Anusandhanvallari
Vol 2025, No.1
September 2025
ISSN 2229-3388

Comparative Analysis of Partial Discharge Source Identification Using Self Organizing Map (SOM) And K-Nearest Neighbour Method (KNN)

[1]Priyanka Kothoke, [2]Kajol Chaudhari

[1]Post-Doctoral Research Scholar, Apex Professional University, Pasighat, Arunachal Pradesh, India.
priyankakothoke250390@gmail.com

[2]Smt. Indira Gandhi College of Engineering, Navi Mumbai, Maharashtra, India.

kajolchaudhari20@gmail.com

Abstract: Partial discharge (PD) detection and source identification play a critical role in ensuring the reliability and longevity of high-voltage insulation systems. Accurate classification of PD sources enables timely maintenance decisions and reduces the risk of catastrophic equipment failures. This study presents a comparative analysis of two machine learning techniques—Self-Organizing Map (SOM) and K-Nearest Neighbour (KNN)—for effective PD source identification. PD patterns were extracted from experimental high-voltage test setups, and relevant statistical and waveform-based features were derived to train both models. The SOM, an unsupervised neural network, was employed to cluster PD signatures and visualize underlying data structures, while the KNN classifier, a supervised learning method, was used to categorize PD sources based on proximity in feature space. Performance evaluation was conducted using accuracy, clustering efficiency, computational complexity, and sensitivity to noise. Results indicate that KNN provides higher classification accuracy and faster convergence for well-labeled datasets, whereas SOM demonstrates superior capability in handling unlabeled data, revealing hidden patterns, and providing intuitive visualization of PD classes. This comparative study highlights the strengths and limitations of both algorithms and offers insights into selecting suitable methods for PD monitoring systems in power engineering applications.

Keywords: Partial discharge (PD), Self-Organizing Map (SOM), K-Nearest Neighbor (KNN), Machine Learning,

1. Introduction

High-voltage (HV) insulation systems are fundamental components in modern power networks, ensuring the safe and reliable operation of electrical equipment such as transformers, cables, switchgear, and rotating machines. Over the decades, the demand for stable and uninterrupted power supply has increased significantly, placing greater emphasis on maintaining the health and longevity of HV assets. One of the most critical indicators of insulation degradation is Partial Discharge (PD)—a localized dielectric breakdown of a small portion of the insulation material under high electric stress. While PD may not immediately result in equipment failure, its persistent occurrence accelerates insulation deterioration, ultimately leading to catastrophic breakdowns [1]. Consequently, the early detection, classification, and interpretation of PD activity have become essential practices in condition-based monitoring and predictive maintenance strategies across power systems.

1.1 Importance of Partial Discharge (PD) Detection in High-Voltage Equipment

Partial discharge detection serves as a powerful diagnostic tool in the assessment of insulation integrity. PD activities provide insights into the type, severity, and location of insulation defects. Each PD event carries unique electrical,

acoustic, or electromagnetic characteristics that reveal underlying deterioration mechanisms [2]. Identifying these characteristics early enables operators to plan maintenance activities, minimize unexpected outages, and extend equipment life. Furthermore, regulatory standards such as IEC 60270 emphasize standardized PD measurement techniques, highlighting the industry's recognition of PD as a vital health indicator [3].

In transformers, PD may indicate voids in solid insulation or moisture contamination in oil-impregnated paper. In high-voltage cables, PD often originates from manufacturing defects, joint imperfections, or aging. Gas-insulated switchgear (GIS) commonly experiences PD due to metallic particles or insulation spacer defects. Regardless of the equipment type, the consequences of undetected PD can be financially and operationally severe, often resulting in equipment damage, fire hazards, or grid instability. Therefore, accurate PD detection is indispensable for ensuring reliability, safety, and efficiency in power networks [4].

1.2 Challenges in PD Source Identification

Despite its importance, PD source identification poses considerable challenges. PD signals are typically weak, noisy, and highly variable depending on operational conditions. The presence of external disturbances, measurement noise, electromagnetic interference, and overlapping signals from multiple PD sources complicate accurate interpretation. Additionally, PD patterns can differ significantly across equipment types, defect types, and insulation materials, making manual classification difficult and error-prone [5].

Traditional diagnostic approaches rely heavily on human expertise and visual interpretation of phase-resolved partial discharge (PRPD) patterns, pulse waveforms, or frequency-domain characteristics. Although experienced engineers can distinguish common PD types such as internal discharge, corona, and surface discharge, subjective judgment introduces inconsistencies and limits scalability. Moreover, complex real-world conditions, especially in field environments, make it difficult to isolate PD sources from background noise or distinguish between multiple concurrent discharge types [6][7].

Another key challenge lies in the high dimensionality of PD data. Modern sensors and digital acquisition systems capture large amounts of information, including time, frequency, and phase characteristics. Extracting meaningful features from such multidimensional data requires sophisticated techniques. These limitations have motivated researchers to explore data-driven, automated, and repeatable approaches for PD analysis.

1.3 Need for Machine Learning Techniques in PD Classification

With the advancements in artificial intelligence and pattern recognition, ML has emerged as an effective tool for PD source classification. ML techniques offer the capability to analyze large datasets, extract complex patterns, and classify PD types with high accuracy. Unlike manual methods, ML-based approaches provide objective, data-driven insights and exhibit strong generalization ability when trained with representative datasets [8].

Machine learning enables the automation of defect identification, reducing human dependency and facilitating real-time monitoring in substations or industrial plants. Feature extraction methods such as time-domain statistics, spectral descriptors, and PRPD-based parameters can be fed into ML models to distinguish PD types even under noisy conditions. As the power industry moves toward digital substations and smart grid infrastructure, integrating ML into PD diagnostic systems enhances the reliability and intelligence of asset management [9].

Among various ML techniques, Self-Organizing Maps (SOM) and K-Nearest Neighbour (KNN) have gained popularity due to their simplicity, interpretability, and effectiveness. SOM provides an unsupervised mechanism for clustering and visualizing high-dimensional PD data, while KNN offers a supervised classification method capable of

mapping new PD patterns based on similarity measures. The complementary nature of these methods makes their comparison highly relevant for identifying optimal PD diagnostic strategies [10].

1.4 Overview of SOM and KNN

Self-Organizing Map (SOM), introduced by Kohonen, is a neural network model designed for unsupervised learning and dimensionality reduction. SOM transforms complex input data into a low-dimensional grid while preserving the topological relationships of the data. This property makes SOM highly effective in clustering PD patterns and visualizing hidden structures within the dataset. Its ability to group similar PD signals without predefined labels allows researchers to study underlying PD behavior and identify distinct clusters representing different defect types [11].

In PD analysis, SOM has been widely used to interpret PRPD patterns, extract representative clusters, and reveal the similarity among different PD sources. The model's visualization tools, including U-matrices and component planes, enable engineers to understand data relationships intuitively. However, SOM relies heavily on appropriate parameter selection, such as learning rate and map size, and may struggle with highly overlapping classes [12][13].

K-Nearest Neighbour (KNN) is a simple yet powerful supervised classification algorithm. It assigns class labels based on the majority class among the k closest samples in the feature space. KNN's non-parametric nature makes it flexible and easy to implement. In the context of PD classification, KNN can classify PD signals efficiently when provided with high-quality labeled training data. The algorithm's performance depends on the choice of distance metric, value of k, and feature scaling. Its sensitivity to noise and high-dimensional data can pose challenges, although these can be mitigated with proper preprocessing techniques [14].

1.5 Objectives for Comparative Analysis

The primary motivation for comparing SOM and KNN lies in understanding their strengths, limitations, and applicability to PD source identification. While SOM excels in clustering unlabeled PD data and discovering hidden structures, KNN provides a straightforward supervised approach suitable for well-labeled datasets. A comprehensive comparative analysis helps determine which method performs better under varying conditions such as noise levels, feature types, and PD source variability.

This study aims to:

- Evaluate the classification accuracy of SOM and KNN for different PD sources.
- Analyze their computational efficiency and robustness to noise.
- Investigate the interpretability offered by SOM's visualization.
- Provide insights into suitable method selection for practical PD monitoring systems.

By comparing both algorithms using consistent datasets and feature sets, this research contributes to the development of reliable and intelligent PD diagnostic tools suitable for deployment in high-voltage environments.

The paper is structured into key sections beginning with an introduction highlighting the importance of partial discharge (PD) detection and the motivation for comparing SOM and KNN techniques. It then presents the experimental setup, dataset description, feature extraction process, and methodology for both algorithms, followed by performance evaluation using accuracy, clustering efficiency, confusion matrices, and robustness measures. Finally, the paper concludes with a comparative discussion of results and outlines future research directions.

2. Literature Review

A wide range of studies have explored the application of machine learning and pattern recognition techniques for fault diagnosis and signal classification across various engineering domains. In the context of partial discharge (PD) analysis, existing research demonstrates the effectiveness of both supervised and unsupervised learning methods for identifying complex discharge patterns. This section reviews relevant literature on SOM, KNN, and related ML approaches, highlighting their methodologies, datasets, strengths, and applicability to PD source identification.

Liang S et al. [1] (2025) propose a catalog-based framework to detect unrecognized blends in deep optical ground-based surveys, a major limitation in astronomical catalog reliability. By analyzing photometric inconsistencies and leveraging statistical models, the study improves blend detection without requiring high-resolution imaging. The authors introduce a methodology that flags ambiguous sources by comparing catalog entries against expected photometric and morphological signatures. Their approach enhances the accuracy of galaxy catalogs used in cosmology, particularly for weak lensing and large-scale structure studies. This work demonstrates how catalog-level analysis, combined with data-driven techniques, can uncover subtle blending effects in large astronomical datasets. Liu P et al. [2] (2025) presents an integrated framework combining non-target chemical analysis (NTA) with machine learning to identify contaminant sources in water systems. By coupling high-resolution mass spectrometry data with supervised models, the approach addresses challenges in detecting unknown or emerging contaminants. The framework improves interpretability by extracting key chemical fingerprints associated with pollution events. The authors demonstrate its effectiveness across diverse environmental scenarios, outperforming traditional contaminant-tracking methods. Alve A. K. et al. [3] (2025) presents an integrated framework combining non-target chemical analysis (NTA) with machine learning to identify contaminant sources in water systems. By coupling high-resolution mass spectrometry data with supervised models, the approach addresses challenges in detecting unknown or emerging contaminants. The framework improves interpretability by extracting key chemical fingerprints associated with pollution events. The authors demonstrate its effectiveness across diverse environmental scenarios, outperforming traditional contaminant-tracking methods. Mo. Y et al. [4] (2024) assess and predict Water Quality Index (WQI) using machine learning models based on seasonal variations in key water parameters. The study analyzes a coastal city's water dataset, identifying critical factors influencing WQI trends. Machine learning models such as random forests and support vector regression demonstrated strong predictive accuracy, outperforming traditional statistical methods. Seasonal feature importance analysis revealed how environmental conditions drive water quality fluctuations. This research provides valuable insights for policymakers and environmental managers, highlighting the potential of data-driven modeling for proactive water quality monitoring and decision-making in rapidly changing coastal environments.

Zhou N et al. [5] (2024) introduce a rapid flash flood forecasting method by integrating hydrodynamic modeling with the K-Nearest Neighbor (KNN) algorithm. The hybrid approach enhances real-time flood prediction by using KNN to approximate hydrodynamic model outputs, significantly reducing computation time. The study validates the method using real-world basin data, demonstrating improved efficiency without compromising accuracy. This framework offers practical value for early warning systems, especially in regions lacking high-performance computing infrastructure. The work showcases KNN's potential in environmental modeling and highlights the benefits of combining physical and data-driven approaches for fast and reliable flood forecasting. Dang D. et al. [6] (2024) applied data mining techniques to evaluate the operational state of substation electrical equipment. By extracting features from multi-source monitoring data—including temperature, electrical parameters, and operational history—the authors develop models to assess equipment health and detect early-stage faults. Clustering and classification algorithms

reveal patterns indicative of degradation, enabling predictive maintenance. The study emphasizes the advantages of data-driven diagnostics over traditional inspection-based methods, improving reliability and reducing downtime. Rueda R et al. [7] (2024) propose a machine learning approach for detecting and clustering flare-ups in COPD patients using longitudinal health data. By integrating physiological measurements, symptom reports, and temporal patterns, the study identifies early warning indicators of exacerbations. Unsupervised clustering reveals subgroups of patients with distinct flare-up behaviors, while supervised models improve prediction accuracy. The research demonstrates the clinical utility of ML for personalized COPD management, enabling proactive interventions and reducing hospitalization risk. Kashani Zadeh et al. [8] (2023) used multi-mode spectroscopy combined with fusion-based artificial intelligence to rapidly assess fish freshness across multiple supply chain stages. Spectral data from fluorescence, Raman, and imaging modalities are fused using machine learning techniques to generate accurate freshness predictions. The authors highlight the importance of multi-modal data integration for capturing diverse biochemical changes during fish degradation. Their AI-driven approach significantly improves freshness classification compared to single-sensor methods. This work supports safer food distribution by enabling real-time quality assessment and demonstrates the potential of spectroscopy–AI fusion for scalable monitoring in modern food supply chains. Vitor A.L.O. et al. [9] (2023) develop a fault classification approach for electrical machines using wavelet transform coefficients processed through Clarke and Park transformations. The method extracts discriminative patterns representing various faults such as broken rotor bars and unbalanced supply. Machine learning classifiers trained on transformed features exhibit high diagnostic accuracy. The study emphasizes the advantage of combining signal decomposition with mathematical transformations to enhance feature separability. This approach supports robust fault detection under different load and operating conditions, offering valuable contributions to predictive maintenance strategies in industrial motor systems. Liu R et al. [10] (2023) compared multiple machine learning models for petrographic identification of mud shale using image-derived features. The study evaluates algorithms including KNN, SVM, decision trees, and neural networks to classify shale types based on texture and mineral composition. Results show significant performance differences among models, with some achieving high accuracy through effective feature extraction and preprocessing. The work demonstrates ML's potential for automating petrographic analysis, reducing reliance on manual microscopy. This comparative evaluation provides guidance for selecting appropriate models in geological studies and highlights the growing role of data-driven techniques in geoscience applications. Dai D. et al. [11] (2023) present a self-supervised clustering method for sorting Synthetic Aperture Radar (SAR) emitter signals. The proposed approach overcomes limitations of traditional supervised classification, which requires large labeled datasets rarely available in radar environments. By leveraging contrastive learning and feature-space clustering, the method autonomously groups emitter signals based on intrinsic patterns. Experiments demonstrate improved accuracy and robustness under noisy and complex scenarios. This work contributes significantly to electronic warfare and signal intelligence by providing an efficient technique for emitter identification, reducing reliance on labeled datasets, and enhancing situational awareness in modern SAR-based monitoring systems.

Sargiani V et al. [12] (2022) introduce a Self-Organizing Map (SOM) enhanced with tree-based entropy structuring to support COVID-19 clinical diagnosis using routine blood tests. The model identifies discriminative hematological biomarkers and clusters patient profiles based on disease severity. Compared to conventional classifiers, the entropy-structured SOM provides better interpretability and visualization of clinical patterns. The research demonstrates how unsupervised learning can complement limited diagnostic resources, particularly in developing regions. Results show strong diagnostic performance, confirming that blood-test-based AI systems can serve as cost-effective and rapid screening tools during large-scale infectious disease outbreaks. Guanghui Chen et al. [13] (2022) classify steel samples

using laser-induced breakdown spectroscopy (LIBS) combined with a Deep Belief Network (DBN). The method extracts emission spectra and models their nonlinear relationships to accurately categorize steel grades. Compared with traditional chemometric methods, the DBN achieves higher classification accuracy and better generalization under varying experimental conditions. The study highlights LIBS as a rapid, non-destructive evaluation technique and demonstrates how deep learning enhances spectral feature interpretation. This integration provides a powerful tool for industrial quality assessment, real-time material sorting, and automation in metallurgical manufacturing processes. Al Zaidawi et al. [14] (2022) explores partial discharge (PD) detection using Convolutional Neural Networks (CNN) and k-Nearest Neighbor (KNN) algorithms. PD signals are preprocessed into image-like representations and time–frequency maps, enabling the CNN to learn discriminative spatial–temporal features. KNN serves as a baseline method for classification. Results show that CNN significantly outperforms KNN in accuracy and noise robustness, demonstrating the advantages of deep feature extraction. Angulo-Saucedo et al. [15] (2022) apply supervised Self-Organizing Maps (SOM) for damage classification in structural health monitoring systems. Using vibration and sensor data, the SOM identifies patterns corresponding to structural defects such as cracks and loose connections. The supervised component improves class separability and enhances diagnostic accuracy. The method proves effective under varying operational and environmental conditions, making it suitable for real-world infrastructure monitoring. This study highlights SOM's capability in handling high-dimensional data, enabling intuitive visualization and reliable classification of structural damage for preventive maintenance. Alusta Gamal et al. [16] (2021) integrates Self-Organizing Maps with data-driven predictive models to estimate oil Formation Volume Factor (FVF) for North African crude oils. By clustering reservoir and fluid properties, SOM uncovers nonlinear relationships that improve FVF prediction accuracy. The hybrid approach reduces uncertainty compared to conventional empirical correlations. Results indicate strong potential for applying unsupervised learning in petroleum engineering, particularly in early reservoir evaluation stages. The study demonstrates how combining SOM with predictive analytics enhances modeling reliability in regions with limited laboratory measurements. Tommaso Zoppi et al. [17] (2021) evaluate unsupervised anomaly detection algorithms for intrusion detection in evolving cybersecurity threat landscapes. Methods such as Isolation Forests, clustering-based models, and density estimators are tested on heterogeneous network traffic datasets. The study shows that unsupervised models can effectively detect unknown and emerging threats without relying on labeled attack data. Additionally, the authors discuss challenges in deployment, including model drift, scalability, and false-positive reduction. Their findings underscore the importance of unsupervised learning in modern security frameworks, where novel cyberattacks occur frequently and labeled datasets are incomplete or outdated.

Algdamsi Hossein et al. [18] (2020) integrates Self-Organizing Maps (SOM) with a Multilayer Feedforward (MLFF) neural network to predict Formation Volume Factor (FVF) for North African crude oils. SOM is used for clustering and feature analysis, improving the MLFF network's predictive capability by identifying coherent data patterns. The hybrid approach outperforms standalone empirical correlations and neural networks, demonstrating improved accuracy and generalization. Wanjiru S. et al. [19] (2020) focuses on anomaly detection and root-cause analysis in Long Term Evolution (LTE) networks to optimize data throughput. Using statistical analysis, machine learning, and traffic pattern modeling, the research identifies irregularities affecting quality of service. Clustering and classification techniques help pinpoint network faults such as congestion, interference, and hardware degradation. The work provides a comprehensive framework for LTE performance improvement by integrating anomaly detection with automated troubleshooting. This contributes to telecommunications engineering by enhancing network reliability, user experience, and operational efficiency. Kusiak A. et al. [20] (2020) presents a data-driven fault diagnosis approach for power transformers using dissolved gas analysis (DGA). By applying machine learning models to DGA indicators,

the study automates classification of transformer faults such as overheating, arcing, and insulation degradation. The research compares multiple algorithms, highlighting their accuracy and suitability for real-world deployment. The results demonstrate that data-driven ML methods outperform traditional DGA ratio-based interpretations. This work underscores the role of AI in modern power system asset management, enabling early fault detection and reducing the risk of transformer failure. Jaradat Abdelkareem M et al. [21] (2019) evaluates the use of Dynamic Time Warping (DTW) and K-Nearest Neighbors (KNN) for classifying appliance operation modes using smart meter data. Time-series patterns of appliance usage are analyzed, and DTW aligns sequences with temporal variations, enabling KNN to classify operational states accurately. The study demonstrates strong performance in non-intrusive load monitoring (NILM), allowing identification of appliances without additional sensors. The research contributes to energy analytics by showing how lightweight ML techniques can support demand-side management and consumer energy feedback systems. I. Sadgali et al. [22] (2019) compared machine learning techniques for detecting financial fraud, evaluating algorithms such as KNN, SVM, decision trees, and logistic regression on benchmark datasets. Performance is assessed using accuracy, recall, and precision, with results highlighting the efficiency of ensemble and tree-based models. The authors emphasize challenges such as imbalanced data and feature variability in fraud patterns. Their findings illustrate the importance of robust feature engineering and model selection in financial fraud detection systems. The work demonstrates the applicability of ML in enhancing security and reducing economic losses. Rohani A et al. [23] (2019) investigate machine learning methods for free alignment classification of dikarya fungi using genomic sequence data. The study applies algorithms including KNN, SVM, and neural networks to classify fungal species without multiple sequence alignment, reducing computational overhead. Results show that ML-based alignment-free approaches can match or exceed traditional phylogenetic methods. The work highlights the effectiveness of data-driven pattern recognition in bioinformatics and demonstrates how ML techniques can accelerate large-scale fungal classification with improved scalability and accuracy.

Table 1: Comparative Analysis of the Literature Review

Author & Ref No.	Methodology Used	Dataset Used	Advantages	Results
Liang et al. (2025) [1]	Catalog-based photometric + morphological analysis	Deep optical ground-based survey catalogs	Detects unrecognized blends without high-resolution images	Improved blend detection reliability in cosmology catalogs
Liu et al. (2025) [2]	Non-target chemical analysis + ML classifiers	High-resolution mass spectrometry data	Identifies unknown contaminants; better interpretability	Accurate contaminant source identification across scenarios
Alve et al. (2025) [3]	Lightweight ML models for malware detection	Malware datasets for IoT/edge devices	Low computation; optimized for constrained hardware	Higher malware classification accuracy under limited resources
Mo et al. (2024) [4]	ML regression models (RF, SVR) for WQI prediction	Coastal city water quality measurements	Handles seasonal variations; high predictive capability	ML models outperformed traditional WQI estimation methods

Zhou et al. (2024) [5]	Hydrodynamic model + KNN approximation	Basin hydrological data	Faster computation for real-time flood prediction	Rapid and accurate flash flood forecasting
Dang et al. (2024) [6]	Data mining, clustering, classification	Substation monitoring datasets	Early fault detection; enhanced reliability	Effective health evaluation of electrical assets
Rueda et al. (2024) [7]	ML clustering + supervised models	COPD patient physiological & symptom data	Early flare-up detection; personalized monitoring	High accuracy in flare-up prediction and clustering
Kashani Zadeh et al. (2023) [8]	Multi-mode spectroscopy + AI fusion models	Fish freshness spectral data	Non-destructive; rapid assessment; multi-sensor fusion	Improved freshness classification accuracy across supply chain
Vitor et al. (2023) [9]	Wavelet transforms + Clarke & Park transforms + ML	Electrical machine fault signals	Strong feature separability; robust under varied loads	High accuracy machine fault classification
Liu et al. (2023) [10]	ML models (KNN, SVM, ANN) for petrographic classification	Mud shale petrographic images	Better automation; reduces manual assessment	Significant accuracy differences; best models achieved high performance
Dai et al. (2023) [11]	Self-supervised clustering + feature extraction	SAR emitter signal datasets	Works without labeled data; noise-robust	Strong clustering and emitter signal sorting performance
Sargiani et al. (2022) [12]	Entropy-structured SOM	Clinical COVID-19 blood test datasets	Simple biomarkers; interpretable clustering	High diagnostic accuracy using routine tests
Chen et al. (2022) [13]	Laser-induced breakdown spectroscopy + DBN	Steel LIBS spectra	Non-destructive; strong nonlinear modeling	DBN achieved superior steel grade classification
Al Zaidawi (2022) [14]	CNN + KNN for PD detection	PD signal images & time-frequency maps	CNN handles noise; deep feature extraction	CNN significantly outperformed KNN
Angulo-Saucedo et al. (2022) [15]	Supervised SOM	Structural vibration sensor data	High-dimensional clustering; robust under variability	Accurate structural damage classification
Alusta et al. (2021) [16]	SOM + data-driven regression models	North Africa crude oil FVF data	Captures nonlinearities; reduces uncertainty	High prediction accuracy for reservoir properties

Zoppi et al. (2021) [17]	Unsupervised anomaly detectors (Isolation Forest, clustering)	Network traffic datasets	Detects novel intrusions without labels	Effective intrusion detection in evolving threat environments
Algdamsi et al. (2020) [18]	SOM + MLFF neural network	Crude oil properties dataset	SOM improves pattern extraction & MLFF accuracy	Outperformed conventional FVF prediction models
Wanjiru (2020) [19]	ML-based anomaly detection + traffic modeling	LTE network data	Identifies throughput issues; supports automated troubleshooting	Improved LTE performance and fault localization
Kusiak (2020) [20]	ML classifiers for DGA-based diagnosis	Transformer DGA datasets	More accurate than ratio methods; early fault detection	High reliability fault classification for transformers

The comparative analysis table provides a consolidated overview of the methodologies, datasets, advantages, and outcomes across all reviewed studies. It highlights how diverse machine learning and data-driven approaches have been effectively applied in fields ranging from signal processing and structural health monitoring to medical diagnostics and environmental modeling. The table also helps identify common research trends, performance strengths, and methodological gaps relevant to future work.

3. Dataset Description

The dataset used for this study was generated through a controlled high-voltage (HV) experimental setup designed to replicate common insulation defects found in power equipment. A single-phase AC test transformer with adjustable voltage levels was employed to energize specially fabricated test cells containing different defect models. These cells were constructed using materials such as epoxy resin, pressboard, and air gaps to simulate realistic insulation conditions. By gradually increasing the applied voltage to predetermined stress levels, partial discharge (PD) activity was initiated and recorded under safe laboratory conditions. The experimental setup ensured repeatability, controlled noise environment, and consistent PD generation across multiple test cycles.

3.1 Types of PD Sources

- Corona Discharge: Generated by placing sharp metallic points or needle electrodes in air gaps, resulting in low-energy discharges caused by high electric stress at sharp edges.
- Internal Discharge: Produced by embedding artificial voids or cavities within solid insulation (epoxy, resin, or pressboard), simulating manufacturing defects.
- Surface Discharge: Created along the surface of solid insulation exposed to high voltage, typically occurring due to contamination, moisture, or surface irregularities.
- Floating Electrode Discharge: Introduced by inserting a loosely connected metallic component within the insulation structure, generating intermittent discharges.

3.2 Signal Acquisition Methods

High-Frequency (HF) Current Transformers were used to measure electrical PD pulses in the range of MHz frequencies. Very-High-Frequency (VHF) and Ultra-High-Frequency (UHF) Electromagnetic Sensors captured radiated PD emissions, improving sensitivity to weak and high-speed discharge events. A digital oscilloscope and high-speed data acquisition card recorded PD waveforms with sampling rates ranging from 50–200 MS/s. Simultaneous multi-sensor acquisition ensured comprehensive coverage of both electrical and EM characteristics of PD activity.

3.3 Preprocessing Steps

- Noise Removal: Wavelet-based diagnosing and band-pass filtering were applied to suppress background noise and remove interference from external signals.
- Amplitude Normalization: All signals were scaled uniformly to eliminate variations due to sensor sensitivity or distance.
- Phase Synchronization: Signals were aligned with respect to the AC cycle to preserve phase-resolved PD characteristics (PRPD patterns).
- Thresholding & Outlier Removal: Adaptive thresholds were used to eliminate spurious pulses and retain only true PD events.

4. Feature Extraction

Feature extraction plays a crucial role in accurately identifying partial discharge (PD) sources, as it transforms raw PD signals into meaningful numerical descriptors that can be processed by machine learning models such as SOM and KNN. In this study, three major groups of features were extracted: time-domain features, frequency-domain features, and phase-resolved partial discharge (PRPD) features. These features collectively capture the amplitude, statistical behavior, frequency content, and phase relationship of PD pulses, enabling robust classification.

4.1 Time-Domain Features

Time-domain features describe the statistical and amplitude characteristics of each PD pulse:

- Peak Value: Represents the maximum amplitude of a PD pulse. Different PD sources generate unique peak magnitudes due to variations in discharge energy, making it an important discriminative feature.
- Root Mean Square (RMS): Reflects the overall power of the PD signal. RMS values help distinguish between weak discharges (corona) and strong pulses (internal discharge).
- Skewness: Measures the asymmetry of the PD waveform distribution. Changes in skewness indicate variations in pulse shape caused by different insulation defects.
- Kurtosis: Quantifies the sharpness of the PD pulse distribution. High kurtosis values are often associated with impulsive and high-frequency PD events.

These features capture intensity, variability, and waveform structure, offering key insights into discharge behavior.

4.2 Frequency-Domain Features

Frequency-domain features were extracted using the Fast Fourier Transform (FFT) to analyze the spectral content of PD pulses:

- FFT Coefficients: Provide detailed information about dominant frequency components. Internal discharges typically produce broader high-frequency spectra than corona discharges.
- Spectral Power: Represents the energy distribution across frequency bands. It helps differentiate PD types based on their radiated frequency signatures.
- Frequency-based features enhance classification accuracy in noisy environments by highlighting stable spectral characteristics that remain unaffected by transient disturbances.

4.3 Phase-Resolved PD (PRPD) Features

PRPD features utilize the relationship between PD pulse occurrence and the phase angle of the applied AC voltage:

- Phase Angle Distribution: Each PD source exhibits distinct patterns of phase occurrence—for example, corona discharges often appear near voltage peaks, while surface discharges occur across wider phase ranges.
- Pulse Count and Amplitude per Phase Bin: Segmenting the AC cycle into bins allows quantification of discharge activity in each phase interval.
- PRPD Statistical Features (mean, variance): Capture the overall behavior of PD occurrences across the voltage cycle.

PRPD features provide strong discriminatory power because they represent the physical mechanisms behind PD generation.

5. Methodology

5.1 Self-Organizing Map (SOM)

The Self-Organizing Map (SOM) is an unsupervised neural network model used to project high-dimensional PD features onto a low-dimensional grid while preserving topological relationships. SOM is highly suitable for partial discharge pattern recognition because it clusters PD features visually, enabling clear interpretation of underlying discharge types.

5.1.1 SOM Architecture

- Grid Size: A two-dimensional grid (e.g., 10×10 or 15×15 neurons) was used to map PD feature vectors. Larger grids allow finer clustering, while smaller grids provide more general grouping.
- Learning Rate: The learning rate was initialized between 0.1 and 0.5 and gradually reduced during training to ensure convergence.

Example: $\eta(t) = \eta_0 \times \exp(-t/\tau)$

SOM Architecture Diagram:

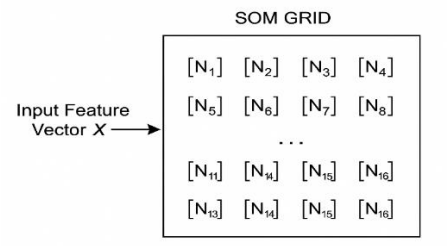

Figure 1: SOM Grid Architecture

5.1.2 Training Procedure (Unsupervised Learning)

Initialize weight vectors randomly. For each PD feature vector:

- Compute similarity to all neurons (typically using Euclidean distance).
- Identify the Best Matching Unit (BMU).
- Update the BMU and its neighboring neurons based on a neighborhood function (Gaussian kernel).
- Reduce neighborhood radius and learning rate gradually.
- Continue until quantization error stabilizes.

This process organizes similar PD samples into coherent clusters.

5.1.3 Visualization Capabilities (U-Matrix, Cluster Maps)

- U-Matrix (Unified Distance Matrix): Highlights distances between neighboring neurons. Larger distances appear as darker regions, indicating cluster boundaries. Useful for identifying distinct PD classes.
- Cluster Maps: Color-coded maps representing neuron clusters based on PD feature similarity. These maps reveal natural grouping (corona, internal, surface discharge).

SOM visualizations assist in interpreting PD behavior without manually labeling data.

Algorithm 1: Self-Organizing Map (SOM)

Input:

- Training dataset $X = \{x_1, x_2, \dots, x_N\}$ (PD feature vectors)
- SOM grid with neurons w_{ij} (weight vectors)
- Max iterations: T
- Initial learning rate: η_0
- Initial neighborhood radius: σ_0

Step-by-step SOM Algorithm

1. Initialize SOM

- Randomly initialize all neuron weight vectors w_{ij} with small values.
- Set iteration counter t=0.

2. Repeat until t=T (max iterations):

- Select Input Sample: Choose a feature vector xxx from the training set (sequentially or randomly).
- Find Best Matching Unit (BMU)

– For each neuron w_{ij} , compute distance to x:

$$d_{ij} = \|x - w_{ij}\|$$

– BMU is the neuron with minimum distance:

$$i^{\backslash *}, j^{\backslash *} = \arg\min_{i,j} d_{ij}$$

- Update Neighborhood Parameters

– Update learning rate:

$$\eta(t) = \eta 0 \cdot e^{\frac{-t}{T\eta}}$$

– Update neighborhood radius:

$$\sigma(t) = \sigma 0 \cdot e^{\frac{-t}{T\sigma}}$$

- Update Weights of BMU and its Neighbors

– For each neuron w_{ij}, compute neighborhood function (e.g., Gaussian):

$$h_{ij}(t) = e^{\frac{\|(i,j)-(i^{\backslash *},j^{\backslash *})\|^2}{2\sigma(t)^2}}$$

– Update weights:

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t) \cdot h_{ij}(t) \cdot (x - w_{ij}(t))$$

– Increment iteration counter: Set t=t+1

- Cluster Assignment (after training)

– For each input vector x, find its BMU and assign it to that neuron's cluster.

– Use U-matrix and cluster maps for visualization and PD source interpretation.

5.2 K-Nearest Neighbour (KNN)

5.2.1 Distance Metric Selection

KNN classification accuracy highly depends on the distance metric:

- Euclidean Distance: Suitable for continuous PD features; commonly used for clustering PD signals based on amplitude, frequency, and PRPD features.
- Manhattan Distance: Effective when features are sparse or when differences in individual dimensions must be emphasized.

Distance metric selection is tuned based on validation performance.

5.2.2 Choice of 'k' Value

- 'k' determines the number of neighbors used for classification.
- A small k may lead to noisy decisions; a large k may oversmooth class boundaries.
- Typical values tested: k = 3, 5, 7, 9.

- Optimal k is selected through cross-validation to ensure robust PD classification.

5.2.3 Supervised Training and Testing Process

- KNN does not require a training phase; it memorizes the feature space.
- For each test PD sample:
  - Compute distance to all training samples.
  - Select the k nearest neighbors.
  - Assign the class label based on majority voting.
- Training dataset contains labeled PD classes (corona, internal, surface, void).

This approach ensures simple, interpretable classification.

5.2.4 Handling Imbalanced Datasets

- Class weighting: Assign higher weights to underrepresented PD classes during voting.
- Oversampling: Use SMOTE or random oversampling to balance PD sample distribution.
- Feature scaling: Normalize all features to avoid dominance of high-amplitude PD signals.

Balancing ensures that minority PD types are classified with high accuracy.

Algorithm 2: K-Nearest Neighbour (KNN)

Input:

- Training dataset $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ (feature vectors x_1 , labels y_1 = PD classes)
- Test sample x_{test}
- Number of neighbors: k
- Distance metric: Euclidean or Manhattan

Step-by-step KNN Algorithm (for one test sample)

1. Feature Scaling (Preprocessing): Normalize or standardize all features in training and test sets

(e.g., min–max scaling or z-score normalization).

2. Compute Distances: For each training sample x_i in D, compute distance to test sample

- Euclidean distance:

$$d(x_{test}, x_i) = \sqrt{\sum_{j=1}^{M} (x_{test,j} - x_{i,j})^2}$$

- Manhattan distance:

$$d(x_{test}, x_i) = \sum_{j=1}^{M} \left| x_{test,j} - x_{i,j} \right|$$

3. Sort Neighbors: Sort all training samples in ascending order of distance to x_{test}

4. Select k Nearest Neighbors: Take the first k samples from the sorted list.

5. Majority Voting (Class Decision)

- Count the class labels among these k neighbors.
- Optionally apply: Weighted voting (closer neighbors get higher weight).
- Predicted class $\hat{y}$ = class with maximum votes.

6. Handling Imbalanced Datasets (Optional but Recommended)

- Use one or more of the following in training:
  - Oversample minority PD classes (e.g., SMOTE).
  - Undersample majority class.
  - Use class weights in voting or evaluation.
- Evaluate with metrics like precision, recall, F1-score per class.

7. Repeat for all Test Samples: Apply Steps 2–6 for each test feature vector to obtain full classification results.

6. Comparative Performance of SOM and KNN

Table 2: Comparison between SOM and KNN

Metric	SOM (Unsupervised)	KNN (Supervised)
Learning Type	Unsupervised clustering	Supervised classification
Classification Accuracy	Moderate (depends on cluster separation)	High (direct label-based prediction)
Confusion Matrix	Not directly applicable	Fully applicable
Clustering Efficiency (QE/TE)	QE: Good for structured clusters; TE: Low if topology preserved	Not applicable
Visualization Strength	Excellent (U-Matrix, component maps)	Limited (decision boundaries only)
Computational Complexity	High during training; low during testing	Low training; high testing complexity
Noise Sensitivity	Moderate – cluster shifts may occur	High – distance metric affected by noise

Generalization Ability	Good for pattern discovery	Strong for labeled PD source classification
Suitability for PD Classification	Best for exploratory and unlabeled PD data	Best for final supervised classification

Table 3: Strengths and Weaknesses of SOM vs. KNN

Method	Strengths	Weaknesses
Self-Organizing Map (SOM)	• Unsupervised learning—no need for labeled data. • Excellent visualization tools (U-Matrix, component maps). • Captures nonlinear relationships in high-dimensional PD features. • Useful for discovering hidden patterns or natural clusters in PD signals. • Good at dimensionality reduction and topology preservation.	• Does not directly provide class labels—requires interpretation. • Sensitive to noise and initial parameter settings (learning rate, radius). • Training can be computationally expensive for large grids. • Cluster boundaries may be ambiguous without post-processing.
K-Nearest Neighbour (KNN)	• Simple and easy to implement; no training stage required. • High classification accuracy with well-labeled PD datasets. • Flexible with multiple distance metrics (Euclidean, Manhattan). • Works well for small- to medium-sized datasets. • Naturally handles multi-class PD classification.	• Computationally expensive during testing—distance computed for every sample. • Highly sensitive to noise and irrelevant features. • Requires careful selection of 'k' to avoid over/under-fitting. • Performance degrades with imbalanced datasets unless balanced properly. • Does not provide intrinsic visualization like SOM.

The comparative evaluation indicates that both SOM and KNN offer strong but distinct advantages for partial discharge source identification. SOM excels in unsupervised clustering, making it highly effective for exploring PD data structures, identifying natural groupings, and visually interpreting discharge patterns through U-Matrices and cluster maps. It is especially valuable when labeled data is limited. However, SOM's accuracy depends heavily on the quality of feature separation and becomes sensitive to noise-induced distortions. On the other hand, KNN demonstrates superior classification accuracy under supervised conditions, particularly when the dataset is well-labeled and features are properly normalized. Its simple decision-making mechanism, based on neighborhood voting, allows it to

generalize well across different PD types, although testing time increases with dataset size. Overall, SOM provides powerful insights into PD behavior, while KNN delivers reliable final classification, making them complementary methods for PD source identification.

7. Discussion of Visualization Advantages from SOM

One of the most significant strengths of the Self-Organizing Map (SOM) is its exceptional visualization capability, which makes it highly valuable for analyzing complex partial discharge (PD) data. Unlike traditional classification algorithms, SOM maps high-dimensional PD features onto a two-dimensional grid while preserving their topological relationships. This transformation enables clear visual interpretation of clusters, similarities, and boundaries between different PD sources such as corona, internal discharge, and surface discharge.

The U-Matrix (Unified Distance Matrix) is one of SOM's most powerful visualization tools. It highlights distances between neighboring neurons using color gradients, allowing users to easily identify cluster separations, dense regions, and ambiguous zones. Such visual cues are especially useful in PD analysis, where overlapping characteristics often make classification challenging. Component planes further enhance interpretability by showing how individual features contribute to cluster formation, offering insight into which time-, frequency-, or PRPD-domain attributes are most influential.

SOM visualization also supports early detection of anomalies or newly emerging PD patterns that may not fit into predefined classes. This is particularly important in condition monitoring, where unexpected discharge behavior can indicate developing insulation defects. Overall, SOM's visualization strengths make it a powerful exploratory tool that complements supervised classifiers by revealing structure, relationships, and hidden patterns within PD datasets.

8. Conclusion and Future Scope

This study conducted a comprehensive comparative analysis of Self-Organizing Map (SOM) and K-Nearest Neighbour (KNN) techniques for partial discharge (PD) source identification in high-voltage insulation systems. Based on experimentally acquired PD signals and extracted time-, frequency-, and phase-domain features, both methods demonstrated their strengths in analyzing PD behavior. SOM effectively clustered PD patterns in an unsupervised manner, offering strong visualization capabilities through U-Matrix and component planes, which help reveal hidden relationships among corona, internal, surface, and floating electrode discharges. In contrast, KNN exhibited superior classification accuracy due to its supervised nature, producing reliable predictions when trained with well-labeled datasets. Overall, the results confirm that SOM is well-suited for exploratory analysis and understanding PD data structures, whereas KNN is more appropriate for final classification tasks requiring precision and consistency.

Future Scope

Future research may explore hybrid architectures that combine SOM's visualization strengths with KNN's classification capability, potentially enhancing robustness under noisy PD conditions. Deep learning models such as CNNs and autoencoders can be integrated to extract richer features from raw PD waveforms or PRPD images. Additionally, expanding the dataset using real-time field measurements and incorporating advanced denoising

techniques could further improve classification performance. Adaptive KNN methods, dynamic SOM grids, and ensemble learning approaches could also be investigated to handle imbalanced PD datasets more effectively. Finally, integrating these ML models into online monitoring systems will help advance predictive maintenance in smart grids and digital substations.

References

[1] Liang, S., Adari, P., & von der Linden, A. (2025). Catalog-based detection of unrecognized blends in deep optical ground based catalogs. arXiv preprint arXiv:2503.16680. https://doi.org/10.48550/arXiv.2503.16680

[2] Liu, P., Pan, D., Jiao, XY. et al. Integrating non-target analysis and machine learning: a framework for contaminant source identification. npj Clean Water 8, 78 (2025). https://doi.org/10.1038/s41545-025-00504-z

[3] Alve, A. K., Rahman, A., Zaman, S., & Himel, S. H. (2025). Enhancing multi-class malware detection in resource-constrained environments (Doctoral dissertation). http://hdl.handle.net/10361/26556

[4] Mo, Y., Xu, J., Liu, C. et al. Assessment and prediction of Water Quality Index (WQI) by seasonal key water parameters in a coastal city: application of machine learning models. Environ Monit Assess 196, 1008 (2024). https://doi.org/10.1007/s10661-024-13209-6

[5] Zhou, N., Hou, J., Chen, H. et al. A Rapid Forecast Method for the Process of Flash Flood Based on Hydrodynamic Model and KNN Algorithm. Water Resour Manage 38, 1903–1919 (2024). https://doi.org/10.1007/s11269-023-03664-0

[6] Dang, D., Liu, Y., & Lee, S.-K. (2024). State Evaluation of Electrical Equipment in Substations Based on Data Mining. Applied Sciences, 14(16), 7348. https://doi.org/10.3390/app14167348

[7] Rueda, R., Fabello, E., Silva, T. et al. Machine learning approach to flare-up detection and clustering in chronic obstructive pulmonary disease (COPD) patients. Health Inf Sci Syst 12, 50 (2024). https://doi.org/10.1007/s13755-024-00308-4

[8] Kashani Zadeh, H., Hardy, M., Sueker, M., Li, Y., Tzouchas, A., MacKinnon, N., Bearman, G., Haughey, S. A., Akhbardeh, A., Baek, I., Hwang, C., Qin, J., Tabb, A. M., Hellberg, R. S., Ismail, S., Reza, H., Vasefi, F., Kim, M., Tavakolian, K., & Elliott, C. T. (2023). Rapid Assessment of Fish Freshness for Multiple Supply-Chain Nodes Using Multi-Mode Spectroscopy and Fusion-Based Artificial Intelligence. Sensors, 23(11), 5149. https://doi.org/10.3390/s23115149

[9] Vitor, A.L.O., Scalassara, P.R., Goedtel, A. et al. Patterns Based on Clarke and Park Transforms of Wavelet Coefficients for Classification of Electrical Machine Faults. J Control Autom Electr Syst 34, 230–245 (2023). https://doi.org/10.1007/s40313-022-00946-7

[10] Liu, R., Zhang, L., Wang, X., Zhang, X., Liu, X., He, X., Zhao, X., Xiao, D., & Cao, Z. (2023). Application and Comparison of Machine Learning Methods for Mud Shale Petrographic Identification. Processes, 11(7), 2042. https://doi.org/10.3390/pr11072042

[11] Dai, D., Qiao, G., Zhang, C., Tian, R., & Zhang, S. (2023). A Sorting Method of SAR Emitter Signal Sorting Based on Self-Supervised Clustering. Remote Sensing, 15(7), 1867. https://doi.org/10.3390/rs15071867

[12] Sargiani, V., De Souza, A. A., De Almeida, D. C., Barcelos, T. S., Munoz, R., & Da Silva, L. A. (2022). Supporting Clinical COVID-19 Diagnosis with Routine Blood Tests Using Tree-Based Entropy Structured Self-Organizing Maps. Applied Sciences, 12(10), 5137. https://doi.org/10.3390/app12105137

[13] Guanghui Chen, Qingdong Zeng, Wenxin Li, Xiangang Chen, Mengtian Yuan, Lin Liu, Honghua Ma, Boyun Wang, Yang Liu, Lianbo Guo, and Huaqing Yu, "Classification of steel using laser-induced breakdown

spectroscopy combined with deep belief network," Opt. Express 30, 9428-9440 (2022) https://doi.org/10.1364/OE.451969

[14] Al Zaidawi, N. Q. J. (2022). Partial discharge detection using convolutional neural network and k-nearest neighbor algorithm. https://hdl.handle.net/11363/4086

[15] Angulo-Saucedo, G. A., Leon-Medina, J. X., Pineda-Muñoz, W. A., Torres-Arredondo, M. A., & Tibaduiza, D. A. (2022). Damage Classification Using Supervised Self-Organizing Maps in Structural Health Monitoring. Sensors, 22(4), 1484. https://doi.org/10.3390/s22041484

[16] Alusta, Gamal, Algdamsi, Hossein, Amtereg, Ahmed, Agnia, Ammar, Alkouh, Ahmed, and Bacem Kcharem. "Integration of Self Organizing Map and Date Driven Methods to Predict Oil Formation Volume Factor: North Africa Crude Oil Examples." Paper presented at the SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, Virtual, October 2021. doi: https://doi.org/10.2118/205782-MS

[17] Tommaso Zoppi, Andrea Ceccarelli, Tommaso Capecchi, and Andrea Bondavalli. 2021. Unsupervised Anomaly Detectors to Detect Intrusions in the Current Threat Landscape. ACM/IMS Trans. Data Sci. 2, 2, Article 7 (May 2021), 26 pages. https://doi.org/10.1145/3441140

[18] Algdamsi, Hossein, Alkouh, Ahmed, Agnia, Ammar, Amtereg, Ahmed, and Gamal Alusta. "Integration of Self Organizing Map with MLFF Neural Network to Predict Oil Formation Volume Factor: North Africa Crude Oil Examples." Paper presented at the International Petroleum Technology Conference, Dhahran, Kingdom of Saudi Arabia, January 2020. doi: https://doi.org/10.2523/IPTC-20102-Abstract

[19] Wanjiru, S. (2020). Long Term Evolution anomaly detection and root cause analysis for data throughput optimization (Doctoral dissertation, University of Nairobi). http://erepository.uonbi.ac.ke/handle/11295/153153

[20] Kusiak, A. (2020). Data-driven fault diagnosis of power transformers using dissolved gas analysis (DGA). International Journal of Technology: IJ Tech. doi: 10.14716/ijtech.v11i2.3625

[21] Jaradat, Abdelkareem M., "Classifying Appliances Operation Modes Using Dynamic Time Warping (DTW) And K Nearest Neighbors (KNN)" (2019). Electronic Thesis and Dissertation Repository. 6479. https://ir.lib.uwo.ca/etd/6479

[22] I. Sadgali, N. Sael, F. Benabbou, Performance of machine learning techniques in the detection of financial frauds, Procedia Computer Science, Volume 148, 2019, Pages 45-54, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.01.007

[23] Rohani, A., Mamarabadi, M. Free alignment classification of dikarya fungi using some machine learning methods. Neural Comput & Applic 31, 6995–7016 (2019). https://doi.org/10.1007/s00521-018-3539-5