FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 2 | Pages 209-217 | e-ISSN: 2347-470X

Data Mining in Power System Fault Identification using Artificial Intelligence

Nhung Le Thi Hong[1], Trong Nghia Le[2*], Tan Phung Trieu[3], Hoang Le Thi Thanh[4], and Thanh Nguyen Tan[5]

[1]*Master of Engineering, Faculty of Electrical and Electronic Engineering, HCMC University of Technology and Education; nhunglth@hcmute.edu.vn*
[2*]*Doctor of Philosophy, Faculty of Electrical and Electronic Engineering, HCMC University of Technology and Education; trongnghia@hcmute.edu.vn*
[3]*Master of Engineering, Faculty of Electrical and Electronics Engineering, Cao Thang Technical College, Ho Chi Minh city, Vietnam*
[4]*Master of Engineering, Faculty of Electrical and Electronic Engineering, HCMC University of Technology and Education; hoangltt@hcmute.edu.vn*
[5]*Master of Engineering, Faculty of Electrical and Electronics Engineering, Cao Thang Technical College, Ho Chi Minh city, Vietnam; ntthanh@caothang.edu.vn*

***Correspondence:** *Trong Nghia Le; trongnghia@hcmute.edu.vn*

ABSTRACT- This paper presents a power system fault identification method by simultaneously applying the Kendall and Spearman correlation coefficients for feature selection, combined with an Artificial Neural Network (ANN) to enhance accuracy and optimize training time. Experimental results indicate that Kendall demonstrates superior performance in handling nonlinear data and mitigating the impact of outliers, leading to more optimal fault identification outcomes. Backpropagation Neural Network (BPNN), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) models are trained on datasets after feature selection using both correlation coefficients. The results show that Kendall significantly improves model performance, with BPNN + Kendall achieving RMSE = 0.0001, MAE = 0.0059, and a training time of 0.8946s, outperforming other methods. The proposed approach not only enhances accuracy but also optimizes processing time, reaffirming the effectiveness of Kendall and Spearman in feature selection for power system fault identification.

Keywords: Data preprocessing; Correlation coefficient; Convolutional Neural Network (CNN); Long Short-Term Memory (LSTM); Backpropagation Neural Network (BPNN).

ARTICLE INFORMATION
Author(s): Nhung Le Thi Hong, Trong Nghia Le, Tan Phung Trieu, Hoang Le Thi Thanh, and Thanh Nguyen Tan;
Received:11/03/2025; **Accepted**: 25/04/2025; **Published**: 30/05/2025;
e-ISSN: 2347-470X;
Paper Id: IJEER 1103-05;
Citation: 10.37391/ijeer.130204
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-13/ijeer-130204.html

Publisher's Note: FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

The rapid development of renewable energy and Smart Grids has increased the complexity of power systems, requiring advanced methods for accurate and fast fault identification [1]. Advanced Metering Infrastructure (AMI) [2] generates massive amounts of data, necessitating efficient data mining techniques to maintain grid stability. However, traditional methods such as Fourier Transform and Wavelet Transform face significant limitations when processing nonlinear data and high noise levels [3]. To address this issue, recent studies have focused on ANN as a promising solution for power system fault identification.

ANNs have been widely applied due to their capability to efficiently extract features from measurement data. Among them, BPNN employs the backpropagation algorithm to optimize the model but is prone to overfitting if features are not appropriately selected. CNN can automatically extract spatial features from measurement signals, reducing dependence on preprocessing [4]. Meanwhile, LSTM is designed to handle time-series data, enabling real-time fault identification with high accuracy. Additionally, hybrid models such as CNN-LSTM and CNN-Transformer are being studied to simultaneously leverage spatial and temporal features, enhancing identification performance [5].

A crucial factor influencing the effectiveness of deep learning models is feature selection. Proper feature selection helps reduce overfitting risks, optimize computational costs, and improve model accuracy [6], [7]. In this study, we propose a Kendall-Spearman-based feature selection method to reduce data dimensionality while preserving essential information. Kendall and Spearman are capable of assessing nonlinear relationships between variables, proving more effective than linear selection methods like Pearson [8], [9].

The reviewed studies point out several limitations in traditional and deep learning-based fault detection methods in power systems. Conventional techniques like Fourier and Wavelet

Transforms struggle with nonlinear, noisy data. While deep learning models such as BPNN, CNN, and LSTM improve accuracy, they also face specific challenges: BPNN is prone to overfitting, CNN relies on spatial features, and LSTM depends on time-series structure. Hybrid models aim to combine strengths but increase complexity. A major issue across all approaches is proper feature selection, as poor selection can lead to overfitting and high computational costs.

In this study, we propose a model integrating Kendall-Spearman correlation coefficients with ANN to optimize fault identification in power systems through four main steps: (1) Preparing IEEE 37-Bus data, (2) Feature selection using Kendall-Spearman, (3) Data normalization and classification, and (4) Training CNN, LSTM, and BPNN models. Model performance is evaluated based on RMSE, MAE, and comparison with traditional models. Experimental results show that Kendall achieves the best performance, with the BPNN-Kendall model obtaining RMSE = 0.0001, MAE = 0.0059, and a short training time of 0.8946s, demonstrating its effectiveness in power system fault detection.

In conclusion, this study addresses the lack of efficient nonlinear feature selection methods for power system fault data by integrating Kendall-Spearman correlation with ANN models [10]. Subsequently, CNN, LSTM, and BPNN models perform automatic feature extraction, minimizing other complex preprocessing steps [11],[12]. This approach integrates two powerful techniques: feature selection via Kendall and Spearman, and Artificial Neural Networks, enabling maximum extraction of useful information while eliminating redundant data without compromising accuracy [13].

2. LITERATURE REVIEW

The increasing complexity of modern power systems, particularly with the integration of renewable energy and smart grid technologies, has heightened the demand for more accurate and robust fault detection methods. Traditional techniques such as Fourier and Wavelet transforms are often limited when handling nonlinear, noisy signals commonly found in smart grid data.

Recent advances have shifted attention to machine learning (ML) and deep learning (DL) approaches, especially artificial neural networks (ANNs), due to their capability to model complex patterns. However, the effectiveness of such models heavily relies on data preprocessing, with feature selection emerging as a crucial step. By selecting informative variables, feature selection enhances model accuracy, reduces training time, and improves generalization.

Although Pearson correlation is widely used in feature selection, it assumes linear relationships and fails to capture nonlinear dependencies. In contrast, rank-based correlation measures like Spearman's rho and Kendall's tau offer non-parametric alternatives better suited for complex, high-dimensional datasets. These techniques have shown promise in other domains but remain underutilized in power system fault detection.

Parallelly, DL architectures such as LSTM and CNN have demonstrated high performance in fault classification tasks—LSTM excels at capturing temporal dependencies, while CNN is effective in extracting spatial features. Some studies have explored hybrid models (e.g., CNN-LSTM) for improved performance. Yet, the integration of nonlinear feature selection techniques into these DL frameworks is still an emerging topic. This study addresses this gap by proposing a hybrid approach that combines nonlinear correlation-based feature selection (using Kendall and Spearman methods) with deep learning models (CNN and LSTM) to improve fault detection accuracy and robustness in noisy, real-time environments of modern power systems.

3. METHODOLOGY

3.1. Fault Identification in Power Systems

Power system consists of critical components such as generators, transmission lines, loads, transformers, and protective devices. Faults may arise due to environmental factors, equipment failures, external disturbances, or operational errors, leading to power outages, equipment damage, and fire hazards. Common fault types include overload, short circuit, disconnection, and loss of synchronization.

Traditional and modern fault identification methods are illustrated in *figure 1*. Early and accurate fault detection ensures stable and secure power system operation.

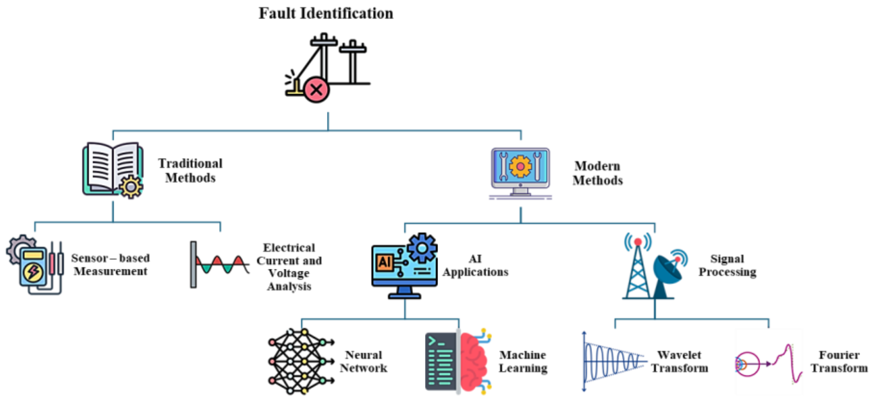


Figure 1. Fault Identification Methods

3.2. Data Mining

3.2.1. Importance of Data Analysis Before Training Neural Network Models

Pre-training data analysis is essential to ensure data quality and enhance model performance. Identifying key variables helps reduce noise and optimize the training process. Removing invalid data prevents errors caused by missing values, outliers, or sensor inaccuracies. Examining relationships among variables provides a better understanding of data structure and improves model efficiency.

3.2.2. Fundamental Steps in Data Mining

The basic steps of data mining are presented in *figure 2*.

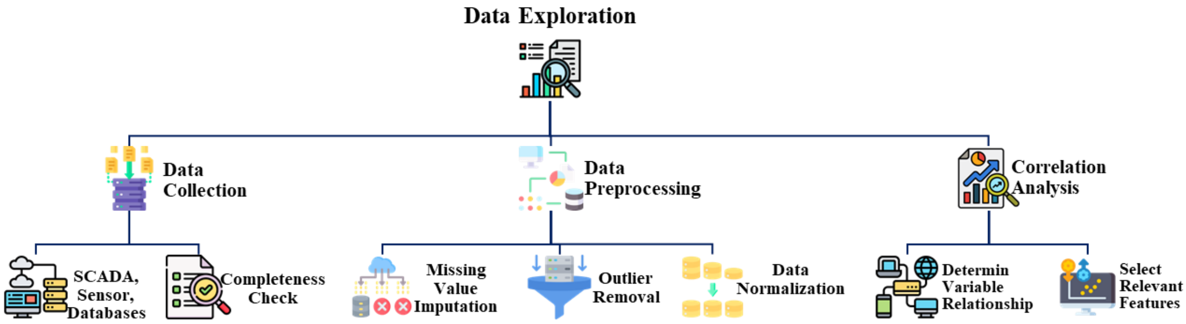


Figure 2. Data Mining Steps

3.2.3. Correlation Analysis and Its Impact on Model Performance

Correlation analysis evaluates the influence of input variables on the target variable, facilitating effective feature selection. Correlation coefficients measure the relationships between variables, which is particularly useful when dealing with nonlinear or noisy data. When two features exhibit high correlation, one can be removed to eliminate redundancy, improving model generalization.

Feature selection based on correlation analysis not only reduces model complexity but also enhances predictive accuracy. This approach allows the model to focus on essential variables, minimize noise, optimize processing speed, and improve forecasting efficiency, especially when handling large datasets.

3.3. Spearman and Kendall Correlation Coefficients

The Spearman and Kendall correlation coefficients measure the degree of association between two variables, particularly useful when data does not follow a normal distribution. These coefficients are commonly used to evaluate relationships between input factors and the target variable. The formulas for Spearman's correlation coefficient (ρ) and Kendall's correlation coefficient (τ) are presented in equations (2.1) and (2.2), respectively.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (2.1)$$

$$\tau = \frac{C-D}{\sqrt{(C+D+T)\cdot(C+D+U)}} \qquad (2.1)$$

ρ operates based on the principle of measuring the monotonic relationship between two variables by comparing their ranks rather than their actual values. In contrast, *τ* evaluates the level of agreement between the rankings of two variables by considering the number of concordant and discordant pairs in the observations. ρ and τ support feature selection by assessing the significance of each variable. Features with a high correlation to the target variable are retained to enhance prediction accuracy, while less relevant features may be removed to prevent redundancy. When two features exhibit a high correlation, one of them can be eliminated to reduce dimensionality without losing critical information. The analysis and evaluation of the strength or weakness of the Peason, Spearman and Kendall values to support the decision to remove or retain the feature has been mentioned by many works [14,15] and is presented in *table 1*. In this paper, values with Strong correlation are retained.

This approach enhances data preprocessing efficiency, optimizes computational speed, and improves model accuracy.

Table 1. Possible cut/removal values for rank-based correlation statistics

Strength	Pearson	Spearman	Kendall
Negligible	0.00	0.00	0.00
Weak	0.10	0.10	0.01
Moderate	0.40	0.38	0.26
Strong	0.70	0.68	0.49
Very strong	0.90	0.89	0.71

3.4. Error Evaluation in Fault Identification

Error is a crucial criterion for assessing the performance of fault identification models in power systems. Two commonly used metrics for measuring model accuracy are Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The formulas for RMSE and MAE are presented in Equations (2.3) and (2.4), respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_1)^2} \qquad (2.3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_1| \quad (2.4)$$

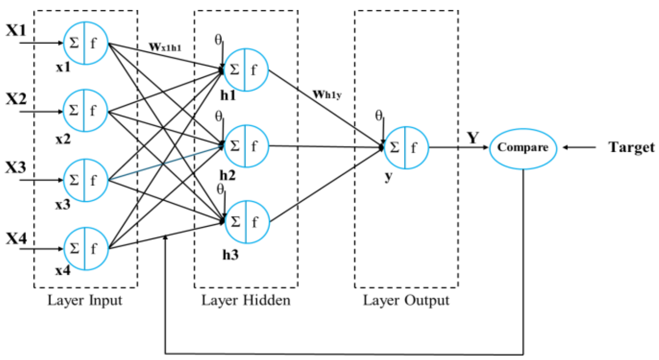
While RMSE is sensitive to outliers due to the squaring of errors, MAE provides a more intuitive measure of the average error.

Error metrics play a critical role in evaluating the accuracy of fault identification models in power systems. RMSE and MAE quantify the deviation between predicted and actual values, facilitating model comparison, optimal model selection, and parameter tuning to enhance performance. In particular, RMSE's sensitivity to outlier's aids in detecting anomalies in data, thereby improving the reliability of the system.

3.5. Application of Neural Networks in Power System Fault Identification

3.5.1. Overview of BPNN and LSTM

ANN are widely applied in power system fault identification, with BPNN and LSTM being two prominent architectures. BPNN is a feedforward neural network that utilizes the backpropagation algorithm to optimize weights by minimizing the error between predicted outputs and actual values. It is well-suited for classification and regression tasks, capable of learning complex relationships between input and output data [16].



LSTM is a variant of Recurrent Neural Networks (RNN) designed to handle long-term sequential data by retaining essential information while discarding irrelevant details. Through its gating mechanism, LSTM effectively addresses fault identification problems where data exhibits temporal dependencies and nonlinear relationships [17].

Both models offer distinct advantages: BPNN is simple and easy to implement but may struggle with long-term sequential data, whereas LSTM overcomes this limitation with superior memory retention capabilities.

3.5.2. Overview of CNN

CNNs are deep learning models specialized in processing grid-like data structures such as images and time-series signals. These models automatically extract features using learned filters, reducing the need for manual feature engineering. CNNs operate through convolutional filters and nonlinear layers to capture complex patterns [16]. A typical CNN architecture consists of:

- **Convolutional Layer**: Performs convolution operations between learned filters and input data to extract local features.
- **Pooling Layer**: Reduces the spatial dimensions of feature maps from the convolutional layer, decreasing the number of parameters and enhancing generalization.
- **Activation Layer**: Applies a nonlinear activation function to increase the model's representational capacity.
- **Fully Connected Layer**: Connects all neurons and produces the final output.

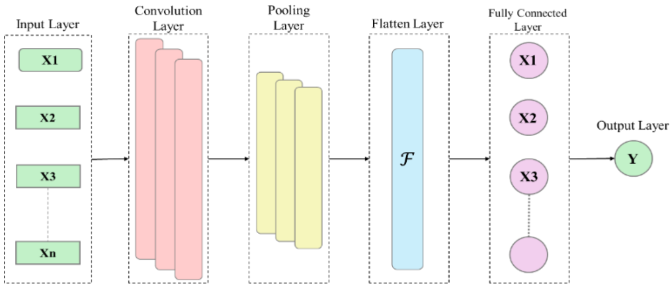


Figure 5. CNN Architecture

3.5.3. Comparison of CNN, LSTM, and BPNN

CNN is well-suited for image-based or matrix-structured signal data due to its strong feature extraction capabilities. LSTM is optimized for time series data, making it effective for analyzing and forecasting fault trends. BPNN has a simple architecture and is easy to implement but is less effective for data with complex temporal dependencies. The choice of model depends on the characteristics of the data and the requirements of the fault identification system.

Table 2. Comparison of CNN, LSTM, and BPNN

Criteria	CNN	LSTM	BPNN
Time Series Processing	Limited (requires signal transformation or 1D CNN)	Good (remembers long-term dependencies)	Poor (lacks temporal memory)
Feature Extraction	Excellent (efficient in spatial and local feature learning)	Inferior to CNN (depends on input data)	Average (requires preprocessed features)
Computational Efficiency	High (fewer parameters due to convolution mechanism)	Moderate (more parameters due to state retention)	High (simple architecture)
Main Applications	Fault image recognition, waveform signal analysis	Load forecasting, fault sequence analysis	Fault classification, nonlinear regression

4. APPROACH AND MODEL DEVELOPMENT

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 2 | Pages 209-217 | e-ISSN: 2347-470X

4.1. Data Collection and Processing

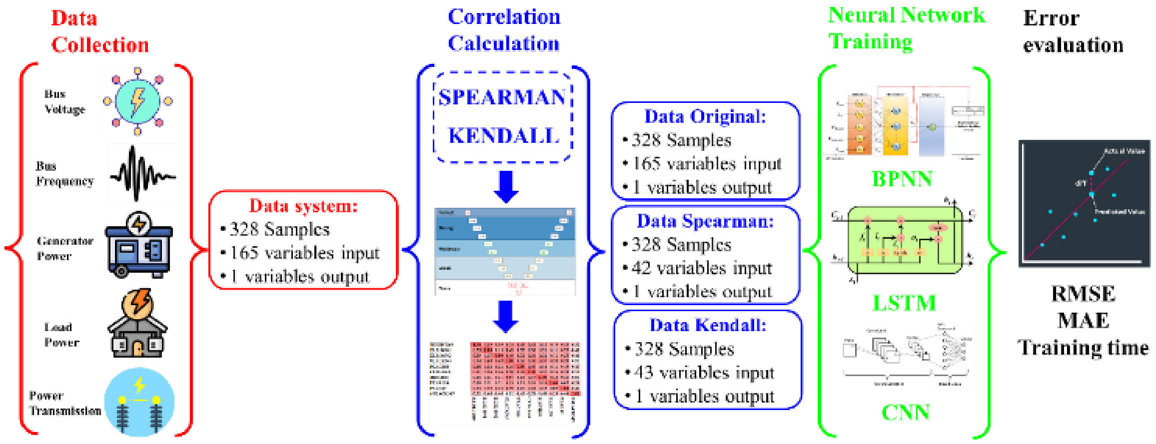


Figure 6. Description of the Data Collection and Processing Procedure

The data processing procedure consists of the following steps:

Step 1: Data collection is performed by simulating the IEEE-37 Bus power grid model using PowerWorld Simulator Version 21. Fault scenarios involving generator outages, except for the main generator at the slack bus, are simulated. Each scenario represents a load level ranging from 100% to 60%, resulting in a total of 328 data samples with 165 variables.

Step 2: From the initial analysis in Section 3, the frequency at the Slack Bus is compared with various electrical characteristics in the power system, including (U_{Bus}, f_{Bus}, $P_{Generator}$, P_{Load}, P_{Branch}).

Step 3: The correlation between variable pairs is evaluated using Spearman's and Kendall's correlation coefficients. In the case of Spearman, it is greater than 0.68 and Kendall is greater than 0.49. This results in three datasets: Original Data (Or_D), Spearman-filtered Data (Sp_D) , and Kendall-filtered Data (K_D).

Step 4: The dataset is split into 80% for training and 20% for testing. Neural network architectures for BPNN, LSTM, and CNN are constructed and trained on each processed dataset.

The basic network configuration parameters are as follows:

- BPNN: 1 input layer, 1 output layer, 1 hidden layer with 8 hidden neurons using the Levenberg-Marquardt training algorithm.
- LSTM: 1 input layer, 1 output layer with the structure: fullyConnectedLayer(1) + regressionLayer. The hidden layers consist of 3 sequential LSTM layers with 200 hidden units on each layer, followed by 3 LSTM layers + 1 fullyConnectedLayer(50) + 1 reluLayer.
- CNN: 1 input layer, 1 output layer with the structure: fullyConnectedLayer(1) + regressionLayer. The hidden layer consists of 1 CNN layer with 64 hidden units.

Step 5: The trained models are evaluated based on **RMSE and MAE** on the test set (20%). Additionally, training time is considered as a performance metric.

4.2. Evaluating the Relationship between Power Grid Characteristics and Slack Bus Frequency

Evaluating the Relationship between Power Grid Characteristics and Slack Bus Frequency using ρ and τ is evaluated in *tables 3* and *table 4*.

Table 3. Relationship Evaluation Based on Kendall's Correlation Coefficient (τ)

Pg		P(load)				Pb							
Generator 1	0.32	Load1	0.00	Load14	0.02	Branch 1	0.07	Branch 15	0.19	Branch 29	0.00	Branch 43	0.62
Generator 2	0.46	Load2	0.03	Load15	0.03	Branch 2	0.07	Branch 16	0.24	Branch 30	0.00	Branch 44	0.12
Generator 3	0.46	Load3	0.04	Load16	0.06	Branch 3	0.05	Branch 17	0.26	Branch 31	0.05	Branch 45	0.24
Generator 4	0.92	Load4	0.02	Load17	0.08	Branch 4	0.04	Branch 18	0.30	Branch 32	0.06	Branch 46	0.02
Generator 5	0.43	Load5	0.02	Load18	0.26	Branch 5	0.27	Branch 19	0.06	Branch 33	0.06	Branch 47	0.02
Generator 6	0.51	Load6	0.29	Load19	0.01	Branch 6	0.27	Branch 20	0.14	Branch 34	0.12	Branch 48	0.01
Generator 7	0.60	Load7	0.02	Load20	0.05	Branch 7	0.19	Branch 21	0.14	Branch 35	0.16	Branch 49	0.23
Generator 8	0.29	Load8	0.03	Load21	0.00	Branch 8	0.14	Branch 22	0.14	Branch 36	0.05	Branch 50	0.29

Research Article | Volume 13, Issue 2 | Pages 209-217 | e-ISSN: 2347-470X

Generator 9	0.40	Load9	0.03	Load22	0.22	Branch 9	0.28	Branch 23	0.27	Branch 37	0.24	Branch 51	0.29
		Load10	0.01	Load23	0.04	Branch 10	0.11	Branch 24	0.14	Branch 38	0.02	Branch 52	0.25
		Load11	0.03	Load24	0.01	Branch 11	0.22	Branch 25	0.01	Branch 39	0.03	Branch 53	0.20
		Load12	0.25	Load25	0.03	Branch 12	0.27	Branch 26	0.01	Branch 40	**0.59**	Branch 54	0.01
		Load13	0.00	Load26	0.12	Branch 13	0.15	Branch 27	0.22	Branch 41	0.02	Branch 55	0.09
						Branch 14	0.15	Branch 28	0.21	Branch 42	**0.67**	Branch 56	0.20
		U						**f**					
		U-Bus1	0.04	U-Bus13	0.28	U-Bus25	0.01	f-Bus1	**0.99**	f-Bus13	**0.73**	f-Bus25	**0.99**
		U-Bus2	0.01	U-Bus14	0.01	U-Bus26	0.05	f-Bus2	**0.98**	f-Bus14	**0.71**	f-Bus26	**0.87**
		U-Bus3	0.04	U-Bus15	0.03	U-Bus27	0.04	f-Bus3	**0.92**	f-Bus15	**0.82**	f-Bus27	**0.94**
		U-Bus4	0.03	U-Bus16	0.02	U-Bus28	0.01	f-Bus4	**0.86**	f-Bus16	**0.81**	f-Bus28	**0.95**
		U-Bus5	0.02	U-Bus17	0.31	U-Bus29	0.04	f-Bus5	**0.87**	f-Bus17	**0.97**	f-Bus29	**0.88**
		U-Bus6	0.01	U-Bus18	0.11	U-Bus30	0.06	f-Bus6	**0.84**	f-Bus18	**0.97**	f-Bus30	**0.77**
		U-Bus7	0.31	U-Bus19	0.08	U-Bus31	0.04	f-Bus7	**0.68**	f-Bus19	**0.95**	f-Bus31	**0.71**
		U-Bus8	0.01	U-Bus20	0.06	U-Bus32	0.01	f-Bus8	**0.74**	f-Bus20	**0.93**	f-Bus32	**0.77**
		U-Bus9	0.01	U-Bus21	0.08	U-Bus33	0.23	f-Bus9	**0.80**	f-Bus21	**0.93**	f-Bus33	**0.75**
		U-Bus10	0.03	U-Bus22	0.10	U-Bus34	0.04	f-Bus10	**0.87**	f-Bus22	**0.72**	f-Bus34	**0.73**
		U-Bus11	0.02	U-Bus23	0.29	U-Bus35	0.02	f-Bus11	**0.90**	f-Bus23	**0.99**	f-Bus35	**0.82**
		U-Bus12	0.03	U-Bus24	0.04	U-Bus36	0.02	f-Bus12	**0.87**	f-Bus24	**0.90**	f-Bus36	**0.99**
						U-Bus37	0.10					f-Bus37	**1.00**

Table 4. Relationship Evaluation Based on Spearman's Correlation Coefficient (ρ)

Pg		Pl				Pb							
Generator 1	0.50	Load1	0.07	Load14	0.03	Branch 1	0.17	Branch 15	0.33	Branch 29	0.08	Branch 43	**0.80**
Generator 2	0.40	Load2	0.09	Load15	0.03	Branch 2	0.17	Branch 16	0.34	Branch 30	0.08	Branch 44	0.30
Generator 3	0.40	Load3	0.04	Load16	0.10	Branch 3	0.13	Branch 17	0.52	Branch 31	0.11	Branch 45	0.39
Generator 4	**0.99**	Load4	0.05	Load17	0.09	Branch 4	0.13	Branch 18	0.41	Branch 32	0.25	Branch 46	0.20
Generator 5	0.63	Load5	0.05	Load18	0.39	Branch 5	0.49	Branch 19	0.11	Branch 33	0.25	Branch 47	0.20
Generator 6	**0.70**	Load6	0.41	Load19	0.07	Branch 6	0.49	Branch 20	0.23	Branch 34	0.30	Branch 48	0.06
Generator 7	**0.78**	Load7	0.04	Load20	0.12	Branch 7	0.34	Branch 21	0.23	Branch 35	0.31	Branch 49	0.38
Generator 8	0.39	Load8	0.04	Load21	0.08	Branch 8	0.29	Branch 22	0.23	Branch 36	0.22	Branch 50	0.45
Generator 9	0.58	Load9	0.04	Load22	0.31	Branch 9	0.51	Branch 23	0.39	Branch 37	0.39	Branch 51	0.45
		Load10	0.07	Load23	0.05	Branch 10	0.26	Branch 24	0.27	Branch 38	0.14	Branch 52	0.52
		Load11	0.04	Load24	0.07	Branch 11	0.44	Branch 25	0.07	Branch 39	0.14	Branch 53	0.32
		Load12	0.37	Load25	0.03	Branch 12	0.39	Branch 26	0.07	Branch 40	**0.72**	Branch 54	0.10

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 2 | Pages 209-217 | e-ISSN: 2347-470X

Load13	0.08	Load26	0.10	Branch 13	0.30	Branch 27	0.33	Branch 41	0.20	Branch 55	0.30
				Branch 14	0.30	Branch 28	0.33	Branch 42	0.84	Branch 56	0.33
U						f					
U-Bus1	0.04	U-Bus13	0.28	U-Bus25	0.01	f-Bus1	0.99	f-Bus13	0.73	f-Bus25	0.99
U-Bus2	0.01	U-Bus14	0.01	U-Bus26	0.05	f-Bus2	0.98	f-Bus14	0.71	f-Bus26	0.87
U-Bus3	0.04	U-Bus15	0.03	U-Bus27	0.04	f-Bus3	0.92	f-Bus15	0.82	f-Bus27	0.94
U-Bus4	0.03	U-Bus16	0.02	U-Bus28	0.01	f-Bus4	0.86	f-Bus16	0.81	f-Bus28	0.95
U-Bus5	0.02	U-Bus17	0.31	U-Bus29	0.04	f-Bus5	0.87	f-Bus17	0.97	f-Bus29	0.88
U-Bus6	0.01	U-Bus18	0.11	U-Bus30	0.06	f-Bus6	0.84	f-Bus18	0.97	f-Bus30	0.77
U-Bus7	0.31	U-Bus19	0.08	U-Bus31	0.04	f-Bus7	0.68	f-Bus19	0.95	f-Bus31	0.71
U-Bus8	0.01	U-Bus20	0.06	U-Bus32	0.01	f-Bus8	0.74	f-Bus20	0.93	f-Bus32	0.77
U-Bus9	0.01	U-Bus21	0.08	U-Bus33	0.23	f-Bus9	0.80	f-Bus21	0.93	f-Bus33	0.75
U-Bus10	0.03	U-Bus22	0.10	U-Bus34	0.04	f-Bus10	0.87	f-Bus22	0.72	f-Bus34	0.73
U-Bus11	0.02	U-Bus23	0.29	U-Bus35	0.02	f-Bus11	0.90	f-Bus23	0.99	f-Bus35	0.82
U-Bus12	0.03	U-Bus24	0.04	U-Bus36	0.02	f-Bus12	0.87	f-Bus24	0.90	f-Bus36	0.99
				U-Bus37	0.10					f-Bus37	1.00

5. RESULTS AND ANALYSIS

Feature Selection Results Using ρ and τ is evaluated in *tables 5* and *table 6.*

Table 5. Feature Selection Results Using Kendall's Correlation Coefficient (τ)

Pg		Pb		f											
Generator 4	0.99	Branch 40	0.72	f-Bus1	0.99	f-Bus7	0.68	f-Bus13	0.73	f-Bus19	0.95	f-Bus25	0.99	f-Bus31	0.71
Generator 6	0.70	Branch 42	0.84	f-Bus2	0.98	f-Bus8	0.74	f-Bus14	0.71	f-Bus20	0.93	f-Bus26	0.87	f-Bus32	0.77
Generator 7	0.78	Branch 43	0.80	f-Bus3	0.92	f-Bus9	0.80	f-Bus15	0.82	f-Bus21	0.93	f-Bus27	0.94	f-Bus33	0.75
				f-Bus4	0.86	f-Bus10	0.87	f-Bus16	0.81	f-Bus22	0.72	f-Bus28	0.95	f-Bus34	0.73
				f-Bus5	0.87	f-Bus11	0.90	f-Bus17	0.97	f-Bus23	0.99	f-Bus29	0.88	f-Bus35	0.82
				f-Bus6	0.84	f-Bus12	0.87	f-Bus18	0.97	f-Bus24	0.90	f-Bus30	0.77	f-Bus36	0.99
														f-Bus37	1.00

Table 6. Feature Selection Results Using Spearman's Correlation Coefficient (ρ)

Pg		Pb		f											
Generator 4	0.99	Branch 40	0.72	f-Bus1	0.99	f-Bus8	0.74	f-Bus14	0.71	f-Bus20	0.93	f-Bus26	0.87	f-Bus32	0.77
Generator 6	0.70	Branch 42	0.84	f-Bus2	0.98	f-Bus9	0.80	f-Bus15	0.82	f-Bus21	0.93	f-Bus27	0.94	f-Bus33	0.75
Generator 7	0.78	Branch 43	0.80	f-Bus3	0.92	f-Bus10	0.87	f-Bus16	0.81	f-Bus22	0.72	f-Bus28	0.95	f-Bus34	0.73
				f-Bus4	0.86	f-Bus11	0.90	f-Bus17	0.97	f-Bus23	0.99	f-Bus29	0.88	f-Bus35	0.82
				f-Bus5	0.87	f-Bus12	0.87	f-Bus18	0.97	f-Bus24	0.90	f-Bus30	0.77	f-Bus36	0.99
				f-Bus6	0.84	f-Bus13	0.73	f-Bus19	0.95	f-Bus25	0.99	f-Bus31	0.71	f-Bus37	1.00

From *table 5* and *table 6*, the training results indicate that Spearman's correlation coefficient reduces the dataset to 42 features, including 3 generator variables, 3 power transmission variables, and 36 bus frequency variables. Meanwhile, Kendall's correlation coefficient reduces the dataset to 43 features, consisting of 3 generator variables, 3 power transmission variables, and 37 bus frequency variables. These refined datasets are then used for training in different neural network architectures (BPNN, LSTM, CNN) to evaluate their effectiveness and applicability.

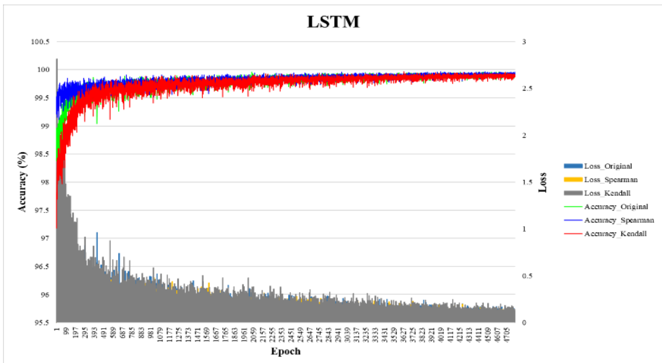
The comparison results of error and training time of the cases using BPNN, LSTM, CNN combined with Org_D, K_D and Sp_D are presented in *table 7*.

Table 7. Comparison of Error Metrics and Training Time for BPNN, LSTM, CNN

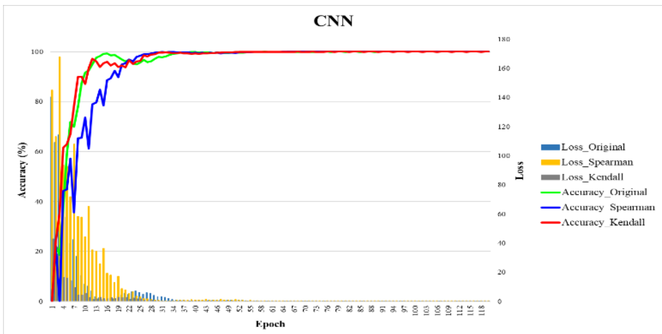
	BPNN			LSTM			CNN		
	Or_D	K_D	Sp_D	Or_D	K_D	Sp_D	Or_D	K_D	Sp_D
RMSE	0.0612	**0.0001**	0.0002	0.6407	**0.2994**	0.3464	0.5510	**0.4137**	0.4569
MAE	0.0612	**0.0059**	0.0075	0.5042	**0.2992**	0.3462	0.5510	**0.4149**	0.4569
Training Time (s)	1.0549	0.8946	**0.7719**	250.49	**221.54**	229.61	**1.4246**	1.7967	1.8312

The results indicate that BPNN with Kendall correlation achieved the best performance, exhibiting the lowest error values (RMSE = 0.0001, MAE = 0.0059) and the shortest training time (0.8946s). While LSTM demonstrated strong learning capabilities, its training time was significantly higher. On the other hand, CNN maintained relatively stable accuracy but showed inconsistencies in training time. These findings confirm that BPNN combined with Kendall correlation is the optimal choice for fault identification tasks when utilizing an 80% training set and a 20% test set.

Moreover, there is no standard range of RMSE and MAE values that can be universally considered "good." In [18], an RMSE ranging from 0.2 to 0.5 is considered acceptable for relatively accurate prediction models, especially when the data is normalized in the range [0,1]. Therefore, cross-comparison between methods is appropriate.



(a) LSTM



(b) CNN

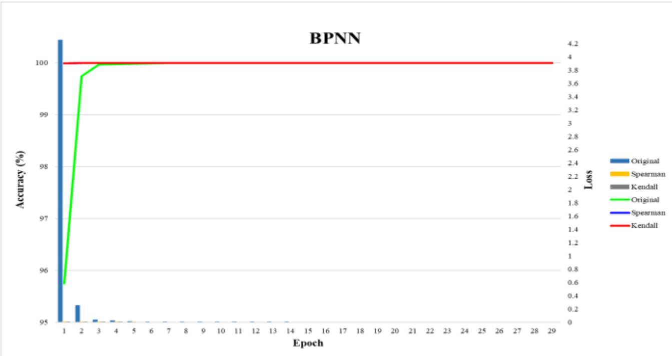


Figure 7. Model Error and Performance Graph

Ultimately, BPNN trained on the dataset filtered using the Kendall coefficient yielded optimal performance. Moreover, when the same dataset was applied to other networks, the results were also superior compared to using the original dataset or the dataset processed using Spearman correlation. Feature selection significantly reduced the dataset dimensionality, thereby alleviating computational demands on neural networks, enhancing forecasting accuracy, and considerably shortening training time.

6. CONCLUSION

The research findings confirm that data filtering using the Kendall correlation coefficient optimizes model performance, reduces error, and shortens training time. The training process was conducted with 80% of the data allocated for training and 20% for validation. Among the three evaluated models, BPNN with Kendall filtering produced the best results, with the lowest error rates (RMSE = 0.0001, MAE = 0.0059) and a short training time of 0.8946 seconds.

This underscores the critical role of data preprocessing in improving model accuracy and efficiency for fault identification tasks. Additionally, the findings highlight that the BPNN model exhibits high potential for real-world applications due to its fast-training capability and low error rates, particularly when using the 80%-20% data split strategy.

Although this study was conducted in a simulated environment, the proposed approach demonstrates strong potential for deployment in real-time monitoring systems. To extend this work, future research will involve testing the model with real-world SCADA data to better evaluate its robustness and practical applicability.

Model performance was assessed using RMSE and MAE—standard and interpretable metrics that directly quantify prediction error magnitudes in regression-based and time-series tasks. While statistical significance testing such as the t-test can complement these metrics by verifying differences between models, our primary evaluation focused on the magnitude of error, which is more intuitive and practical for engineering applications.

Confidence intervals and robust analysis under varying operational conditions were not addressed in the current scope but are recognized as valuable additions for comprehensive model evaluation. These aspects will be considered in future work using real-world SCADA datasets.

Acknowledgement: This work belongs to the project grant number: T2024-120 funded by Ho Chi Minh City University of Technology and Education, Vietnam.

REFERENCES

[1] M. Khalid, "Smart grids and renewable energy systems: Perspectives and grid integration challenges," Energy Strategy Reviews, vol. 51, p. 101299, Jan. 2024.

[2] J. S. Guerrero-Prado, W. Alfonso-Morales, and E. F. Caicedo-Bravo, "A Data Analytics/Big Data Framework for Advanced Metering Infrastructure Data," Sensors, vol. 21, no. 16, p. 5650, Aug. 2021.

[3] S. Khatoon, Ibraheem, M. Shahid, G. Sharma, E. Çelik, E. Bekiroğlu, M. F. Ahmer, and Priti, "Near Real-Time Load Forecasting of Power System Using Fuzzy Time Series, Artificial Neural Networks, and Wavelet Transform Models," Electric Power Components and Systems, vol. 51, no. 10, pp. 796–810, Aug. 2023.

[4] X. Wang, N. Fatehi, C. Wang, and M. H. Nazari, "Deep Learning-Based Weather-Related Power Outage Prediction with Socio-Economic and Power Infrastructure Data," in Proc. 2024 IEEE Power & Energy Society General Meeting (PESGM), Jul. 2024.

[5] S. Madasthu, A. A. Mamun, A. Abbas, E. Abbate, B. Chowdhury, and R. Cox, "Ensemble Deep Learning Model for Power System Outage Prediction for Resilience Enhancement," in Proc. 2023 North American Power Symposium (NAPS), Oct. 2023.

[6] N. Ahmed, M. Assadi, Q. Zhang, and A. A. Ahmed, "Assessing impact of borehole field data's input parameters on hybrid deep learning models for heating and cooling forecasting: A local and global explainable AI analysis," in IOP Conference Series: Materials Science and Engineering, vol. 1294, no. 1, p. 012056, 2023.

[7] A. F. Amiri, A. Chouder, H. Oudira, S. Silvestre, and S. Kichou, "Improving Photovoltaic Power Prediction: Insights through Computational Modeling and Feature Selection," Energies, vol. 17, no. 13, p. 3078, June 2024.

[8] M. V. Suntrakumar, S. S. B. S. Ahmad, M. L. Samsudin, and M. R. Othman, "Comparative Analysis of Filter, Wrapper, and Embedded Feature Selection Methods to Predict Load Progression in Combined Cycle Power Plant (CCPP) Startup Across Diverse Startup Classes," in Proc. 2024 8th Int. Conf. Power Energy Syst. Appl. (ICoPESA), June 2024.

[9] N. Das, S. Deb, and A. K. Goswami, "Day-ahead Wind Power Prediction using Optimised XGBoost and Correlation Analysis based Noise Reduction Technique," in Proc. 2022 IEEE Int. Conf. Power Electron., Drives Energy Syst. (PEDES), Dec. 2022.

[10] M. Feng, S. Wang, H. Yu, Q. Guo, and B. Li, "Enhancing Fault Detection in Power Load Management Systems: A Neural Network Approach with Seq2Seq-CNN-LSTM Models", Mar. 2024.

[11] J. B. Thomas, S. G. Chaudhari, S. K. V., and N. K. Verma, "CNN-Based Transformer Model for Fault Detection in Power System Networks," IEEE Trans. Instrum. Meas., vol. 72, no. 2504210, Jan. 2023.

[12] H. A. Alghamdi, "A Time Series Forecasting of Global Horizontal Irradiance on Geographical Data of Najran Saudi Arabia," Energies, vol. 15, no. 3, p. 928, Jan. 2022.

[13] H. Karimipour, A. Dehghantanha, R. M. Parizi, K.-K. R. Choo, and H. Leung, "A Deep and Scalable Unsupervised Machine Learning System for Cyber-Attack Detection in Large-Scale Smart Grids," IEEE Access, vol. 7, pp. 80778–80788, May 2019.

[14] P. Schober, C. Boer, and L. A. W. Schwarte, "Correlation coefficients: Appropriate use and interpretation" Anesthesia & Analgesia, vol. 126, no. 5, pp. 1763–1768, May 2018.

[15] A. Zhang, D. Sun, X. Wang, and Y. Gao, "Assessment of water resources carrying capacity in Tianjin City based on ecological footprint theory," Sustainability, vol. 11, no. 12, p. 3271, Jun. 2019.

[16] P. Zhang, Z. Cui, Y. Wang, và S. Ding, "Application of BPNN optimized by chaotic adaptive gravity search and particle swarm optimization algorithms for fault diagnosis of electrical machine drive system," Electrical Engineering, vol. 104, no. 2, pp. 819–831, Apr. 2022.

[17] A. Siami-Namini and N. Tavakoli, "RNN-LSTM: From applications to modeling techniques and beyond," Journal of King Saud University - Computer and Information Sciences, vol. 36, no. 5, pp. 102068, 2024.

[18] Srivastava, A. K., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., Gaiser, T., & Rahimi, J. (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-06249-w