Inverter-Dependent Transmission System Protection Using a Joint Random Forest Classifier

Yue Chen
Power Systems Engineering Center
National Renewable Energy Laboratory
Golden, Colorado, USA
yue.chen@nrel.gov

Soham Chakraborty
Power Systems Engineering Center
National Renewable Energy Laboratory
Golden, Colorado, USA
soham.chakraborty@nrel.gov

Jing Wang
Power Systems Engineering Center
National Renewable Energy Laboratory
Golden, Colorado, USA
jing.wang@nrel.gov

***Abstract*—This paper presents a data-driven transmission system protection scheme based on a random forest classifier. Using voltage and current phasors and their sequence components, the classifier jointly outputs the fault status (fault/non-fault), fault type, and fault location. A moving-window architecture and a fault confirmation logic were employed for real-time implementation and provide a tunable trade-off between data efficiency, detection speed, and security. The evaluation on 871 PSCAD simulated fault cases shows 100% accuracy for fault detection, 97% accuracy for fault type identification, and 94.9% accuracy for fault location classification, with the decision time less than 2.5 cycles.**

***Index Terms*—Data-driven protection, inverter-based resources, random forest, multitask fault classification.**

I. Introduction

Traditional transmission line protection schemes (time-overcurrent, distance, and differential) are coordinated under assumptions of well-defined fault currents from synchronous generators and relatively stable source impedances, which enable reliable relay operations using current magnitudes, apparent impedance, and sequence components [1]; however, the penetration of inverter-based resources (IBRs)—such as utility-scale photovoltaics, wind, and battery energy storage—is fundamentally challenging these assumptions. IBRs exhibit fast dynamics and strictly limited fault currents because their grid interface is governed by vendor-specific power electronic control algorithms rather than synchronous machine physics [2]. As a result, the assumptions of high and predictable fault current and characteristic sequence behavior no longer hold in many IBR-rich systems. Numerous studies and reports have documented underreach, overreach, or failure of distance and overcurrent elements, and the misoperation of directional and ground fault functions, when a significant portion of the fault current is contributed by IBRs [3]–[5].

This work was authored by NREL for the U.S. Department of Energy (DOE), operated under Contract No. DE-AC36-08GO28308. Funding provided by the U.S Department of Energy Office of Energy Efficiency and Renewable Energy Solar Energy Technologies Office Agreement Number 41215. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

To address these issues, recent work has explored advanced protection concepts, such as time domain and traveling wave line elements [6]–[8], adaptive or self-tuning protection settings [9], and machine learning-based fault classification for IBR-connected transmission lines [10]–[12]. Machine learning techniques are attractive because they can learn complex, nonlinear relationships between measured signals and fault classes, potentially capturing subtle patterns that are not exploited by conventional protection algorithms.

Motivated by these developments, this paper proposes an inverter-dependent transmission system protection scheme based on a random forest classifier. Unlike most existing data-driven schemes that require separate models for each task (e.g., [13]), the proposed classifier simultaneously outputs the fault status (fault/non-fault), the fault location, and the fault type within a single unified model. To enable practical real-time deployment, we used a moving window of the most recent measurements for data acquisition and online analysis, and we applied a fault confirmation logic to consecutive decisions: A fault is declared only when consistent classifications are obtained over multiple detection steps. Together, the moving-window approach and confirmation logic provide a tunable trade-off between data efficiency, detection speed, and security.

The main contributions of this work are as follows:

- Transmission line protection in IBR-rich systems is formulated as a multitask classification problem, and a random forest-based algorithm is developed to produce the fault status, the fault type, and the fault location within a single *joint* classifier. This design avoids multiple task-specific models and shows that the joint model can reliably support all three protection decisions.
- An online implementation framework tailored for relay applications is proposed, combining moving-window data acquisition with a fault confirmation logic. This framework offers tunable knobs—including window length, decision time step, and confirmation requirement—to balance data efficiency, detection speed, and security.
- The effectiveness of the proposed scheme is demonstrated on test data from a PSCAD-modeled transmission system, achieving fast and accurate classification on the fault status, the fault type, and the fault location across diverse scenarios.

979-8-3315-8138-1

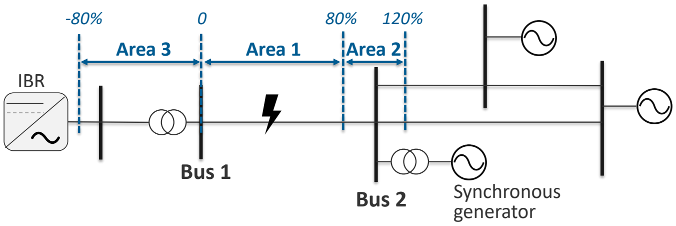

Fig. 1. Transmission system under study.

II. Power System and Feature Engineering

A. Power System Simulation

Fig. 1 shows the schematics of the transmission network we employed to demonstrate our approach. This is modified from a realistic transmission system, where we focus on the protection of the transmission line between Bus 1 and Bus 2. Bus 1 connects to a 14-MVA battery energy storage system, and Bus 2 connects to multiple synchronous generators.

To obtain data for this study, we conducted extensive simulations in PSCAD/EMTDC (with a 79-kHz time resolution) to examine diverse fault scenarios, as illustrated in Fig. 2:

- **Fault types:** single-phase-to-ground (AG, BG, CG), phase-to-phase (AB, BC, CA), double-phase-to-ground (ABG, BCG, CAG), and three-phase (ABC, ABCG)
- **Fault locations:** As shown in Fig. 1, Area 1 covers the forward line section from Bus 1 to 80% of the line (Bus 1–Bus 2), including eight fault locations in $(0, 80]$ in Fig. 2. Area 2 spans the remaining forward section from 80% to 120% of the line length (eight locations in $(80, 120]$ in Fig. 2). Area 3 represents the reverse direction from Bus 1, from -80% to 0 (8 locations in $[-80, 0)$ in Fig. 2).
- **Fault impedance:** fault impedance from 0 to 10 Ω, representing low- to high-impedance faults.

B. Data Acquisition

In this work, we consider distributed relay decision-making, where the relay in each bus processes its own measurements and makes decisions without the measurements of other relays. In this study, we focus on Bus 1, which connects with the IBRs and is the challenging part for conventional protection. At Bus 1, three-phase voltage and current magnitudes $v_{\mathrm{a}}(t)$, $v_{\mathrm{b}}(t)$, $v_{\mathrm{c}}(t)$, $i_{\mathrm{a}}(t)$, $i_{\mathrm{b}}(t)$, and $i_{\mathrm{c}}(t)$ are recorded at 9.9 kHz. In addition, we recorded the information for each fault case, including the fault time, fault type, and fault location.

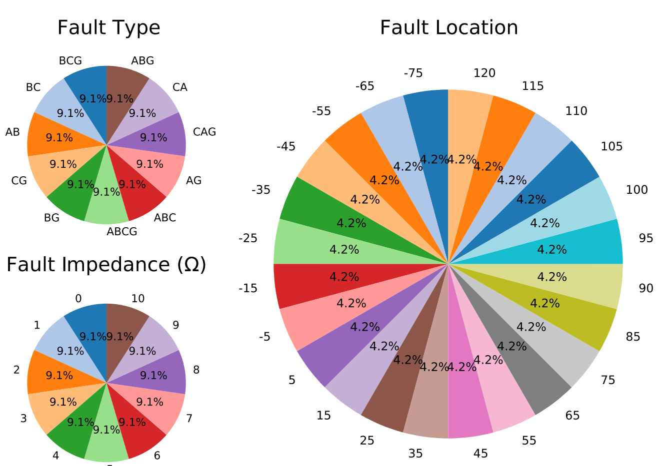

Fig. 2. Distribution of fault scenarios in the study.

Dataset size and split: In total, 2,904 distinct fault cases were simulated, covering 11 fault types, 24 fault locations, and 11 fault impedances. The data were partitioned 70:30 for training and testing.

C. Feature Extraction

For effective protection, the features used in data-driven algorithms must capture the electrical state of the system under both normal and faulted conditions. In this work, features including phasors and sequence components are extracted from three-phase voltages and currents, which provide a compact and physically meaningful representation of the system behavior.

1) Three-Phase Phasors: The instantaneous three-phase voltages $v_{\mathrm{a}}(t)$, $v_{\mathrm{b}}(t)$, $v_{\mathrm{c}}(t)$ and currents $i_{\mathrm{a}}(t)$, $i_{\mathrm{b}}(t)$, $i_{\mathrm{c}}(t)$, are represented as complex phasors. For a block of N samples, the phasors for phase $x \in \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$ are computed as:

$$V_{\mathrm{x}} = \frac{2}{N} \sum_{n=0}^{N-1} v_{\mathrm{x}}[n]\, e^{-j2\pi f_0 n / f_{\mathrm{s}}}, \; I_{\mathrm{x}} = \frac{2}{N} \sum_{n=0}^{N-1} i_{\mathrm{x}}[n]\, e^{-j2\pi f_0 n / f_{\mathrm{s}}} \tag{1}$$

where f_{s} and f_0 denote the sampling and fundamental frequencies, respectively.

2) Sequence Components: To capture system asymmetry and detect faults, the three-phase phasors are transformed into symmetrical components: zero sequence (V_0, I_0), positive sequence (V_1, I_1), and negative sequence (V_2, I_2), defined as:

$$V_0 = (V_{\mathrm{a}} + V_{\mathrm{b}} + V_{\mathrm{c}})/3, \; I_0 = (I_{\mathrm{a}} + I_{\mathrm{b}} + I_{\mathrm{c}})/3, \tag{2}$$

$$V_1 = (V_{\mathrm{a}} + aV_{\mathrm{b}} + a^2 V_{\mathrm{c}})/3, \; I_1 = (I_{\mathrm{a}} + aI_{\mathrm{b}} + a^2 I_{\mathrm{c}})/3, \tag{3}$$

$$V_2 = (V_{\mathrm{a}} + a^2 V_{\mathrm{b}} + aV_{\mathrm{c}})/3, \; I_2 = (I_{\mathrm{a}} + a^2 I_{\mathrm{b}} + aI_{\mathrm{c}})/3, \tag{4}$$

where $a = e^{j120^\circ}$ is the phase rotation operator. The magnitudes and angles of these components—i.e., $|V_0|$, $|V_1|$, $|V_2|$, $\angle V_0$, $\angle V_1$, $\angle V_2$ for voltages and $|I_0|$, $|I_1|$, $|I_2|$, $\angle I_0$, $\angle I_1$, $\angle I_2$ for currents—are employed as features for fault classification.

3) Relative Angle Representation: Instead of using the absolute angles defined here, relative angles are employed as features. Specifically, each phase's phasor angle is expressed relative to the phase-a voltage angle, and the sequence-component angles are referenced to the positive-sequence voltage component. This representation ensures that all angular measurements are expressed in a consistent reference frame, reduces sensitivity to measurement offsets, and enhances the robustness of the features for data-driven classification.

4) Justification: The use of phasors, sequence components, and relative angles is motivated by several considerations. Phasors condense high-sample-rate signals into a few representative complex values, enabling dimensionality reduction

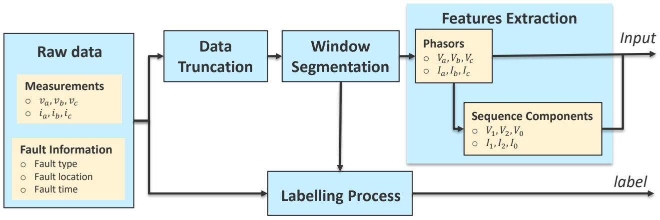

Fig. 3. Training data construction process.

without loss of key information. Sequence components effectively present unbalances and fault signatures in the system. Relative angles preserve the voltage-current and inter-phase relationships independent of reference choice, improving both physical interpretability and algorithm robustness.

III. Random Forest-Based Fault Classification

The section describes the design of the random forest-based protection scheme that simultaneously (i) discriminates between normal and faulted system conditions; and for faulted cases, (ii) identifies the fault type and (iii) determines the faulted area. Rather than employing separate classifiers for fault detection, fault type identification, and fault area classification, we adopt a joint multitask classifier that integrates all three decisions within a single model. This joint formulation reduces the overall model complexity and more effectively exploits the statistical dependencies among the three tasks.

A. Training Data

The raw data obtained from the PSCAD simulations consist of long time-series records that are not directly suitable for training. This section describes the construction of the training dataset, as summarized in Fig. 3.

For each simulated fault case, the dataset contains three-phase voltage and current measurements over the full simulation horizon. To improve efficiency, the data were truncated to a five-cycle interval around the fault, with two cycles before inception and three cycles after. Because the proposed scheme operates on a fixed-length moving window, the truncated data were segmented using the same window length, i.e., from each five-cycle data, one-cycle windows were randomly sampled to form the training dataset. This randomization enables the generation of multiple training samples from one fault case and enhances robustness to variations in the fault inception time and window alignment.

Following the data truncation and segmentation, feature extraction is performed to capture key fault response characteristics. As described in Section II-C, voltage and current phasors are computed, along with their corresponding positive-, negative-, and zero-sequence components.

Each window data sample is then assigned a class label to define the target outputs for supervised learning, thereby formulating the problem as a multiclass classification task. If the entire window lies before the fault inception, it is labeled as *non-fault*; otherwise, it is labeled according to the combination of fault type and faulted area. For example, the label *ABC_Area1* represents an ABC fault located in Area 1.

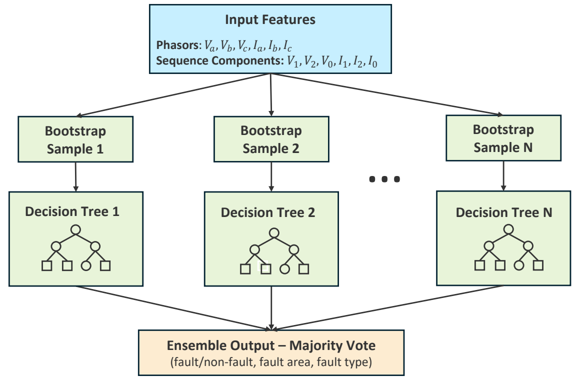

Fig. 4. Random forest algorithm for fault classification.

By pairing the extracted features with their corresponding class labels, the resulting dataset is prepared for training data-driven fault classification models.

B. Random Forest Classifier

The random forest classifier was employed for the fault classification tasks. As illustrated in Fig. 4, the algorithm creates multiple subsets of the training data using random sampling with replacement (called bootstrap samples). Each subset is used to train a separate decision tree. When building each tree, instead of evaluating all features at each split, it chooses a random subset of features. Among these features, the algorithm evaluates possible splits using a an impurity-based criterion (e.g., Gini impurity, entropy) and choose the best split. For the classification, each tree independently predicts a class label, and the final prediction is determined by a majority vote across all trees. This architecture enhances robustness to noisy and unbalanced data and mitigates overfitting relative to a single decision tree. See, e.g., [14] for further details on the random forest algorithm.

During training, key hyperparameters—including the number of trees, the maximum tree depth, the minimum number of samples required to split a node, the minimum number of samples per leaf, and the split criterion—were optimized via a grid search procedure. Each hyperparameter configuration was evaluated using stratified K-fold cross-validation, ensuring that each fold preserved the class distribution across fault types. The macro F1-score was adopted as the performance metric to account for class imbalance and to provide a balanced assessment across all fault categories. The random forest model achieving the highest cross-validated macro F1-score was selected as the final classifier.

This training strategy enables the random forest to capture nonlinear relationships among the phasor- and sequence-component features while maintaining good generalization to unseen operating conditions. The combination of ensemble learning, hyperparameter optimization, and cross-validation ensures a robust and reliable fault classification framework suitable for transmission system protection applications.

C. Online Testing

The online classification procedure is illustrated in Fig. 5 and comprises the following main components.

979-8-3315-8138-1

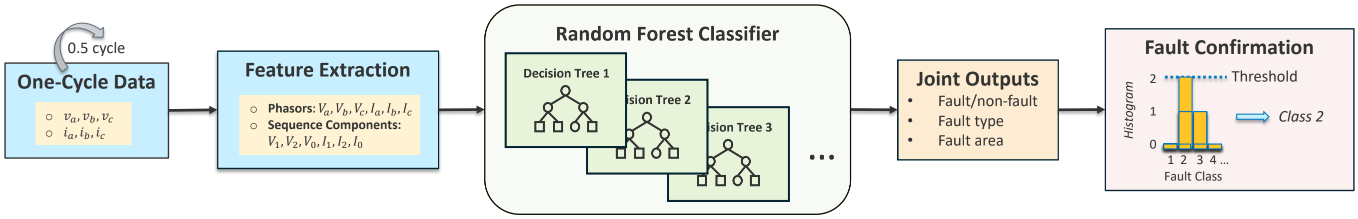

Fig. 5. Online fault classification procedure.

1) Moving-Window Data Acquisition: For real-time fault detection, a moving-window scheme is employed. A sliding window of one cycle of the most recent three-phase voltage and current measurements is continuously collected, and the window is advanced every 0.5 cycle. This configuration enables near-real-time monitoring and fast protection response.

2) Real-Time Fault Classification: At each detection instant, the most recent one-cycle data are updated. The same feature-extraction procedure used during the model training (cf. Section II-C) is then applied to compute the phasor- and sequence-component features. The resulting feature vector is fed to the pretrained random forest classifier to produce the fault classification results.

3) Classifier Output and Fault Confirmation: At every detection step, the classifier outputs three pieces of information: fault/non-fault status, fault location, and fault type; however, due to the limited size of the data window, the input might not fully capture the characteristics of a fault event—particularly at its onset. In addition, transient or partial responses might resemble other fault types or locations, potentially leading to misclassification.

To mitigate this issue, a fault confirmation mechanism is introduced. A fault is only confirmed if the same classification result (fault type and location) is obtained in multiple consecutive detection steps. The choice of the confirmation threshold represents a trade-off between detection speed and classification reliability. In this study, a threshold of two identical classification results is adopted—that is, a fault is confirmed only after the same fault type and location are identified in two different detection steps.

IV. Numerical Study

This section presents the numerical results for the proposed random forest-based protection scheme. The objectives are to (i) quantify the accuracy of the fault detection, the fault type identification, and the fault location classification; (ii) assess the fault detection time under diverse fault scenarios; and (iii) compare the performance against separate single-task classifiers.

TABLE I
HYPERPARAMETERS OF THE RANDOM FOREST CLASSIFIER.

Hyperparameter	Value
Split criterion	Entropy
Number of trees	200
Maximum tree depth	20
Maximum features per split	$\sqrt{N_{features}}$
Minimum samples per leaf	1
Minimum samples per split	2

A. Random Forest Model Configuration

The random forest model used in the numerical experiments was obtained via a grid search over its hyperparameters. The final configuration is summarized in Table I. In particular, the model employs an entropy-based split criterion, a maximum tree depth of 20, and an ensemble size of 200 decision trees.

B. Online Testing Results

Fig. 6 illustrates the online fault classification process for an AB fault that occurred at 3 seconds in Area 2. The classifier performed a fault classification every 0.5 cycle using the most recent one-cycle window of measurements. Predicted labels are shown as orange 'x' markers, and ground-truth labels are indicated by blue dots. As expected, no fault was detected before the fault inception at 3 seconds. At the first decision instant following the fault inception, the classifier did not detect the fault due to insufficient fault data. At the next decision instant (0.5 cycle later), it correctly identified the AB fault in Area 2. When this output was repeated at the next decision step, the confirmation criterion was met, and the fault was confirmed with the correct type and location.

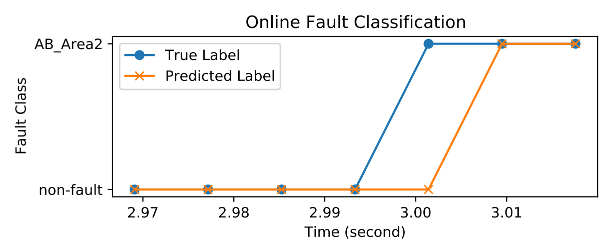

Fig. 6. Example of online fault classification.

The data-driven protection scheme was evaluated on 871 test cases. Under the joint framework, each classification result consists of three types of information: fault status (fault/non-fault), fault type, and fault location. The corresponding accuracies are reported in Table II: The fault/non-fault detection achieves 100%, the fault type identification achieves 97%, and the fault location classification achieves 94.9%. These results suggest that the proposed framework provides highly accurate fault detection, type identification, and localization

TABLE II
FAULT CLASSIFICATION ACCURACY.

	Fault Status	Fault Type	Fault Area
Accuracy	100%	97%	94.9%

979-8-3315-8138-1

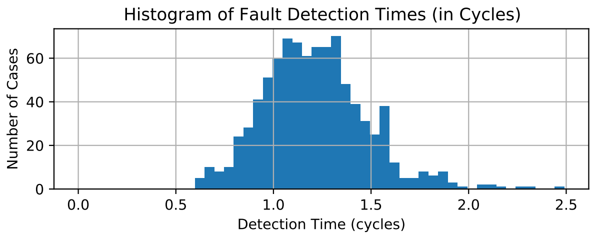

Fig. 7. Histogram of fault detection times.

for inverter-dependent transmission systems. In addition, the fault detection times for the 871 test cases are summarized in Fig. 7. All fault cases were successfully detected and confirmed in less than 2.5 cycles (0.0417 seconds), with a mean value of 1.21 cycles (0.0202 seconds).

Fig. 8 shows the confusion matrices for the fault type and fault area classifications, respectively. Each matrix entry reports the number of test cases for which the predicted class matches (diagonal entries) or differs from (off-diagonal entries) the true class. For the fault type classification, most misclassifications arise from confusion between ABC and ABCG faults, likely due to their highly similar system responses within the one-cycle data window. For the fault area classification, the confusion matrix indicates perfect classification in Area 3 (the reverse direction), but there are some misclassifications between Area 1 and Area 2. The results suggest that incorporating additional location-sensitive features may further enhance fault area classification.

C. Comparison with Separate Classifiers

To benchmark the proposed joint fault classification framework, we also trained three task-specific (separate) classifiers, each dedicated to fault detection, fault type identification, or fault area classification. The results on the same test dataset are summarized in Table III. Both approaches achieve strong performance: The joint model matches the separate models on fault status (100%), is within 0.1 percentage points (pp) on fault type, and yields a 0.8-pp improvement on fault area. These findings indicate that joint learning offers modest gains for the fault localization task without degrading detection or type classification. Beyond accuracy, the joint approach consolidates parameters and inference into a single network, simplifying deployment and potentially reducing computational and memory cost, while separate models might remain preferable when task-specific interpretability or independent updates are required.

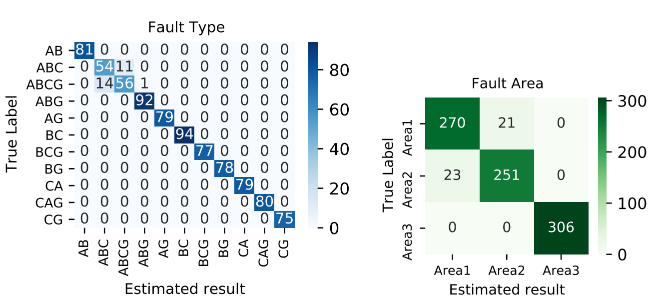

Fig. 8. Confusion matrix for the fault type and fault area classification.

TABLE III
FAULT CLASSIFICATION ACCURACY
(JOINT VS. SEPARATE CLASSIFIERS)

Task \ Classifier	Joint	Separate	Δ (pp)
Fault status	100%	100%	0.0
Fault type	97%	97.1%	−0.1
Fault area	94.9%	94.1%	+0.8

V. CONCLUSION

This paper introduced a joint random forest classifier for transmission line protection in IBR-dependent systems. By using a single model with phasor- and sequence-component features, embedded in a moving-window structure with fault confirmation logic, the scheme achieved fast and accurate decisions on fault status, fault type, and fault location while avoiding the complexity of designing and coordinating multiple separate protection functions.

Future work will focus on (i) integrating additional location-sensitive features to further improve fault location accuracy and (ii) validating the approach using hardware-in-the-loop studies to assess robustness to measurement noise, model mismatch, and communication constraints.

ACKNOWLEDGMENT

The authors would like to thank Abu Shouaib Hasan for his help with the PSCAD simulations and data acquisition.

REFERENCES

[1] S. H. Horowitz and A. G. Phadke, *Power System Relaying*, 3rd ed. John Wiley & Sons, 2008.
[2] X. Lyu *et al.*, "Impact of inverter-based resources on grid protection," Pacific Northwest National Laboratory, Tech. Rep. PNNL-36069, 2024.
[3] R. Chowdhury and N. Fischer, "Transmission line protection for systems with inverter-based resources – part i: Problems," *IEEE Transactions on Power Delivery*, vol. 36, no. 4, pp. 2416–2425, 2021.
[4] J. D. Hernández-Santafé and E. Sorrentino, "Problems and solutions concerning the distance protection of transmission lines connected to inverter-based resources," *Energies*, vol. 18, no. 6, 2025.
[5] M. Nagpal *et al.*, "Protection challenges and practices for interconnecting inverter based resources to utility transmission systems," IEEE Power & Energy Society PSRC, Tech. Rep. PES-TR81, 2020.
[6] E. O. Schweitzer, B. Kasztenny, and M. V. Mynam, "Performance of time-domain line protection elements on real-world faults," in *Proc. 69th Annu. Conf. for Protective Relay Engineers (CPRE)*, 2016.
[7] C. L. Peralta and H. N. V. Pico, "A time-domain protection approach for ac transmission systems with grid-forming resources," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, p. 9000318, 2024.
[8] F. V. Lopes, M. J. B. B. Davi, and M. Oleskovicz, "Assessment of traveling wave-based functions in inverter-based resource interconnecting lines," in *Proc. Int. Conf. on Power Systems Transients (IPST)*, 2023.
[9] S. Paladhi and A. K. Pradhan, "Adaptive distance protection for lines connecting converter-interfaced renewable plants," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2020.
[10] K. A. Kharusi, A. E. Haffar, and M. Mesbah, "Fault detection and classification in transmission lines connected to inverter-based generators using machine learning," *Energies*, vol. 15, no. 15, p. 5475, 2022.
[11] P. K. Bera, S. R. Pani, C. Isik, and R. C. Bansal, "A hybrid intelligent system for protection of transmission lines connected to pv farms based on linear trends," *Electric Power Systems Research*, 2024, early access.
[12] R. Fan, T. Yin, K. Yang, J. Lian, and J. Buckheit, "New data-driven approach to bridging power system protection gaps with deep learning," *Electric Power Systems Research*, vol. 208, p. 107863, 2022.
[13] Y. Chen, S. Chakraborty, A. Zamzam, and J. Wang, "End-to-end microgrid protection using distributed data-driven methods," *Applied Energy*, vol. 391, p. 125797, 2025.
[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

979-8-3315-8138-1