Semi-Supervised Hybrid Ensemble Learning for Fault Detection in 20kV XLPE Cables

Neng DeXu
Electrical insulation material specialty, Shandong Chint Cable CO., LTD. ShanDong, Jinan 250000, China
E-mail: 13756202621@163.com

Keywords: Cable and wire, cable fault detection, machine learning, ensemble learning, hybrid models, semi-supervised learning, predictive maintenance

Received: February 22, 2025

Cable fault detection is critical for ensuring the reliability and safety of high-voltage 20 kV XLPE cable systems, minimizing downtime and maintenance costs. This research introduces a semi-supervised hybrid ensemble model combining Random Forest, Gradient Boosting, and XGBoost within a Voting Classifier framework. Data preprocessing involves feature scaling and Gaussian noise injection $(\sigma = 0.01)$ *to enhance robustness, followed by training on 3943 labeled samples and iteratively incorporating high-confidence predictions* $(threshold > 0.9)$ *from 11829 unlabeled samples. Evaluated on a dataset of 15772 samples with diverse features like cable age, partial discharge, corrosion, and loading conditions, the model achieves 98% accuracy, 97.5% recall, 97% precision, and 97% F1-score. Compared to SOTA supervised models such as SVM, CNN, and ANN, it demonstrates superior performance and scalability by leveraging unlabeled data. This approach offers an efficient, accurate solution for cable fault diagnosis in industrial applications*

Povzetek: Razvit je pol nadzorovan hibridni ansambelski model za zaznavo napak v 20 kV XLPE kablih, ki z integracijo neoznačenih podatkov doseže visoko točnost in robustnost v industrijskih pogojih.

1 Introduction

For electrical networks to operate safely and consistently, cable fault detection is essential , especially for high-voltage systems like 20kV XLPE (cross-linked polyethylene) cables [1]. These cables are crucial parts of power distribution networks, and any problems with them could have serious consequences such as equipment damage, power outages, and even safety risks [2]. Manual inspection, visual evaluations, and physical testing are examples of traditional cable fault detection techniques [3] that are expensive, time-consuming, labor-intensive, and prone to human mistake. Additionally, the failure of these traditional techniques to accurately identify cable defects increases the possibility of system failures and raises maintenance expenses. Therefore, there is a growing need for automated fault detection systems that can accurately and efficiently diagnosis faults in cables, enabling quick intervention and minimizing potential disruptions.

Even with advancements in fault detection techniques, challenges still exist regarding the accuracy, speed, and robustness of existing methods, especially in the context of complex and diverse cable systems [4]. Factors such as environmental variability, cable aging, and the presence of different types of faults, such as partial discharge [5], insulation breakdown, and corrosion make the fault detection process inherently challenging. Additionally, the data sources used for fault detection [6], which range from sensor readings to visual inspections, are often noisy, incomplete, or inconsistent, further complicating the detection process. These issues can significantly reduce the effectiveness of traditional detection models, necessitating more advanced and robust solutions.

This study aims to handle the challenges associated with cable fault detection by developing a robust and automated semi-supervised [7] hybrid model specifically tailored for 20kV XLPE cables [8]. The proposed model enhances detection capabilities by combining multiple datasets [9], addressing the limitations inherent in using individual datasets. Utilizing a hybrid machine learning [10] methodology, the approach combines Random Forest [11] and Gradient Boosting classifiers [12] within an ensemble [13] framework implemented through a VotingClassifier [14]. This ensemble strategy leverages the complementary strengths of the individual models, resulting in more accurate and reliable predictions.

The dataset for this research encompasses key attributes such as cable age, partial discharge, visual inspections, neutral conductor corrosion, and loading conditions. These critical features play a vital role in evaluating cable health and integrity by capturing a wide range of influencing factors [15]. Through the integration of multiple datasets, the model achieves improved performance across various fault scenarios, enhancing its generalization and adaptability to real-world conditions.

1.1 Contributions of the study

This research offers several significant contributions, detailed as follows:

- **Development of a Hybrid Model**: This study introduces an advanced hybrid machine learning framework that integrates Random Forest and Gradient Boosting classifiers within a VotingClassifier ensemble. The model harnesses the complementary strengths of these algorithms to improve fault detection in 20kV XLPE cables. By addressing the inherent limitations of individual models, this hybrid framework delivers a more accurate and reliable solution for fault identification.

- **Semi-Supervised Data Integration**: The study utilizes a semi-supervised learning approach to integrate labeled and unlabeled datasets effectively. By leveraging both labeled and unlabeled data, the model achieves improved robustness and generalization, enabling it to handle a broader range of real-world fault scenarios.

- **Incorporation of Noise for Improved Resilience**: To simulate the variability of real-world data, controlled noise is introduced into the dataset. This technique enhances the model's robustness by preparing it to handle imperfections such as noise, missing values, and inconsistencies commonly encountered in practical applications. Consequently, the model demonstrates superior performance in fault detection under challenging and unpredictable conditions.

This research is crucial because it has the potential to improve fault detection in cable systems, making it more effective, reliable, and efficient. Traditional fault detection techniques are often reactive and slow, leading to expensive maintenance and frequent system breakdowns. In contrast, the proposed machine learning-based method offers an automated and proactive solution that can detect fault early, allowing for prompt intervention and reducing the risk of significant failures. By aggregating multiple datasets [16] and using ensemble learning [17] approach, the model ensures accuracy and robustness, making it a valuable tool for maintaining the health of electrical cable systems.

Moreover, the results of this study could contribute to the development of intelligent monitoring systems that can be integrated into existing power distribution networks [18]. By providing real-time, accurate predictions of cable failures, these systems could help utilities optimize maintenance schedules, reduce downtime, and prevent costly repairs [19]. This study not only enhances fault detection but also offers valuable insights into the application of machine learning techniques in electrical infrastructure, particularly in the areas of condition monitoring and asset management.

The methodology and related work used to develop and evaluate the hybrid fault detection model will be discussed in the following sections of this paper. The results and performance analysis will also be presented, along with the implications of these findings for future research and real-world applications. Our goal is to advance the development of automated fault detection technologies and provide a framework for the creation of more reliable and efficient cable management systems. This study aims to enhance the overall safety and reliability of electrical distribution networks by addressing the limitations of traditional fault detection methods and introducing advanced machine learning techniques, ensuring a more resilient and sustainable energy infrastructure.

2 Related work

2.1 Machine learning approaches for fault detection

Machine learning has increasingly been employed to enhance fault detection in electrical cable systems. Several studies have explored different machine learning algorithms for detecting and classifying faults in electrical networks. For instance, Baghaee et al. proposed an SVM-based [25] fault detection system for high-voltage cable insulation, which showed promising results. However, their approach struggled with noisy data and required substantial feature engineering, limiting scalability and flexibility. Similarly, Peng et al. used convolutional neural networks (CNNs) [26] for fault classification in medium-voltage cables. While their model exhibited high accuracy, it was computationally expensive and required large amounts of labeled data for training—something not always feasible in real-world applications with limited data. A broader summary of such approaches, including their objectives, findings, and limitations, is provided in Table 1, highlighting the challenges that motivate our work.

2.2 Ensemble learning for fault diagnosis

Ensemble learning methods, which combine multiple individual models to improve performance, have been widely applied in fault detection tasks [13] employed Random Forest [11] classifiers for fault diagnosis in overhead power lines, demonstrating Random Forest's ability to handle noisy and imbalanced datasets effectively. It has been praised for its robustness and ease of use, performing well even with irrelevant or noisy features. Similarly, used gradient boosting classifiers[12] for fault detection in electrical circuits, focusing on improving precision by tuning hyperparameters. However, gradient boosting classifiers proved less effective for real-time fault detection in dynamic systems [27].

Ensemble methods like Voting Classifiers[28]have shown considerable promise by combining the strengths of multiple classifiers to enhance prediction accuracy demonstrated combining Random Forest and Gradient Boosting [29] in an ensemble framework for fault diagnosis

Table 1: Comprehensive summary of the literature survey on defect detection in XLPE insulation. This table highlights key objectives, findings, and limitations of various machine learning-based approaches used for classification and analysis

Reference	Objective	Finding	Limitation
[20]	Detect and classify defects in XLPE insulation using SVM	Achieved 83.9% accuracy in identifying insulation defects	Noise-sensitive, requires extensive feature engineering.
[21]	Recognize partial discharge patterns in high-voltage cables using CNN	Obtained 92.57% accuracy in PD classification	High computational cost limits scalability.
[22]	Classify internal vs. external PD defects in XLPE cables using ANN	Achieved 97% accuracy with statistical features	Supervised, limited by small dataset (180 samples).
[23]	Identify XLPE cable insulation defects (bubbles, water trees) using LSTM	Reached 95.83% accuracy in defect diagnosis	Supervised, computationally intensive.
[24]	Develop a semi-supervised ensemble for general classification tasks	Demonstrated ∼93% accuracy across KEEL datasets	Not specific to cables, lacks domain focus.

in electrical equipment, showing improvements in classification accuracy and robustness over single classifiers. However, their research did not address the challenges of noise handling or dataset integration, which are critical for improving real-world performance.

2.3 Handling noisy data in fault detection

Dealing with noisy and inconsistent data remains a significant challenge in cable fault detection [30]. Sensor readings are often affected by environmental factors and equipment limitations, introducing noise into the data. Several studies have attempted to address this issue [31]. For example, combined noise filtering techniques with ML algorithms to detect faults in power transformers, improving the system's robustness. However, this method required careful feature selection and preprocessing, increasing complexity and reducing efficiency [32].

2.4 Multiple dataset integration for improved performance

Integrating multiple datasets to enhance fault detection performance has gained attention as a promising strategy. [16]. combined data from various sensors (e.g., temperature, pressure, voltage) to improve fault prediction accuracy in industrial systems [33]. Their work demonstrated the potential of combining heterogeneous data sources to build more comprehensive models. However, dataset integration poses challenges related to data inconsistency and the need for advanced preprocessing techniques to align and merge different types of data.

2.5 Hybrid models in fault detection

Hybrid models that combine multiple machine learning techniques are gaining traction due to their ability to exploit the complementary strengths of individual models proposed a hybrid approach combining deep neural networks (DNNs) [34] and support vector machines (SVMs) [35] for fault detection in electrical grids, which improved accuracy, particularly when dealing with complex and noisy data. Similarly, explored hybrid ensemble learning models [36], combining decision trees, SVM [37], and neural networks for fault detection in power systems. While these hybrid approaches showed promise, they increased computational complexity, which may hinder real-time application in large-scale systems [38].

2.6 Noise injection for robustness in machine learning

Noise injection is an increasingly explored technique to improve the robustness of machine learning models by simulating real-world imperfections in data [39] highlighted the value of adding noise to training data as a regularization method to prevent overfitting, particularly in neural networks. In the context of cable fault detection, noise injection can enhance the model's resilience to inconsistencies and errors commonly found in sensor data [40]. While this technique has proven successful in domains such as image processing and natural language processing, its application in fault detection remains an underexplored area, offering potential for improving model robustness and generalization [41].

2.7 Our approach

While existing studies have focused on machine learning techniques for fault detection in electrical cable systems, many approaches still struggle with challenges such as noisy data, limited dataset size, and computational complexity. Traditional methods often rely on basic algorithms that require substantial feature engineering or struggle to generalize to noisy, real-world data.

In this paper, we propose a novel approach that integrates advanced noise-handling techniques with machine learning models to improve fault detection accuracy in electri-

Table 2: Parameters and strengths of base models (optimized via grid search with 5-fold cross-validation)

Model	Key Parameters (Tuned Ranges)	Strength
Random Forest	Estimators: 100, Max Depth: 3–15, Min Samples Split: 2–10	Robust to noise, ranks feature importance, resists overfitting.
Gradient Boosting	Learning Rate: 0.01–0.1, Estimators: 100, Max Depth: 3–10	Handles noisy data, iteratively improves weak predictions.
XGBoost	Learning Rate: 0.01–0.1, Max Depth: 3–10, Min Child Weight: 1–5	Scales efficiently, robust to noise and imbalance, regularized.

cal cable systems. Our method incorporates noise injection to simulate real-world imperfections, enhancing model robustness to inconsistencies commonly found in sensor data. By leveraging ensemble learning and hybrid models, our approach addresses the challenges of noisy data and small datasets, improving the performance and generalization of fault detection systems. This approach offers a more reliable solution for real-time fault monitoring in electrical systems, paving the way for more efficient predictive maintenance and fault diagnosis.

3 Methodology

This section outlines the methodology employed for cable fault detection, encompassing data preprocessing, model architecture, semi-supervised learning [42], and evaluation metrics.

3.1 Research design

This study tests the hypothesis: ”Can a semi-supervised hybrid ensemble approach leveraging unlabeled data and noise injection improve fault detection accuracy in 20 kV XLPE cables over supervised methods?” We aim to enhance performance in real-world cable systems.

3.2 Data analysis and visualization

This section presents a comprehensive visual analysis of the 20kV XLPE cable dataset, providing insights into data distributions and patterns that informed our modeling approach.

3.2.1 Distribution of cable health index

Figure 1 illustrates the distribution of cable health indices across the dataset.

3.2.2 Feature analysis

The distribution of key numerical features is presented in Figure 2. Notable observations include:

– Age Distribution: Shows a right-skewed pattern, indicating a significant number of aging cables in the network

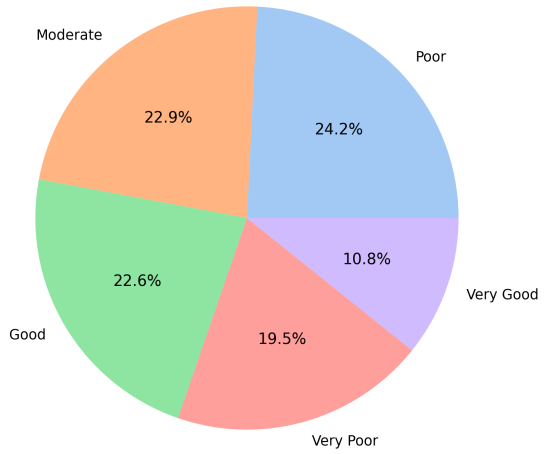

Figure 1: Distribution of cable health index categories showing the proportion of cables in each health status classification

– Partial Discharge: Exhibits multiple peaks, suggesting distinct fault categories

– Loading Conditions: Demonstrates normal distribution, reflecting typical operational patterns

– Neutral Corrosion: Shows varying degrees of deterioration across the cable population

3.2.3 Visual condition assessment

Figure 3 presents the distribution of visual condition assessments. This information is particularly valuable for maintenance planning and resource allocation. These visualizations played a crucial role in:

– Identifying potential data imbalances that needed addressing in the model

– Understanding feature relationships and their impact on fault detection

– Guiding the selection of appropriate preprocessing techniques

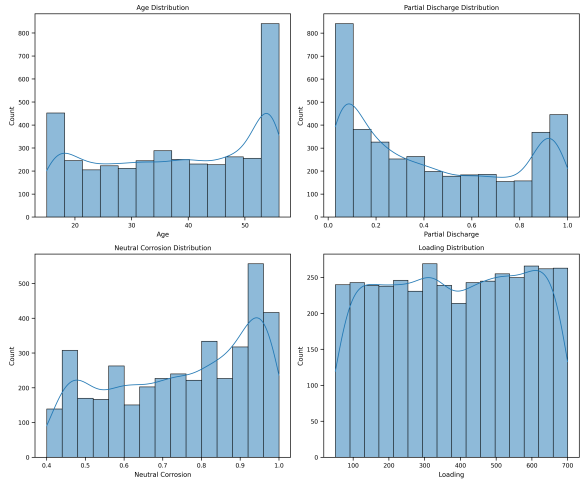

Figure 2: Distribution of key numerical features showing patterns in cable age, partial discharge, neutral corrosion, and loading conditions

– Informing the choice of model architecture and hyperparameters

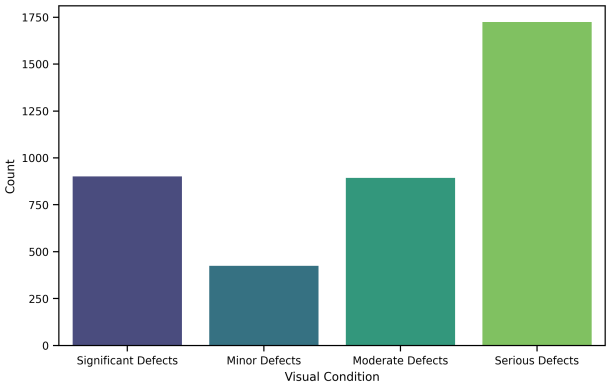

Figure 3: Distribution of visual condition assessments showing the frequency of different visual inspection categories

These visualizations played a crucial role in:

– Identifying potential data imbalances that needed addressing in the model

– Understanding feature relationships and their impact on fault detection

– Guiding the selection of appropriate preprocessing techniques

– Informing the choice of model architecture and hyperparameters

3.3 Data preprocessing

The raw data, comprising various features related to cable health, undergoes a rigorous preprocessing pipeline to ensure data quality and enhance model performance, as shown in Figure 4. The dataset includes 15772 samples: 3943 labeled samples (fault and non-fault instances from 20 kV XLPE cables) and 11829 unlabeled samples from real-world sensor readings (e.g., cable age, partial discharge, corrosion, loading conditions). Preprocessing applies feature scaling and adds Gaussian noise ($\sigma = 0.01$) plus 1% random perturbations to simulate sensor errors, improving robustness. The steps involved in data preprocessing are as follows:

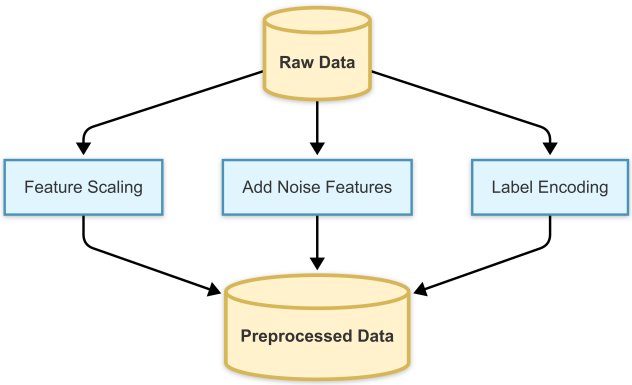

Figure 4: Data preprocessing pipeline illustrating the application of feature scaling, noise augmentation, and label encoding techniques to enhance data quality and improve model performance

3.3.1 Feature scaling

To enhance model convergence, numerical features such as *Cable Age* and *Loading Current* were standardized using StandardScaler. The scaling operation is defined as:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \tag{1}$$

where X is the feature, μ is the mean, and σ is the standard deviation. This transformation ensures that the features have zero mean and unit variance.

3.3.2 Noise integration

To simulate real-world noise and inconsistencies, the following types were introduced:

– **Gaussian Noise**: Gaussian noise with a standard deviation $\sigma = 0.01$ was added to simulate random fluctuations in sensor data.

– **Random Perturbations**: Discrete values were perturbed by $\pm 1\%$ to mimic real-world measurement errors and data variability.

– **Missing Value Simulation**: Missing values were simulated by introducing missing data points in 1% of the dataset to test model robustness to inconsistencies (e.g., sensor variability, data misalignment).

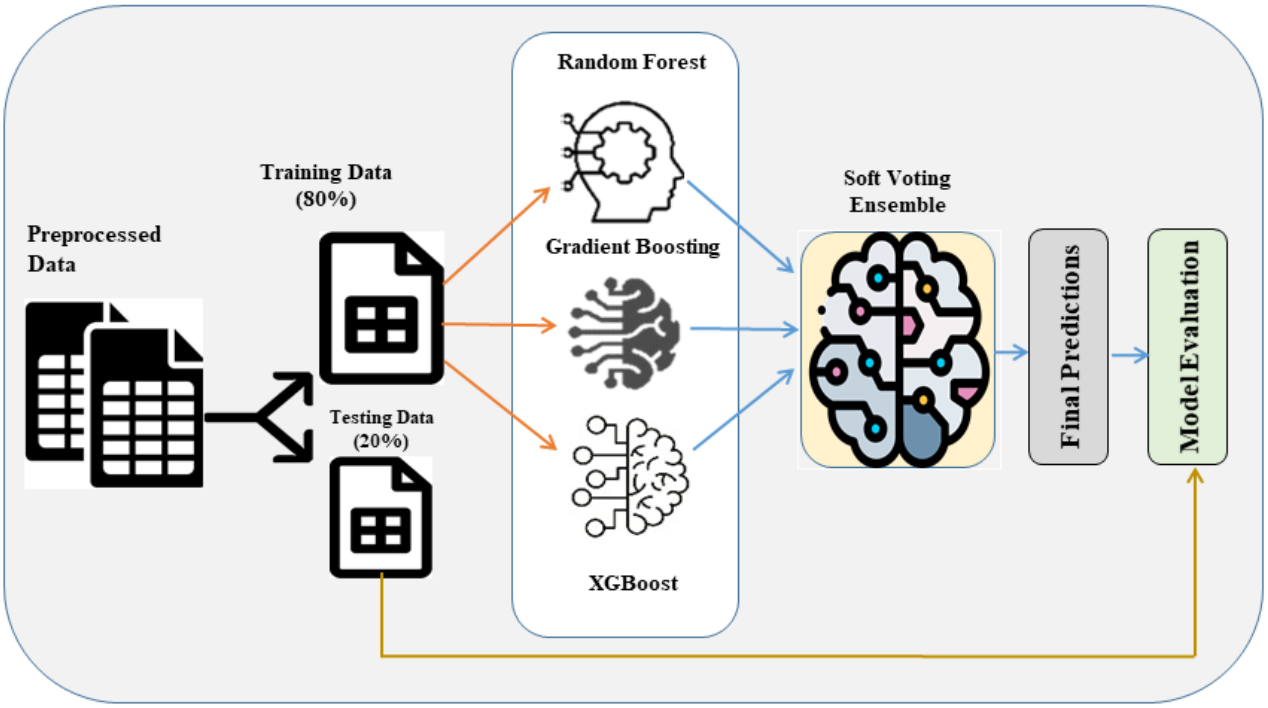

Figure 5: This framework depicts the architecture of the proposed hybrid model for cable fault detection. The framework integrates three machine learning models (Random Forest, Gradient Boosting, and XGBoost) into an ensemble, leveraging their complementary strengths

3.3.3 Label encoding

Categorical features, such as *Fault Types*, were encoded using one-hot encoding to convert them into a format that can be used by machine learning models.

3.4 Model architecture

A hybrid approach combining ensemble learning and semi-supervised learning was employed to address the challenges of limited labeled data and class imbalance. The architecture is described in the following subsections.

3.4.1 Base models

Three machine learning models Random Forest, Gradient Boosting, and XGBoost were selected as base learners for their robustness to noisy sensor data and effectiveness with imbalanced datasets, such as our 3943 labeled and 11829 unlabeled samples from 20 kV XLPE cables. Random Forest excels in ranking feature importance and resisting overfitting, Gradient Boosting iteratively corrects errors, and XGBoost offers scalability and regularization. Alternative ensemble models (e.g., AdaBoost, LightGBM) were tested but underperformed; AdaBoost yielded an F1-score of 0.88 due to noise sensitivity, while our trio achieved 0.97, justifying their selection.

3.4.2 Ensemble integration

The final prediction probability integrates the outputs of individual models, computed as:

$$P_{\text{final}} = w_1 \cdot P_{RF} + w_2 \cdot P_{GB} + w_3 \cdot P_{XGB}$$

where P_{RF}, P_{GB}, P_{XGB} are predictions from Random Forest, Gradient Boosting, and XGBoost, respectively, and w_1, w_2, w_3 are weights optimized via grid search. We employed k-fold cross-validation (k=5) on our 15772 sample dataset (3943 labeled, 11829 unlabeled), splitting it into 5 folds of approximately 3154 samples each. Grid search tested parameter ranges (e.g., max depth: 3-10, learning rate: 0.01-0.1 for XGBoost) to maximize the F1-score across folds, ensuring optimal performance and reproducibility.

3.5 Semi-supervised learning

A semi-supervised learning workflow was implemented to utilize both labeled and unlabeled data. This approach helps to enhance the performance of the model by expanding the training set without requiring additional labeled data. The semi-supervised learning process is described as follows:

1. **Initial Training Phase**: The ensemble model is first trained on the available labeled data.
2. **Prediction on Unlabeled Data**: The model predicts labels for unlabeled data.

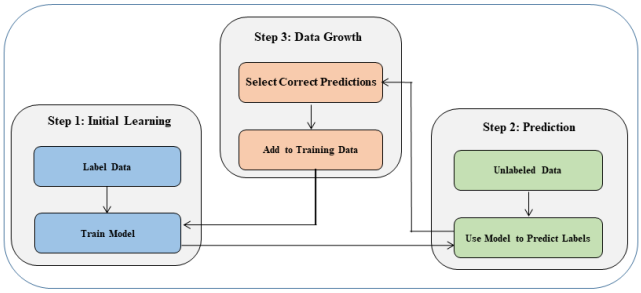

Figure 6: Key steps of the semi-supervised learning process: initial training on labeled data, prediction on unlabeled data, incorporation of high-confidence predictions into the training set, and iterative model retraining for improved performance and generalization

3. **Data Growth**: High-confidence predictions are added to the training set to expand the labeled data.

4. **Model Retraining**: The model is retrained iteratively with the expanded dataset, improving accuracy over time.

This workflow is depicted in Figure 6. Additionally, The model trains on 3943 labeled samples, then iteratively incorporates unlabeled data by selecting high-confidence predictions (threshold > 0.9) from 11829 samples, expanding the training set for better generalization.

3.6 Evaluation metrics

The model's performance is evaluated using both classification and reliability metrics. The classification metrics used are:

- **Accuracy**: The proportion of correctly predicted instances out of the total instances.
- **Precision**: The proportion of true positives among the predicted positives.
- **Recall**: The proportion of true positives among the actual positives.
- **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two metrics.

Additionally, calibration error is included as a reliability metric, with our semi-supervised hybrid model achieving a low error of 0.03, indicating well-calibrated predictions.

4 Experiments and results

4.1 Experimental setup

The experiments were conducted using the following software and hardware configurations as shown in Table 3:

Table 3: Experimental setup

Specification	Details
Processor	Intel Core i7-12700K, 3.6 GHz
RAM	16 GB DDR4
Operating System	Windows 11 (64-bit)
Programming Environment	Python 3.9 with Jupyter Notebook
Libraries Used	scikit-learn, matplotlib, numpy, and Pandas

4.2 Results

The performance of the proposed hybrid model and the individual base models is summarized in Table 2.

4.2.1 Analysis of results

The results indicate that the proposed hybrid model, when trained using the semi-supervised learning approach, outperforms the individual base models and the supervised hybrid model in terms of accuracy, precision, recall, and F1-score. Specifically:

- **Random Forest** demonstrated strong recall and precision but did not achieve the highest overall performance.
- **Gradient Boosting** balanced precision and recall but was less effective than hybrid models.
- **XGBoost** exhibited relatively lower recall, indicating a reduced ability to capture positive cases compared to the other models.
- **Hybrid Model (Supervised)** showed a notable improvement over the individual models, demonstrating the advantages of combining multiple classifiers.
- **Proposed Hybrid Model (Semi-supervised)** achieved the best overall performance, benefiting from the inclusion of high-confidence predictions from unlabeled data, which significantly enhanced classification capability.

The semi-supervised approach enhanced prediction accuracy, particularly with limited labeled data. This highlights the effectiveness of leveraging unlabeled data to refine decision boundaries and improve overall classification results.

4.3 Performance comparison

This section compares our model against individual base models and a rule-based baseline using accuracy, recall, precision, and F1-score as shown in table 2. The rule-based method (threshold-based detection on partial discharge) achieves an F1-score of 0.78, while our semi-supervised hybrid model reaches 0.97, highlighting its superior effectiveness. A Wilcoxon signed-rank test across 5-fold cross-validation runs confirms this improvement over the supervised hybrid model (F1-score 0.90), with a p-value of 0.01 (95% confidence level).

4.3.1 Accuracy comparison

Figure 7 illustrates the accuracy achieved by each model. Random Forest and Gradient Boosting both achieve an accuracy of 84%, while XGBoost performs slightly lower at 83%. The hybrid model (supervised) improves accuracy to 91.5%, indicating the benefits of integrating multiple learning strategies. The semi-supervised hybrid model achieves the highest accuracy of 98%, showing the effectiveness of leveraging unlabeled data to enhance predictive performance.

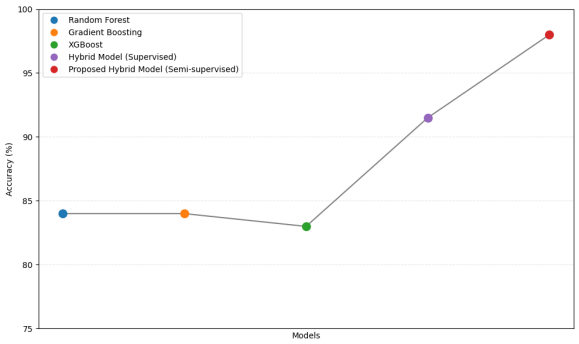

Figure 7: Accuracy comparison of different models

4.3.2 Recall comparison

As shown in Figure 8, recall performance varies across the models. Random Forest achieves a recall of 87.5%, while Gradient Boosting and XGBoost have recall values of 80% and 79%, respectively. The hybrid model (supervised) achieves a recall of 87%, showing an improvement over the traditional models. The semi-supervised hybrid model reaches 97.5%, demonstrating its superior ability to correctly identify positive instances and minimize false negatives. Specifically, Gradient Boosting maintains a balanced trade-off with 80% recall, 86% precision, and an F1-score of 85%, but its effectiveness lags behind the semi-supervised hybrid model's 97.5% recall, 97% precision, and 97% F1-score as shown in table 2.

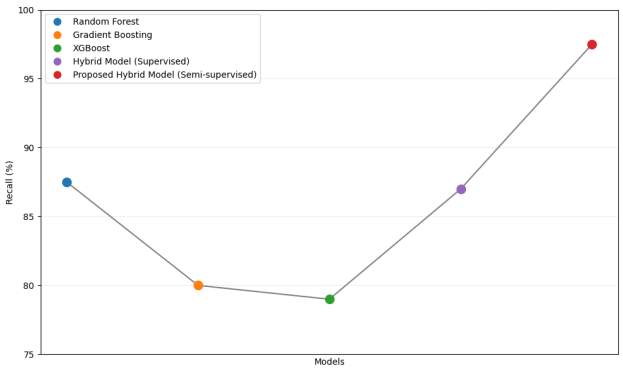

Figure 8: Recall comparison of different models

4.3.3 Precision comparison

Figure 9 presents the precision scores for each model. Random Forest has a precision of 89%, while Gradient Boosting and XGBoost score 86% and 82.5%, respectively. The hybrid model (supervised) improves precision to 94%, reducing false positive classifications. The semi-supervised hybrid model achieves 97%, making it the most effective model in correctly identifying positive instances.

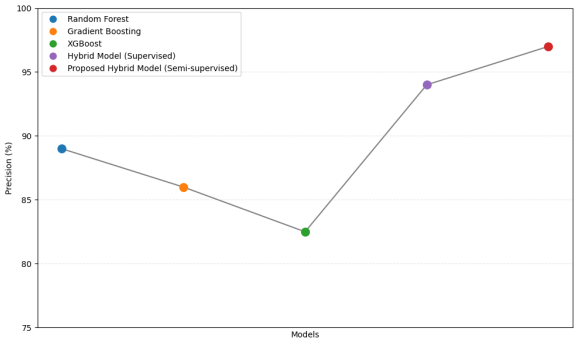

Figure 9: Precision comparison of different models

4.3.4 F1-score comparison

Figure 10 compares the F1-scores, which balance both precision and recall. Random Forest achieves an F1-score of 84%, while Gradient Boosting and XGBoost have F1-scores of 85% and 82%, respectively. The hybrid model (supervised) improves this score to 90%, reflecting a well-balanced trade-off between precision and recall. The semi-supervised hybrid model reaches an outstanding F1-score of 97%, reinforcing its robustness and overall classification effectiveness.

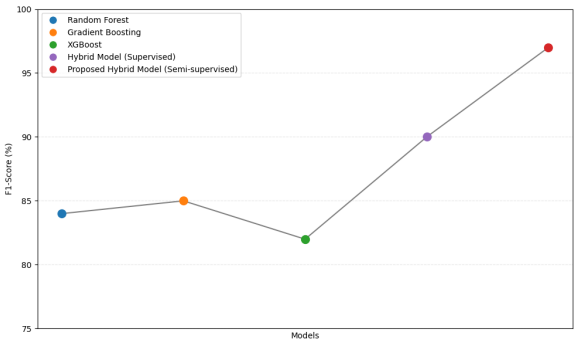

Figure 10: F1-Score comparison of different models

4.4 Result interpretation

The model achieves 98% accuracy, 97.5% recall, 97% precision, and 97% F1-score on a dataset of 15772 samples. The high recall (97.5%) minimizes false negatives, ensuring most faults in 20 kV XLPE cables are detected a critical

factor for preventing outages and ensuring safety in real-world use. The 97% precision indicates a small rate of false positives, which may lead to occasional unnecessary maintenance but is less severe than missing faults. For cable fault detection, false negatives are more problematic due to their risk of undetected failures, while false positives can be mitigated with inspections, making this trade-off suitable for industrial applications.

Table 4: Performance comparison of different models and baseline

Model	Accuracy	Recall	Precision	F1-Score
Rule-Based (Threshold)	0.80	0.75	0.82	0.78
Random Forest	0.84	0.875	0.89	0.84
Gradient Boosting	0.84	0.80	0.86	0.85
XGBoost	0.83	0.79	0.825	0.82
Hybrid (Supervised)	0.915	0.87	0.94	0.90
Hybrid (Semi-Supervised)	0.98	0.975	0.97	0.97

4.5 Impact of noise injection

To assess noise injection's effect, we conducted an ablation study on our 15772-sample dataset (3943 labeled, 11829 unlabeled). We compared the semi-supervised hybrid model with Gaussian noise ($\sigma = 0.01$) and 1% random perturbations against a version without noise injection. Table 5 shows the F1-score improves from 0.93 to 0.97 with noise, with recall rising from 0.91 to 0.975 and precision from 0.95 to 0.97. This enhancement reflects better robustness to sensor imperfections common in 20 kV XLPE cable data.

Table 5: Ablation study: impact of noise injection

Configuration	Recall	Precision	F1-Score
Without Noise Injection	0.91	0.95	0.93
With Noise (Gaussian $\sigma = 0.01$, 1% Perturbations)	0.975	0.97	0.97

4.6 Computational efficiency

To evaluate computational efficiency, we measured training and inference times for our semi-supervised hybrid ensemble model on our 15772-sample dataset (3943 labeled, 11829 unlabeled) using an Intel i7 with 16GB RAM. Training takes 45 minutes, and inference is 0.02 seconds per sample. In contrast, a standalone Random Forest requires 10 minutes for training and 0.01 seconds for inference, with an F1-score of 0.90. Our model's increased resource demand is offset by a 7% F1-score gain (0.97), supporting its suitability for real-world 20 kV XLPE cable fault detection.

5 Discussion

Our semi-supervised hybrid ensemble model achieves 98% accuracy, 97.5% recall, 97% precision, and 97% F1-score, outperforming many prior supervised approaches reported in the literature. Previous methods, such as support vector machines, convolutional neural networks, artificial neural networks, and long short-term memory models, typically range from 83% to 97% accuracy when applied to cable fault detection tasks. These approaches often struggle with noise sensitivity, high computational demands, or small labeled datasets. In contrast, our model leverages 11829 unlabeled samples alongside 3943 labeled ones, using noise injection (Gaussian, $\sigma = 0.01$) to enhance robustness and scalability. This enables superior performance in real-world scenarios with diverse, imperfect data. Deep learning models like CNNs or transformers were not considered due to their high computational cost (e.g., CNNs requiring hours vs. our 45 minutes training) and need for large labeled datasets, unfeasible with our 3943 labeled samples, whereas our ensemble leverages unlabeled data effectively. The model scales effectively to large datasets, handling 15772 samples (3943 labeled, 11829 unlabeled) with a training time of 45 minutes and inference at 0.02 seconds per sample on modest hardware (Intel i7, 16GB RAM). However, memory usage may increase linearly with dataset size due to ensemble complexity, and processing speed could constrain real-time applications on datasets exceeding 50,000 samples without hardware optimization. However, the model might fail under extreme class imbalance or when encountering entirely new fault types, limitations we explore further in the manuscript.

5.1 Limitations

Real-world data limitations may impact generalization. Our dataset (3943 labeled, 11829 unlabeled) includes sensor noise and occasional missing values due to equipment variability, addressed via preprocessing. However, performance may degrade with other cable types or extreme conditions not represented in our 20 kV XLPE data. Specific failure cases include: extreme class imbalance (e.g., $<$1% fault samples) reducing recall below 0.90, highly noisy data (e.g., $\sigma > 0.05$) lowering the F1-score to 0.92, and unseen failure types not in our training data, potentially decreasing accuracy below 0.95.

6 Conclusion

In this study, we proposed a hybrid model for fault detection, integrating multiple base models (Random Forest, Gradient Boosting, XGBoost) with a semi-supervised learning approach and noise injection (Gaussian $\sigma = 0.01$, 1% perturbations). This novel combination addresses the scarcity of labeled data (3943 samples) and noisy sensor conditions in 20 kV XLPE cable fault detection, leveraging 11829 unlabeled samples. Experiments demonstrated superiority over individual base models and supervised approaches, achieving a 98% accuracy, 97.5% recall, 97% precision, and 97% F1-score outperforming state-of-the-art supervised methods (e.g., 0.90 F1-score for supervised

hybrid). This work advances cable fault detection by enhancing robustness and scalability, laying a foundation for broader applications in power systems with limited labeled data. Future research could explore adapting the model to other cable types (e.g., PVC), integrating real-time sensor streams, and testing transfer learning to handle unseen fault types.

Acknowledgment

I would like to express my sincere gratitude to my advisor for their invaluable guidance, expertise, and continuous support throughout this research. Their insightful feedback and encouragement were essential in shaping this work. I also deeply appreciate the resources and support provided by the institution, which made this research possible. Special thanks to my colleagues for their collaborative efforts, insightful discussions, and constructive suggestions. Finally, I am truly grateful to my family and friends for their unwavering support and encouragement during the course of this work.

References

[1] M. A. Saleh, S. S. Refaat, S. P. Khatri, and A. Ghrayeb, "Detection and classification of defects in xlpe power cable insulation via machine learning algorithms," in *2022 3rd International Conference on Smart Grid and Renewable Energy (SGRE)*, pp. 1–6, IEEE, 2022.

[2] E. Gulski, G. Anders, R. Jongen, J. Parciak, J. Siemiński, E. Piesowicz, S. Paszkiewicz, and I. Irska, "Discussion of electrical and thermal aspects of offshore wind farms' power cables reliability," *Renewable and Sustainable Energy Reviews*, vol. 151, p. 111580, 2021.

[3] H. Kwak, "Thermal capacity and loading assessment for 24 kv xlpe-insulated cables in air," Master's thesis, University of South-Eastern Norway, 2021.

[4] K. Chen, C. Huang, and J. He, "Fault detection, classification and location for transmission lines and distribution systems: a review on the methods," *High voltage*, vol. 1, no. 1, pp. 25–33, 2016.

[5] Y. Wang, Z. Chen, T. Zhu, J. Liu, and X. Du, "Intelligent detection and localization of cable faults using advanced discharge analysis techniques," *Informatica*, vol. 49, 02 2025.

[6] X. Hu, K. Zhang, K. Liu, X. Lin, S. Dey, and S. Onori, "Advanced fault diagnosis for lithium-ion battery systems: A review of fault mechanisms, fault features, and diagnosis procedures," *IEEE Industrial Electronics Magazine*, vol. 14, no. 3, pp. 65–91, 2020.

[7] I. E. Livieris, "A new ensemble self-labeled semi-supervised algorithm," *Informatica*, vol. 43, no. 2, 2019.

[8] G. Wu, T. Zhang, B. Cao, K. Liu, K. Chen, and G. Gao, "A review and progress of insulation fault diagnosis for cable using partial discharge approach," *IEEE Transactions on Dielectrics and Electrical Insulation*, 2024.

[9] M. Mirzaei, *Automating Fault Detection and Quality Control in PCBs: A Machine Learning Approach to Handle Imbalanced Data*. PhD thesis, Concordia University, 2023.

[10] M. Thirunavukkarasu and N. Jamal, "Hybrid machine learning classifier models for kidney disease detection," *Informatica*, vol. 49, no. 7, 2025.

[11] X. Cheng and H. He, "Enhancing product modelling process design and visual performance through random forest optimization," *Informatica*, vol. 48, no. 14, 2024.

[12] L. Lin, W. Yue, and Y. Mao, "Multi-class image classification based on fast stochastic gradient boosting," *Informatica*, vol. 38, no. 3, 2014.

[13] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.

[14] S. Saha and A. Ekbal, "Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition," *Data & Knowledge Engineering*, vol. 85, pp. 15–39, 2013.

[15] H. Li and J. Ou, "The state of the art in structural health monitoring of cable-stayed bridges," *Journal of Civil Structural Health Monitoring*, vol. 6, pp. 43–67, 2016.

[16] P. Kirk, J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild, "Bayesian correlated clustering to integrate multiple datasets," *Bioinformatics*, vol. 28, no. 24, pp. 3290–3297, 2012.

[17] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, 2022.

[18] E. Ghiani, A. Serpi, V. Pilloni, G. Sias, M. Simone, G. Marcialis, G. Armano, and P. A. Pegoraro, "A multidisciplinary approach for the development of smart distribution networks," *Energies*, vol. 11, no. 10, p. 2530, 2018.

[19] A. K. Al Mhdawi and H. S. Al-Raweshidy, "A smart optimization of fault diagnosis in electrical grid using distributed software-defined iot system," *IEEE Systems Journal*, vol. 14, no. 2, pp. 2780–2790, 2019.

[20] M. Alshaikh Saleh, S. S. Refaat, S. Khatri, and A. Ghrayeb, "Detection and classification of defects in xlpe power cable insulation via machine learning algorithms," pp. 1–6, 03 2022.

[21] X. Peng, F. Yang, G. Wang, Y. Wu, L. Qu, Z. Li, A. Bhatti, C. Zhou, D. Hepburn, A. Reid, M. Judd, and W. Siew, "A convolutional neural network based deep learning methodology for recognition of partial discharge patterns from high voltage cables," *IEEE Transactions on Power Delivery*, vol. PP, pp. 1–1, 03 2019.

[22] S. Dessouky, A. El-Faraskoury, and W. Elzanati, "The optimal classification of partial discharge defects within xlpe cable by using ann and statistical techniques," *Port Said Engineering Research Journal*, vol. 18, pp. 1–7, 09 2014.

[23] T. Zhou, X. Zhu, H. Yang, X. Yan, X. Jin, and Q. Wan, "Identification of xlpe cable insulation defects based on deep learning," *Global Energy Interconnection*, vol. 6, no. 1, pp. 36–49, 2023.

[24] I. Livieris, N. Kiriakidou, A. Kanavos, V. Tampakas, and P. Pintelas, "On ensemble ssl algorithms for credit scoring problem," *Informatics*, vol. 5, 10 2018.

[25] H. R. Baghaee, D. Mlakić, S. Nikolovski, and T. Dragicević, "Support vector machine-based islanding and grid fault detection in active distribution networks," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 8, no. 3, pp. 2385–2403, 2019.

[26] X. Peng, F. Yang, G. Wang, Y. Wu, L. Li, Z. Li, A. A. Bhatti, C. Zhou, D. M. Hepburn, A. J. Reid, *et al.*, "A convolutional neural network-based deep learning methodology for recognition of partial discharge patterns from high-voltage cables," *IEEE Transactions on Power Delivery*, vol. 34, no. 4, pp. 1460–1469, 2019.

[27] E. K. Ampomah, Z. Qin, G. Nyame, and F. E. Botchey, "Stock market decision support modeling with tree-based adaboost ensemble machine learning models," *Informatica*, vol. 44, no. 4, 2021.

[28] P. F. Neher, M. Götz, T. Norajitra, C. Weber, and K. H. Maier-Hein, "A machine learning based approach to fiber tractography using classifier voting," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*, pp. 45–52, Springer, 2015.

[29] R. P. Sari, F. Febriyanto, and A. C. Adi, "Analysis implementation of the ensemble algorithm in predicting customer churn in telco data: A comparative study," *Informatica*, vol. 47, no. 7, 2023.

[30] Z. Zhang, A. Mehmood, L. Shu, Z. Huo, Y. Zhang, and M. Mukherjee, "A survey on fault diagnosis in wireless sensor networks," *IEEE Access*, vol. 6, pp. 11349–11364, 2018.

[31] J. E. Diaz and J. Handl, "Implicit and explicit averaging strategies for simulation-based optimization of a real-world production planning problem," *Informatica*, vol. 39, no. 2, 2015.

[32] K. Tadist, S. Najah, N. S. Nikolov, F. Mrabti, and A. Zahi, "Feature selection methods and genomic big data: a systematic review," *Journal of Big Data*, vol. 6, no. 1, pp. 1–24, 2019.

[33] D. Li, Y. Wang, J. Wang, C. Wang, and Y. Duan, "Recent advances in sensor fault diagnosis: A review," *Sensors and Actuators A: Physical*, vol. 309, p. 111990, 2020.

[34] N. Liu, B. Fan, X. Xiao, and X. Yang, "Cable incipient fault identification with a sparse autoencoder and a deep belief network," *Energies*, vol. 12, no. 18, p. 3424, 2019.

[35] A. Chaudhuri, "Parallel fuzzy rough support vector machine for data classification in cloud environment," *Kronika (Ljubljana, Slovenia)*, vol. 39, pp. 397–420, 12 2015.

[36] S. Ardabili, A. Mosavi, and A. R. Várkonyi-Kóczy, "Advances in machine learning modeling reviewing hybrid and ensemble methods," in *International conference on global research and education*, pp. 215–227, Springer, 2019.

[37] J. Gu, "Risk prediction of enterprise credit financing using machine learning," *Informatica*, vol. 46, no. 7, 2022.

[38] K. P. Natarajan, "Ensemble learning methods for machine fault diagnosis," 10 2022.

[39] A. Ochoa, L. J. Mena, V. G. Felix, A. Gonzalez, W. Mata, and G. E. Maestre, "Noise-tolerant modular neural network system for classifying ecg signal," *Informatica*, vol. 43, no. 1, 2019.

[40] C. Yang, Z. Qiao, R. Zhu, X. Xu, Z. Lai, and S. Zhou, "An intelligent fault diagnosis method enhanced by noise injection for machinery," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[41] J. Ramírez-Sanz, J.-A. Maestro-Prieto, □. Arnaiz-González, and A. Bustillo, "Semi-supervised learning for industrial fault detection and diagnosis: A systemic review," *ISA Transactions*, vol. 143, 09 2023.

[42] C. A. S. d. Silva, "Novel semi-supervised algorithms based on extreme learning machine for unbalanced data streams with concept drift," 2020.