].	Exploring compressed sensing techniques for high-resolution harmonic monitoring.
[87]	LSTM + SHAP	Automated classification of 8 disturbance types with 0.99 F1-score [87].	Model generalizability to diverse microgrid topologies remains largely untested [87].	Cross-domain transfer learning for standardized deployment across different bus configurations.
[28]	LIME / SHAP	Identification of instability drivers for real-time preventive control [28].	Local explanations may fail to capture complex, global system-wide stability interactions [28].	Multi-level stability explanation frameworks that bridge local and global grid insights.
[68]	Two-Stage STGCN	Robust localization in microgrids with bidirectional energy flows [68].	Requires extensive historical training data for accurate graph edge weights [68].	Zero-shot or few-shot learning for faster deployment in newly installed microgrids.
[52]	RF / Unsupervised	Exceptional accuracy (99.84%) using hybrid supervised-unsupervised learning [52].	The underlying logic remains a "black-box," lacking explainability in the clustering phase [52].	Integrating SHAP or LIME specifically into the unsupervised clustering and detection stage.

BIOGRAPHIES OF AUTHORS

Mohd Hatta Jopri received his B.Eng. from Universiti Teknologi Malaysia (UTM), M.Sc. in Electrical Power Engineering from Rheinisch-Westfälische Technische Hochschule Aachen (RWTH Aachen University), Germany, and Ph.D. from Universiti Teknikal Malaysia Melaka (UTeM). Since 2005, he has served in academia and as a senior researcher at UTeM, with strong contributions in electrical power engineering, power quality, renewable energy, and industry-driven technical education. He is a Professional Engineer registered with the Board of Engineers Malaysia (BEM), a Professional Technologist registered with the Malaysia Board of Technologists (MBOT), a Certified Data Protection Officer (CDPO), and a member of the International Association of Engineers (IAENG). His work integrates engineering research, Technical and Vocational Education and Training (TVET), competency-based education, and industry certification to support human capital development and technology advancement. His research interests include power electronics and drives, power quality analysis, signal processing, machine learning, artificial intelligence (AI), explainable artificial intelligence (XAI), data science, and TVET transformation. He can be contacted at email: hatta@utem.edu.my.

Tole Sutikno is a full professor in the Department of Electrical Engineering at Universitas Ahmad Dahlan, Yogyakarta, Indonesia. He has served in this role since 2023, after his appointment as associate professor in 2008. He received his bachelor's degree from Universitas Diponegoro in 1999, master's degree from Universitas Gadjah Mada in 2004, and Ph.D. in Electrical Engineering from Universiti Teknologi Malaysia in 2016. His doctoral research focused on advanced digital power electronics and intelligent control systems. From 2016 to 2021, he served as Director of the Institute for Scientific Publishing and Publications at Universitas Ahmad Dahlan, where he led initiatives in research visibility, journal management, and international scholarly collaboration. He is the founding leader of the Embedded Systems and Power Electronics Research Group, which collaborates with national and international institutions in fault-tolerant embedded systems, FPGA-based control, and renewable energy integration. His research interests include digital design, industrial electronics, motor drives, robotics, intelligent systems, AI-based automation, embedded systems, power electronics, FPGA applications, and digital libraries. His work focuses on practical applications in industrial and healthcare contexts. He has been listed among the Top 2% of Scientists Worldwide by Stanford University and Elsevier BV from 2021 to the present, based on standardised citation indicators across scientific disciplines. He can be contacted at email: tole@te.uad.ac.id.

Yacine Djeghader was born on 5 July 1980 in Annaba, Algeria. He is currently a professor of Electrical Engineering at the University of Souk Ahras, Algeria. He received his State Engineering degree in Electrical Engineering in 2003, followed by a Master's degree in 2006 and a Ph.D. in 2011 from Annaba University. He was promoted to professor in 2025. His research focuses on power quality improvement, power electronics, active power filters (APF and SAPF), harmonic mitigation, FACTS devices, renewable energy integration, smart grids, and intelligent control and optimization methods for modern power systems. He has authored and co-authored numerous papers in international journals and conferences. He is actively involved in postgraduate supervision and research projects, contributing to the development of efficient, reliable, and sustainable electrical energy systems. He collaborates with national and international researchers and participates in scientific committees and peer review activities in the field of electrical engineering. He can be contacted at email: djeghaderyacine@yahoo.fr.

Mohd Riduan Mohd Shariff was born in Sungai Besar, Selangor, Malaysia, in 1983. He graduated with a Bachelor of Engineering (Electrical and Electronic Engineering) from Universiti Kebangsaan Malaysia (UKM) in 2006 and is currently pursuing an Industry Ph.D. in Electrical Engineering at Universiti Tun Hussein Onn Malaysia (UTHM). With nearly two decades of experience in the refinery and oil & gas industry, he is a Professional Engineer (Ir.) with a Practising Certificate registered with the Board of Engineers Malaysia (BEM) and a certified 132 kV Competent Electrical Engineer recognized by the Energy Commission of Malaysia (ST). He currently serves at Malaysian Refining Company Sdn. Bhd. (MRCSB), where he manages specialized engineering services for electrical power generation, transmission, and distribution systems. His professional interests include electrical safety and governance, power system protection, substation engineering, earthing and lightning protection, electromagnetic interference, infrared thermography, and reliability engineering. He actively contributes to technical publications, industry forums, and competency development initiatives. He can be contacted at email: riduan_shariff@petronas.com or simposius@gmail.com.

Wan Azlan Zainal Abidin is a full professor at the Faculty of Engineering, Universiti Malaysia Sarawak (UNIMAS), with more than 25 years of academic, research, and leadership experience. He began his career at UNIMAS in 1998 as a Tutor and later advanced through the positions of lecturer, senior lecturer, associate professor, and professor. He previously served as Head of the Department of Electronic Engineering, Deputy Director, and Director of the Centre for Graduate Studies at UNIMAS. He obtained his Bachelor of Engineering from the University of Bristol, United Kingdom, Master of Engineering Studies from the University of Technology Sydney, Australia, and Doctor of Engineering from Kyushu University, Japan. His expertise includes renewable energy, solar photovoltaic systems, micro-hydropower, telecommunications, and engineering education. He is actively involved in competency training, postgraduate supervision, and research projects related to sustainable energy and rural electrification. He can be contacted at email: wzaazlan@unimas.my.