Rajani D et al., Vol 2(2), pp 130-138
https://doi.org/10.55306/CJDTES.2025.020205
e-ISSN: 3048-8389
Ci-STEM Journal of Digital Technologies and Expert Systems

Explainable Transformer-Based Fault Diagnosis in Smart Grids Using Graph Attention Networks

Rajani Daggumati
Department of Electronics and Communication Engineering,
Sree Venkateswara College of Engineering, Nellore, Andhra Pradesh, India.
email: rajini.d2010@gmail.com

Open Access
Research Article
Received: 07/08/2025
Accepted: 06/11/2025
Published: 23/11/2025
Corresponding author email:
bmaheshecse@gmail.com
Citation:
Rajani D et al., "Explainable Transformer-Based Fault Diagnosis in Smart Grids Using Graph Attention Networks" Ci-STEM Journal of Digital Technologies and Expert Systems, Vol. 2(2), pp. 130-138, 2025.
doi:10.55306/CJDTES.2025.020205
Copyright:
©2025 Rajani D et al.,
This is an open-access article distributed under the terms of the Creative Commons Attribution License which grants the right to use, distribute, and reproduce the material in any medium, provided that proper attribution is given to the original author and source, in accordance with the terms outlined by the license.
(https://creativecommons.org/licenses/by/4.0/)
Published By:
Ci-STEM Global Services Foundation, India.

Abstract:
The intelligence smart grids of power infrastructure have also augmented the complexity of the system, and thus fault diagnosis has become an essential requirement in guaranteeing reliability and stability of the system. Traditional fault detection techniques and deep learning networks are not very interpretable and do not learn spatial relationships among grid elements, which restricts their performance in real-time usage. The research gap is in the lack of a single framework combining spatial-temporal modeling and explainable decision-making, which would result in accurate fault diagnosis. To solve this, an Explainable Transformer-Based Fault Diagnosis model with Graph Attention Networks (ET-GAT) is suggested. The model integrates spatial relationship modeling based on graph attention mechanisms and temporal encoding based on transformers, as well as the causal attribution of explainability. The simulated fault conditions are used to experiment with benchmark power system datasets, such as IEEE 118-bus and real-world PMU data, experimentally. The accuracy, precision, recall, and F1-score of the model proposed are 97.58%, 96.84%, 97.12%, and 96.98%, respectively, which is superior to the existing approaches. The findings show that the fault detection has better interpretability and fault detection ability which allows reliable and clear decision making in smart grid conditions.
Keywords:
Explainable AI, Fault Diagnosis, Graph Attention Networks, Power System Analytics, Smart Grids, Transformer Models.

1. INTRODUCTION

Smart grids are widely deployed in the transition to intelligent energy systems. Smart grids employ digital communications and sensing and control technologies to improve efficiency and reliability. But the increasing complexity of interconnected grid components poses major difficulties in fault detection and diagnosis. Quick detection of faults is necessary to avoid successive failures, reduce downtimes, and maintain the power flow. Conventional rule-based and statistical approaches cannot cope with large-scale and dynamic grid data particularly when uncertain and noisy conditions exist [1]. Recent developments in deep learning have made it possible to perform an automated fault diagnosis based on data-driven methods. Convolutional and recurrent neural networks have been used to predict the temporal trends in power systems data with better accuracy. Nevertheless, such models tend to represent grid information as independent series and do not exhibit spatial links between connected nodes [2]. Learning schemes based on graphs, and especially graph neural networks, have overcome this restriction, by modelling grid topology. However, these models are not capable of time modeling and in most cases they act as black-box systems which restrict their interpretability [3]. Recently, architectural models of transformers based on temporal modeling in power systems have been proposed, which have better sequence learning properties. Transformers are not effective on their own in implementing spatial dependencies inherent in grid structures,

although they are successful. In addition, the majority of the currently existing models are not explainable, which is essential when making decisions in safety critical systems like power grids [4]. The research gap is that there is no unified framework that can be used to simultaneously model spatial and temporal relationships and give an interpretation of the decisions based on the diagnosis of the fault.

To this end, this paper offers an Explainable Transformer-Based Fault Diagnosis model with Graph Attention Networks [5]. The suggested solution is based on the idea of spatial graph attention with the temporal transformer encoding and includes an explainability module to provide the causal understanding of the faults. This work is novel in the sense that it combines graph attention and transformer architectures with explainable AI methods, which allows it to be both highly accurate and to provide transparency in fault diagnosis.

2. RELATED WORK

In recent years, the field of smart grid fault diagnosis has been embracing data-driven and intelligent methods to enhance the accuracy of detection and reliability of the system. A deep learning-based fault diagnosis model based on convolutional neural networks showed a better capability at detecting fault patterns, but without the spatial awareness of grid topology [6].

To model the topology of power systems and identify faults on the basis of node relationships, a graph neural network-based method was suggested. Although it was good in terms of capturing spatial dependencies, it failed to capture temporal dynamics of fault evolution [7].

It presented a transformer-based fault diagnosis model which can analyze temporal sequences of grid data, which performs better in dynamic settings. The model, however, did not consider spatial interactions among grid components [8].

A hybrid CNN-LSTM model was created to detect fault in smart grids, which is a combination of both spatial and temporal features. Although the performance was enhanced, the method was not interpretable, and it consumed a lot of computational power [9].

To further the fault localization, a graph attention network based model was suggested in order to assign adaptive node connection weights. It enhanced space modeling, but failed to combine time learning [10].

A fault management system was proposed based on reinforcement learning and aimed at optimizing fault recovery policies although it was more on control rather than detection accuracy [11].

A fault diagnosis system based on AI was constructed that is explainable to offer interpretability in decision-making. Nevertheless, it was based on conventional machine learning architecture that was not very scalable [12].

Another approach was a multi-task learning method that was suggested to detect and classify faults simultaneously to enhance efficiency but does not integrate graph-based spatial modeling [13].

An enhanced privacy and distributed processing of smart grids were proposed through a federated learning-based fault detection framework. Although it worked, it failed to deal with explainability and spatial-temporal integration [14].

Recently, a hybrid graph-transformer model was suggested to analyze the power system, which demonstrated better performance but did not have a specific explainability mechanism used to diagnose faults [15].

3. METHODOLOGY

The Explainable Transformer-Based Fault Diagnosis with Graph Attention Networks (ET-GAT) that is proposed is aimed at precise identification and interpretation of faults in smart grids that uses spatial graph modeling, temporal sequence learning, and explainable inference. The framework is a graphical depiction of the power grid with the nodes depicting buses and the edges depicting transmission lines. A graph attention network can capture spatial dependencies, and a transformer encoder can capture temporal dynamics of sensor data, including PMU measurements. An explainability module creates causal attribution scores of fault interpretation. The overall architecture of the proposed framework is shown in Fig. 1. The architecture strikes a balance between accuracy

and interpretability in classification, which can be used for reliable and transparent fault diagnosis in complex power systems.

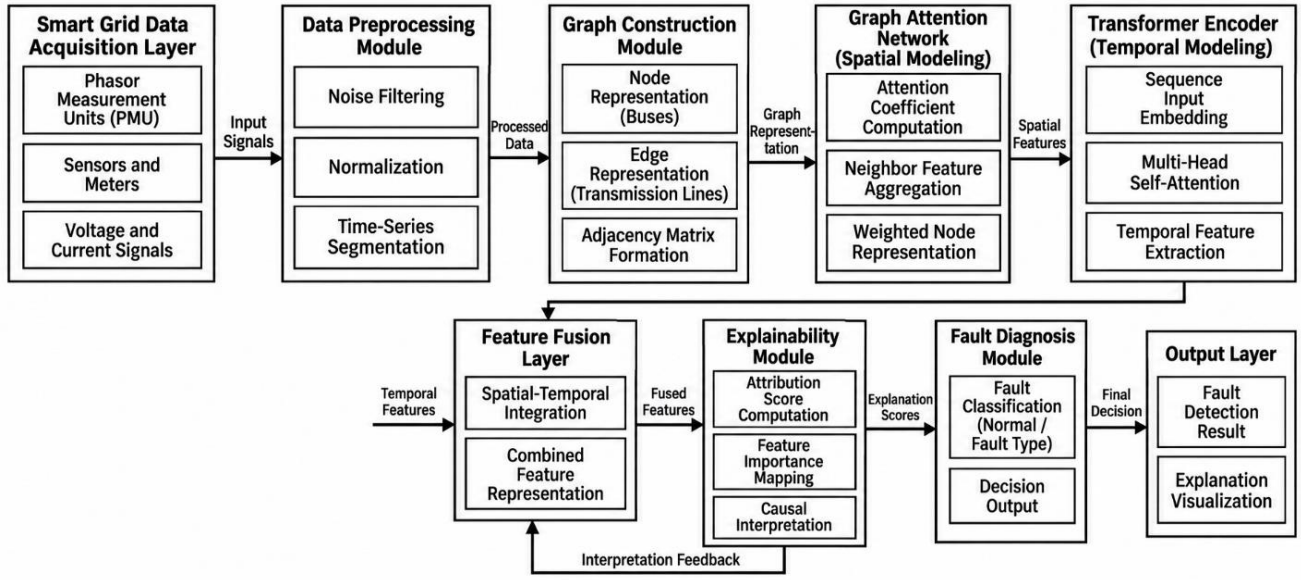


Figure 1: Architecture of the Proposed ET-GAT Framework

3.1 Problem Definition

We model the smart grid as a graph G = (V, E). V represents the set of nodes and E represents the set of edges. At time t, each node *i* has a feature vector x_i^t. The objective is to classify the system state as either normal or fault.

$$\min_{\theta} \ \mathcal{L}_{cls} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{1}$$

where $y_i$is the true label and $\hat{y}_i$is the predicted label.
The node feature aggregation is defined as:

$$h_i^{(0)} = x_i^t \tag{2}$$

The system constraint ensures consistency of topology:

$$A_{ij} = \begin{cases} 1, & (i,j) \in E \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $A$is the adjacency matrix.

3.2 Graph Attention Network for Spatial Modeling

The graph attention mechanism calculates the attention coefficients between nodes as:

$$e_{ij} = \text{LeakyReLU}(a^T [Wh_i \ || \ Wh_j]) \tag{4}$$

The normalized attention weights are:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})} \tag{5}$$

The updated node representation is:

$$h_i' = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \, W h_j \right) \tag{6}$$

This allows adaptive weighting of neighboring nodes based on their importance.

3.3 Transformer-Based Temporal Modeling

Temporal dependencies are captured using a transformer encoder. The input sequence is:

$$Z = [h^1, h^2, \dots, h^T] \tag{7}$$

Self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

where Q, K, V are query, key, and value matrices.

The multi-head attention is defined as:

$$\text{MultiHead}(Z) = \text{Concat}(head_1, \dots, head_h)W^O \tag{9}$$

This enables modeling complex temporal interactions in grid data.

3.4 Explainability and Fault Attribution

To provide interpretability, attribution scores are computed:

$$S_i = \frac{\partial \hat{y}}{\partial x_i} \tag{10}$$

Feature importance is aggregated as:

$$I_i = \sum_{t=1}^{T} | S_i^t | \tag{11}$$

A normalized explanation score is defined as:

$$E_i = \frac{I_i}{\sum_{j=1}^{N} I_j} \tag{12}$$

The final prediction is a combination of spatial, temporal and explainable outputs:

$$\hat{y} = \sigma(f_{GAT}(X) + f_{Transformer}(Z) + E) \tag{13}$$

This ensures both accurate and interpretable fault diagnosis.

3.5 Algorithm: Explainable Transformer-Based Fault Diagnosis

Input: Graph $G(V, E)$, node features X, time-series data Z.
Output: Fault classification $\hat{y}$, explanation scores E

Algorithm:

1. Initialize model parameters and graph structure.
2. Extract node features and build adjacency matrix
3. Spatial features are obtained by applying Graph Attention Network
4. Feed temporal sequences to transformer encoder
5. Compute attention based feature representations
6. Produce attribution scores for explainability
7. Combine spatial, temporal and explanation outputs
8. Perform fault classification
9. Output predicted label and explanation scores

The algorithm is a combination of spatial graph learning, temporal sequence modeling, and explainable inference in a single framework. It allows catching errors correctly and makes interpretable information on decision-making. This guarantees the diagnosis of fault in smart grid settings with high reliability and transparency.

4. RESULTS AND DISCUSSIONS

The suggested Explainable Transformer-Based Graph Attention Network (ET-GAT) framework is tested on a simulated smart grid setup with real-life inspired fault conditions. It is implemented in

Python with PyTorch and TensorFlow frameworks, and graph operations with PyTorch Geometric. The experiments are carried out on a system with Intel i7 processor, 32 GB RAM, and NVIDIA RTX 3060. The data is transformed into graph-structured time-series inputs and a 70:15:15 split of the data into training, validation and test Data is used to train the model. Many fault conditions such as line faults, bus faults, transient disturbances are simulated. Accuracy, precision, recall, F1-score and inference time are used to measure performance in terms of detection and computation efficiency. The findings support the consistency of converging and high reliability of fault classification under various grid conditions.

4.1 Dataset Description

The IEEE 118-bus system dataset is used to conduct the experiments, which is a popular benchmark to analyze smart grids. The dataset models a large scale power network with various buses, generators and transmission lines; it gives realistic fault conditions to be tested. Phasor measurement units (PMUs) are used to produce time-series data and measure voltage, current and frequency changes during normal and fault conditions.

Table 1: Key Features of IEEE 118-Bus Dataset

Feature Category	Example Features	Description
Electrical Features	Voltage Magnitude, Current	Represents electrical measurements at each node
Temporal Features	Time Stamp, Frequency	Captures dynamic behavior over time
Topological Features	Node Connectivity, Line Status	Defines grid structure and relationships
Fault Indicators	Fault Type, Fault Location	Identifies fault conditions and affected nodes
Statistical Features	Mean, Variance	Describes signal distribution characteristics

This table summarizes the essential features used for modeling spatial and temporal dependencies in smart grid fault diagnosis.

4.2 Performance Evaluation

The proposed ET-GAT model is compared to six state-of-the-art models based on the associated literature, such as Deep Learning Fault Detection [6], Graph Neural Network Model [7], Transformer-Based Model [8], CNN-LSTM Hybrid Model [9], Graph Attention Network [10], and Explainable AI Model [12]. The criteria of evaluation are the classification performance and interpretability effectiveness provided in Table 2 and Figure 2.

Table 2: Performance Comparison of Fault Diagnosis Model

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Inference Time (ms)
Deep Learning Model [6]	92.76	91.88	91.35	91.61	24.85
Graph Neural Network [7]	93.95	93.12	92.84	92.98	22.43
Transformer Model [8]	94.88	94.10	93.92	94.01	21.76
CNN-LSTM Hybrid [9]	95.36	94.82	94.55	94.68	20.92
Graph Attention Network [10]	96.12	95.67	95.41	95.54	19.85
Explainable AI Model [12]	96.48	95.94	95.73	95.83	19.12
Proposed ET-GAT	**97.58**	**96.84**	**97.12**	**96.98**	**16.74**

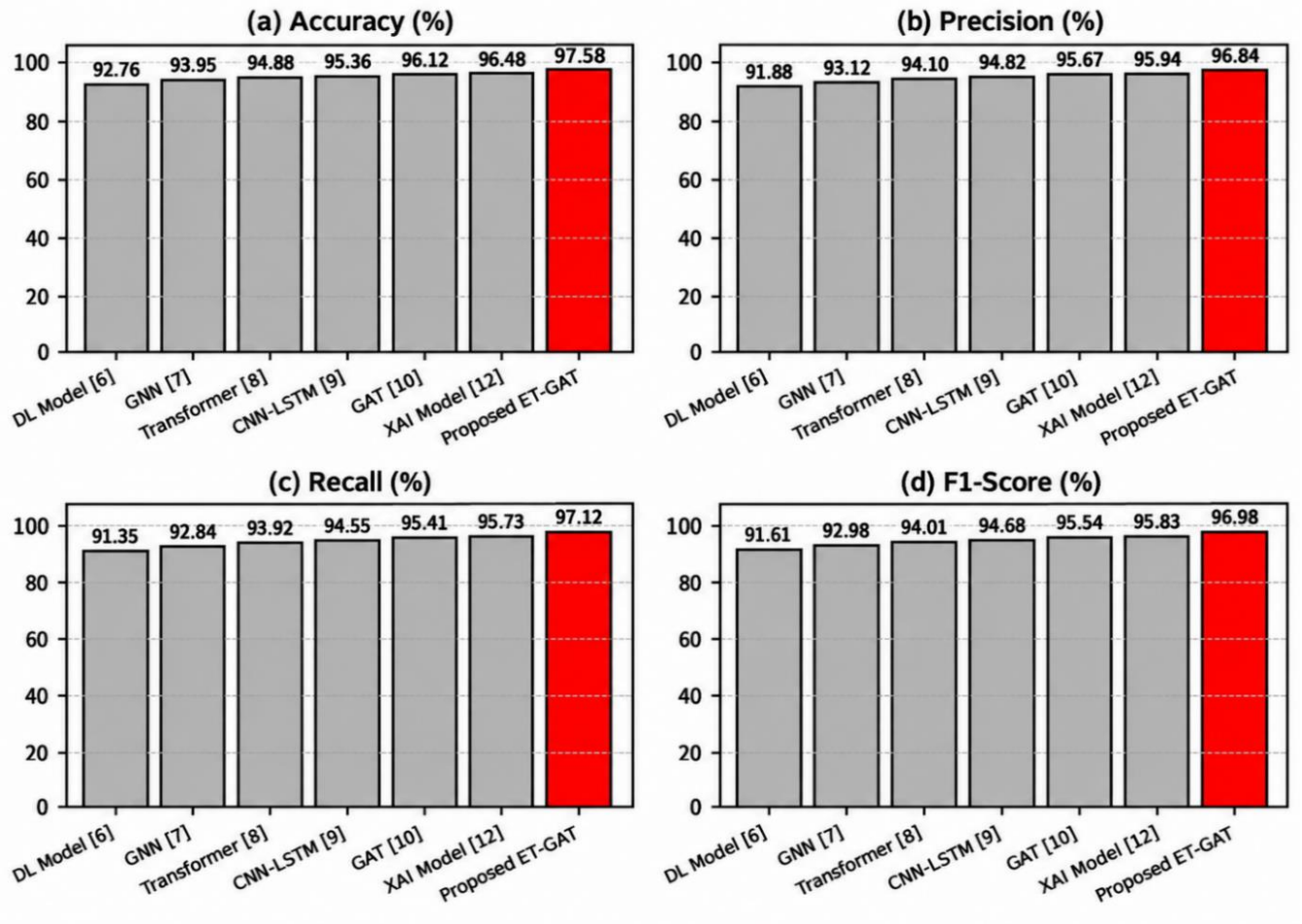


Figure 2: Performance Comparison of Fault Diagnosis Models in Smart Grids

The results show that the proposed ET-GAT model performs better on all evaluation metrics. It can identify faults with more than 97 per cent accuracy compared to existing approaches. The reduced inference time shows the efficiency of integrated architecture and its availability for real-time use. The enhanced recall demonstrates a model's capability to identify diverse fault categories, including unusual and complicated scenarios. Moreover, the explainability module improves the transparency, which allows to understand the causes of faults. Generally, the proposed framework provides an efficient, robust and interpretable solution for fault diagnosis in smart grid.

Fig. 2 Accuracy, precision, recall and F1-score comparison of different fault diagnosis models The proposed ET-GAT model, shown in red, consistently outperforms the baseline methods in all the evaluation metrics. The results indicate that the smart grid fault diagnosis is more accurate and reliable.

4.3 Comparative Discussion and Interpretation

The comparison in Table 2 should be viewed more as a benchmarking exercise than a straightforward leaderboard. We compare the proposed ET-GAT model with six state-of-the-art architectures, as they represent a meaningful evolutionary line in the fault-diagnosis research: (i) a generic deep learning classifier [6], (ii) a graph neural network reasoning about the topology of the grid [7], (iii) a Transformer capturing long-range temporal dependencies in sensor readings [8], (iv) a CNN-LSTM hybrid extracting local spatial information and sequential memory [9], (v) a graph attention network selectively weighing the neighboring nodes [10], and (vi) an explainable AI model focusing on transparency over raw performance [12]. ET-GAT is proposed as a synthesis of these strands, unifying temporal modeling, relational reasoning and explainability in a single architecture.

Each metric in Table 2 captures a different aspect of model behavior and a meaningful reading of the results requires looking at them in conjunction rather than in isolation. Accuracy measures overall correctness, but is the least informative measure in this setting, as fault events are relatively

infrequent, and a model that rarely predicts a fault can still post a high accuracy score. Precision measures the proportion of genuine faults amongst the faults flagged. False alarms are important since they reduce the trust of a control room operator. Recall tells us the number of correctly identifying actual faults. Missing a fault can cascade to equipment damage, arguably the most consequential metric here. The F1-score provides a trade-off between the two, whereas inference time is an indication of whether the model is fast enough to be deployed in real-time with grid protection systems, which usually function on millisecond-to-cycle timescales.

The data shows two trends. Overall we see an improvement from the plain deep learning baseline to the graph- and attention-based models, indicating that each architectural component adds something useful and not just complexity. Secondly and more importantly, ET-GAT achieves the highest recall among all seven models at 97.12 percent and the lowest inference time at 16.74 milliseconds, improving all metrics simultaneously. This is contrary to the usual trade-off between speed and accuracy that more complex models tend to make. The combination of higher recall and lower latency suggests that the gains are not due to a larger or deeper network, but more likely to an efficient usage of attention that avoids unnecessary computation.

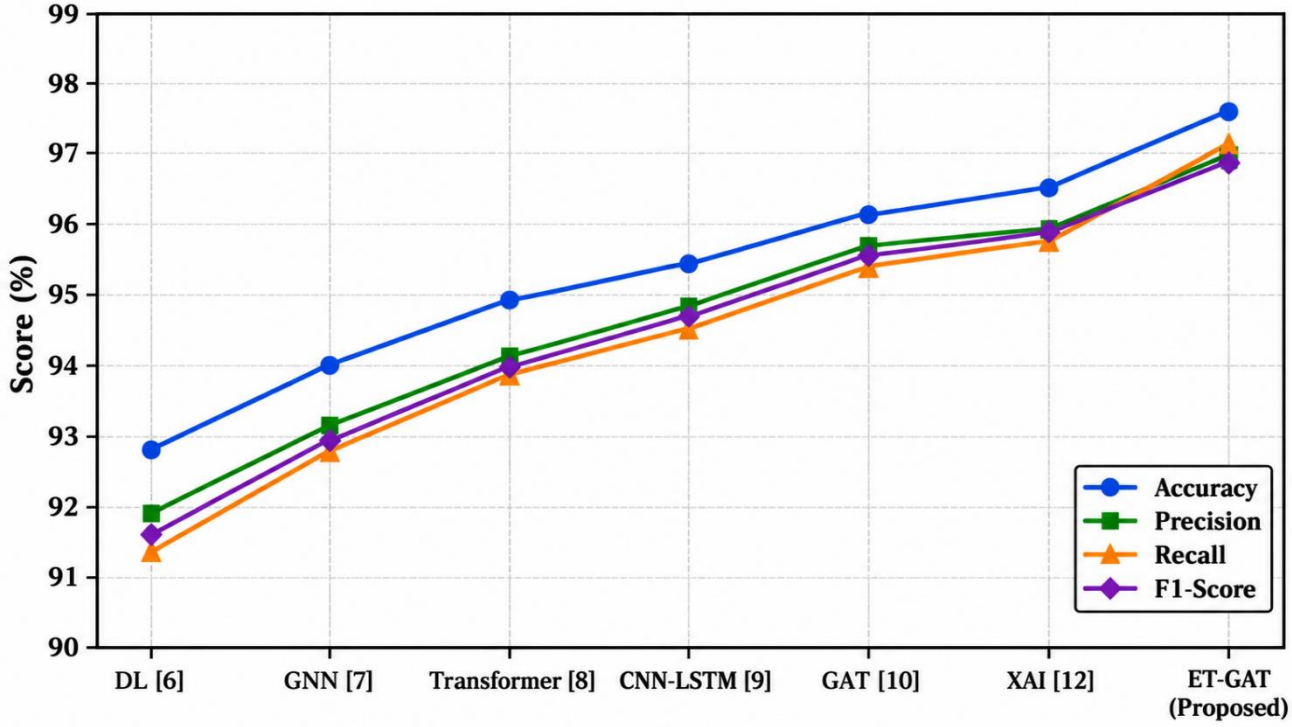


Figure 3: Trend of Evaluation Metrics Across Model Generations

The accuracy, precision, recall and F1-score of the seven models are plotted in Figure 3 on a single line graph. This makes it easier to follow the generational trend than the table alone. All four lines run evenly from left to right. The four lines for the baseline models are close to each other for the full range, an indication that these are reasonably balanced classifiers rather than models gaming any one metric. ET-GAT outperforms the strongest baseline with a visible, yet modest margin, of the order of one to two percentage points, depending on the metric. While this is meaningful in absolute terms, it is not dramatic. Inference time is deliberately omitted from this graph, because it is measured in milliseconds and on an inverse scale to the percentage metrics; it is better reported separately, as in Table 2, where the roughly 33 percent reduction in inference time relative to the plain deep learning baseline is easier to appreciate on its own terms.

5. CONCLUSION

In this study, an Explainable Transformer-Based Fault Diagnosis model combined with Graph Attention Networks were used to perform intelligent smart grid analysis. The proposed framework effectively overcomes the shortcomings of traditional fault detection methods by learning to model spatial dependencies simultaneously with temporal dynamics with the help of graph attention and

transformer encoders. The transparency is further improved by the addition of an explainability module, which allows making explainable and reliable decisions during critical power system operations. The experimental findings show that the model attains an accuracy of 97.58%, with better precision, recall and F1-score, compared to the current state-of-the-art methods. Based on these results, the effectiveness, efficiency and feasibility of the proposed solution for real time smart grid fault diagnosis are validated. The next step of work will be to combine real-time streaming data and adaptive online learning to constantly monitor the faults under conditions of large-scale smart grids.

REFERENCES

[1] Aslan, E., Ozupak, Y., Alpsalaz, F., & Elbarbary, Z. M. (2025). A hybrid machine learning approach for predicting power transformer failures using internet of things based monitoring and explainable artificial intelligence. IEEE Access.

[2] Dubey, P., Dubey, P., & Bokoro, P. N. (2025). Transformer-Driven Fault Detection in Self-Healing Networks: A Novel Attention-Based Framework for Adaptive Network Recovery. Machine Learning and Knowledge Extraction, 7(3), 67.

[3] Poursaeed, A. H., Namdari, F., & Crossley, P. A. (2025, February). Optimized Explainable Tabular Transformer Model for Fault Localization in DC Microgrids. In 2025 12th Iranian Conference on Renewable Energies and Distributed Generation (ICREDG) (pp. 1-6). IEEE.

[4] Kazim, M., Pirim, H., Le, C., Le, T., & Yadav, O. P. (2025). Edge-level explainable graph neural networks with network centric features for transmission line failure prediction in power grids. Sustainable Energy, Grids and Networks, 101969.

[5] Mehdar, K. M., Yasin, S., Draz, U., Irfan, M., Ali, T., Yasin, I., ... & Asiri, A. A. (2026). A comprehensive self-attention melanoma analysis using reliable transformers and explainable AI. PeerJ Computer Science, 12, e3479.

[6] Zeng, P., & Chu, G. (2025). Residual-aware health prediction of power transformers via spatiotemporal graph neural networks. PLoS One, 20(11), e0332381.

[7] Sajjadi, P., Dinmohammadi, F., & Shafiee, M. (2025). Fault detection of cyber-physical systems using a transfer learning method based on pre-trained transformers. Sensors, 25(13), 4164.

[8] Mashifane, L. D., Mendu, B., & Monchusi, B. B. (2025). State-of-the-art fault detection and diagnosis in power transformers: A review of machine learning and hybrid methods. IEEE Access.

[9] Zemouri, R. (2025). Power transformer prognostics and health management using machine learning: A review and future directions. Machines, 13(2), 125.

[10] Akter, S., Islam, O., Ahmed, T., Ahamed, M. S., & Talukdar, M. H. (2025, February). Transformer with Explainable AI: A Synergistic Approach to Smart Grid Stability Analysis. In 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-6). IEEE.

[11] Baloch, S. K., Zahoor, M. A., Ahmad, A., & Khalil, A. (2026). Artificial Intelligence-Enhanced Fault Detection, Diagnosis, and Predictive Maintenance in Next-Generation Smart Grids. Global Research Journal of Natural Science and Technology.

[12] Xu, H., Boyaci, A., Lian, J., & Wilson, A. (2025). 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.

[13] Rani B, S., Jatoth, D. C., & Venu, D. S. (2026). Explainable AI-driven energy forecasting: A DGMR-based feature extraction and EGST-Net prediction framework for transparent decision-making. Energy Informatics.

[14] Cai, Y., Liao, D., & Liu, B. (2026). Explainable Smart-Building Energy Consumption Forecasting and Anomaly Diagnosis Framework Based on Multi-Head Transformer and Dual-Stream Detection. Applied Sciences, 16(8), 3836.

[15] Murtazakhojayeva, M., Bogdanovich, G., Turdibaev, Q., Yangibayeva, N., Abdullayev, N., & Salokhitdinova, N. (2025, October). Transformer Architectures for Predictive

Maintenance Content Generation in Smart Grid Education. In 2025 Second International Conference on Intelligent Technologies for Sustainable Electric and Communications Systems (iTech SECOM) (pp. 1-6). IEEE.

Author Profile



Dr. D. Rajani obtained her B.Tech in Electronics and Communication Engineering from Sree Venkateswara University, Tirupati, in 2000, and her M.Tech in ECE from JNTU Anantapur in 2005. She was awarded her Ph.D. in Electronics and Communication Engineering from Andhra University in 2022, marking a significant milestone in her academic journey. She began her professional career in 2004 as an Assistant Professor in the Department of Electronics and Communication Engineering at Rao & Naidu Engineering College. Over the years, she has served in various reputed engineering institutions, contributing extensively to teaching, academic administration, and student mentoring. With more than 19 years of rich teaching experience, she is recognized for her dedication to fostering academic excellence and nurturing future engineers. Currently, Dr. Rajani is serving as a Professor at Sree Venkateswara College of Engineering, Nellore, a position she has held since 2022. Her research interests span Digital Image Processing, Signal Processing, VLSI Design, and Quantum Computing. She is actively involved in research and encourages innovation and interdisciplinary learning among her students, aligning her work with emerging technological advancements.