A Machine Learning-Based Method for Identifying Critical Distance Relays for Transient Stability Studies

Ramin Vakili and Mojdeh Khorsand *

Abstract: Protective relays play a crucial role in defining the dynamic responses of power systems during and after faults. Therefore, modeling protective relays in stability studies is crucial for enhancing the accuracy of these studies. Modeling all the relays in a bulk power system is a challenging task due to the limitations of stability software and the difficulties of keeping track of the changes in the setting information of these relays. Distance relays are one of the most important protective relays that are not properly modeled in current practices of stability studies. Hence, using the Random Forest algorithm, a fast machine learning-based method is developed in this paper that identifies the distance relays required to be modeled in stability studies of a contingency, referred to as critical distance relays (CDRs). GE positive sequence load flow analysis (PSLF) software is used to perform stability studies. The method is tested using 2018 summer peak load data of Western Electricity Coordinating Council (WECC) for various system conditions. The results illustrate the great performance of the method in identifying the CDRs. They also show that to conduct accurate stability studies, only modeling the CDRs suffices, and there is no need for modeling all the distance relays.

Keywords: distance relays; identifying critical protective relays; modeling protective relays in stability studies; power system protection; random forest classifier; relay misoperation; transient stability study

1. Introduction

The dynamic responses of the major assets in power systems, including generators, loads, and control systems, and the responses of the protection schemes are the two main aspects that define power system behavior during disturbances [1]. Post-analysis of the prior blackouts and major outages show that unforeseen relay misoperations played a significant role in leading the system toward these catastrophic events [2,3]. Therefore, proper modeling of protection systems in transient stability studies is crucial for obtaining a precise analysis of the behavior of the system during disturbances [4].

Various types of protective relays exist in bulk power systems that need to be included in transient stability studies of different disturbances to be able to accurately analyze the dynamic response of the system to the disturbances. Some of these relays, such as underfrequency load shedding (UFLS) and undervoltage load shedding (UVLS) relays, are usually included in transient stability studies. However, although distance relays are frequently used to protect the transmission lines of the system and can play a significant role in defining the topology of the system, they are usually not properly included in stability studies [4]. References [4–8] show that the actual response of the system to different disturbances might not be properly captured if the distance relays in the system are not modeled in these studies. The impact of modeling distance relays on the accuracy of transient stability studies is also highlighted in this paper with several case studies on the test system.
The most straightforward way to capture the behavior of distance relays in transient stability studies is to model all the relays in the system. However, thousands of distance relays are installed in real-world bulk power systems, such as the WECC system. Modeling this large number of distance relays is challenging due to two main reasons [5]:

- Commercial stability software tools, such as GE PSLF, have limitations on the number of dynamic models that can be included in the dynamic files used for stability studies. Hence, including a large number of distance relay models in the dynamic file of the system overwhelms the dynamic file and exceeds the current limitation of the software.
- In bulk power systems, keeping the setting information of thousands of distance relays updated in the dynamic file of the system is a challenging task since protection engineers change these settings for various purposes. The characteristics of the protection functions of distance relays are defined by the reach and the time delays of each of their protection zones. In an outdated dynamic file, the settings of one or more distance relays can be different from their actual values. It is shown in [5] that if the relay settings in the dynamic file are different from the actual settings of the relay, transient stability studies might represent an incorrect assessment of the dynamic response of power systems to major disturbances.

Due to these challenges, the need for a method to find the CDRs for each type of contingency has been identified by the industry [2,4,9]. Note that, in this paper, the distance relays that are likely to operate during a contingency are referred to as the CDRs for that contingency. To address this need, various methods are proposed in the literature.

As a method for identifying CDRs, the Independent System Operator (ISO) of New England monitors the impedance trajectories observed by relays in their planning studies. If any of the impedance trajectories traverse into the zone-3 reach of its related distance relay, the actual setting data of the relay is collected, and the relay is modeled in the planning studies [10]. The zone-3 reach of each distance relay is considered 300% of the impedance of its related transmission lines. The operations of distance relays affect the system behavior assessment for the rest of the simulation. Therefore, after identifying the distance relay that operates first during the transient stability study, the results of the study might not properly reflect the actual dynamic response of the system for the remainder of the simulation. Hence, the rest of the CDRs identified by this method might be incorrect.

References [11,12] have proposed new methods for identifying the distance relays that are at the electrical center of a system and might misoperate due to unstable power swings. A method is proposed in [4] that identifies a general list of CDRs for any contingency. The method is based on solving an optimization problem using the network topology and generator grouping information. The methods proposed in [4,11,12], however, only focus on the distance relay operations that occur due to unstable power swings. Thus, they do not consider the operations in the vicinity of the initiating event, which occur due to the initial impacts of the event. Additionally, the method in [4] provides a general list of CDRs for all the contingencies. However, the type and location of a contingency are the key factors that define the list of CDRs for the contingency. Thus, providing a specific list of CDRs for each contingency is of great importance.

In another method proposed in [13], a link between the power system simulator for engineering (PSS/E) and the computer-aided protection engineering (CAPE) software—a software tool that analyzes the dynamic responses of protective relays—has been developed. This link provides a platform for simultaneous assessment of protection system behavior and dynamic response of power systems during disturbances. During severe disturbances, the initiating event might have system-wide impacts and results in unstable power swings. These unstable power swings can cause the impedance trajectory observed by several distance relays in the system, which can even be located far from the initiating fault location, to enter one of the protection zones of these relays and cause their misoperation [9]. In this approach, however, only the operations of the distance relays close to the initial fault
locations are analyzed and the system-wide effects of disturbances on the distance relays are ignored.

Different methods for identifying CDRs are proposed in references [8,14,15]. The methods proposed in these references consider the location of the initiating events along with the sizes of these events as the factor for finding the CDRs for the events. Therefore, these methods also fail to find the distance relays that operate due to the system-wide effects of the initiating events.

Different methods for designing distance relay schemes are proposed in [16–26], which enables these schemes to detect if an abnormal condition is a stable power swing or a fault. Proper modeling of these schemes in transient stability studies might make it possible to detect distance relay misoperations due to unstable power swings. However, none of these methods can identify all the distance relays operations during a contingency and provide a complete list of CDRs for the contingency.

Reference [7] shows the importance of performing transient stability analysis with proper modeling of the essential elements of power systems, including protective relays, in identifying the system vulnerabilities. The paper also provides an approach based on the minimum voltage evaluation method to identify the CDRs for any contingency. However, the proposed method fails to identify a complete list of CDRs for a contingency under study and properly capture their performance during transient stability studies.

Many research works have proposed different techniques for power system clustering. The clustering is performed for various purposes including but not limited to controlled islanding and providing regional ancillary services [4]. These approaches can also be used for the purpose of identifying CDRs in the system. To do so, the system can be divided into multiple clusters in a way that the CDRs are located in the boundaries of these clusters. The clusters might be selected based on the similarity between the responses of the generators inside each cluster to disturbances. During severe disturbances, the generators within each cluster are in synchronism with each other, but they might lose synchronism with respect to the generators in other clusters. In this condition, which is called loss of synchronism (LOS), significant voltage drop might occur at the electrical center of the system, which, in turn, might lead to distance relay misoperations and unintended islanding in the system. Therefore, these distance relays are CDRs and are required to be modeled in transient stability studies. These distance relay misoperations usually occur at or near the boundaries of the clusters. Thus, these power system clustering methods can be used for the purpose of identifying CDRs.

In this regard, the authors in [27–29] proposed a power system clustering method based on slow-coherency method. References [28,29] also proposed an efficient approach for implementing controlled islanding using the slow-coherency method. Other measurement-based methods are also developed in [30,31] for power system clustering.

Similar methods can be developed to identify CDRs. However, the main drawback of using these power system clustering methods for identifying CDRs is as follows: first, these methods mostly rely on clustering power system based on generator buses. Hence, other buses (such as load buses) are not properly assigned to any cluster. Therefore, it is challenging to find a clear boundary between clusters, and identifying the distance relays on those boundaries is critical. Second, even if these methods accurately identify the distance relays located in the electrical center of the system, they still fail to capture all the distance relay operation in the system and provide a complete list of CDRs for any contingency.

The method proposed in [5] is an iterative algorithm that provides a specific list of CDRs for any contingency using the apparent impedance monitoring and the minimum voltage evaluation methods. Although this method can find all the CDRs for a contingency, it imposes a heavy computational burden on the computer system as several runs of transient stability studies should be performed in this method. This, in turn, leads to excessive simulation time.
A summary of the advantages and disadvantages of the methods proposed in the literature and the method proposed in this paper are provided in Table 1.

Table 1. Advantages and disadvantages of the methods proposed in the literature.

| References | Advantages                                                                 | Disadvantages                                                                 |
|------------|----------------------------------------------------------------------------|------------------------------------------------------------------------------|
| [10]       | Can capture all distance relay operations, fast, easy to implement         | Probability of giving a wrong list of CDRs                                  |
| [4,7,11,12]| Ability to identify distance relay operations in the electrical center    | Unable to identify all CDRs, providing a generic list of CDRs (not contingency-specific list) |
| [13]       | Providing a platform for simultaneous assessment of the response of the system and behavior of relays | Unable to identify all CDRs (only identifying CDRs in the vicinity of fault), computationally heavy |
| [8,14,15]  | Easy implementation, identifying CDRs close to fault location             | Unable to identify all CDRs                                                 |
| [16–26]    | Potentially able to identify distance relay misoperation due to power swings | Unable to identify all CDRs                                                 |
| [27–31]    | Provide probable list of CDRs on the boundaries of the clusters, provide efficient controlled islanding | Unclear boundaries between clusters, Unable to identify all CDRs              |
| [5]        | Ability to accurately identify all the CDRs, Contingency-specific method  | Computationally heavy, excessive simulation time                             |

As seen in Table 1, the major drawbacks of the methods proposed in the literature are as follows:

- Inability to identify all the CDRs for a contingency.
- Inability to provide a contingency-specific list of CDRs.
- Imposing heavy computational burden on the computing system and excessive simulation time.

To address the drawbacks of the methods in the literature, a fast method based on machine learning (ML) is developed in this paper to promptly identifies all the CDRs for any contingency. The method performs transient stability studies on many different types of contingencies and uses the results of these studies to train the ML model. These studies also capture various topologies and operating conditions of the system. After being trained, the model performs an initial early-terminated transient stability study of the contingency under study and utilizes the results to identify the CDRs for that contingency. Compared to the total number of distance relays in the system, the method identifies a far smaller number of distance relays as critical. Therefore, the method eliminates the need for modeling every distance relay existing in the system and the challenges associated with it. Moreover, it is indeed less challenging to keep the updated setting information for this small subset of distance relays in the dynamic files of bulk power systems.

The advantages of the proposed method are summarized as follows:

- Unlike the methods proposed in the literature, the ML-based method developed in this paper can identify all the CDRs for a contingency. This includes the distance relays that operate due to the initial impacts of the contingency and the ones that operate due to the system-wide impacts of the contingency.
- Unlike the other methods in the literature, the proposed method is very fast. Therefore, it can be used in planning studies, where a sheer number of contingencies are studied, to promptly identify the CDRs for the contingencies under study.
- The proposed method is trained considering different topologies and operating conditions of the system under study. Therefore, as the results reveal, the method is robust against changes in the topology and the operating condition of the system and can yield great results even under different topologies and operating conditions of the system.
The rest of this paper is organized as follows. Section 2 presents the proposed ML-based method for identifying CDRs. Section 3 describes the random forest (RF) algorithm. The metrics used to evaluate the performance of the trained RF model are introduced in Section 4. The grid search and the K-fold cross-validation methods are explained in Section 5. Section 6 evaluates the performance of the trained model in terms of the metric used. This section further assesses the performance of the method in identifying the CDRs and capturing the behavior of the system by performing transient stability studies of three major contingencies in the WECC system using the proposed method. Section 7 summarizes the conclusions.

2. The ML-Based Method

The proposed ML-based method works based on the latent correlation between the impedance trajectories observed by distance relays at the time of fault and 1 cycle after that and the behavior of the distance relays for up to 10 s later. The flowchart of the method is provided in Figure 1. The first step in the training stage (the blue rectangle in Figure 1) is to create a comprehensive dataset for training and testing an ML model. To create such a dataset, extensive transient stability studies are performed on different types of contingencies. Different topologies and operating conditions of the system are considered in these studies. The contingencies analyzed in these studies are bus faults, line faults, and generator outage disturbances, which include the outages of one or several transmission lines.

![Figure 1. The flowchart of the proposed method.](image)

To train the ML model, appropriate features and labels need to be extracted from these studies. The conceptual representation of the process of extracting the features and labels is illustrated in Figure 2. This figure shows a simple power system with 4 buses. In this system, a 3-phase fault occurs on the line between bus 2 and 3. The impedance trajectory observed by the distance relay during this contingency is also shown in Figure 2. To extract the required data for training the ML model, first, the real and imaginary parts of the impedance trajectory are sampled at the time of fault ($R_{1,1}$...
and \( X_{1,1} \), which represent the steady-state impedance seen by the relay. Then, the real and imaginary parts of the impedance trajectory observed by the relay at \( \frac{1}{4} \) of a cycle (\( R_{1,2} \) and \( X_{1,2} \)), \( \frac{2}{4} \) of a cycle (\( R_{1,3} \) and \( X_{1,3} \)), \( \frac{3}{4} \) of a cycle (\( R_{1,4} \) and \( X_{1,4} \)), and one cycle after the fault (\( R_{1,5} \) and \( X_{1,5} \)) are sampled, as well. These samples are shown by red, green, yellow, gray, and pink circles on the impedance trajectory plot. Note that \( \frac{1}{4} \) of a cycle is the step of the simulation. These samples are then used as the features of the ML model. The status of the distance relay is monitored for the duration of the transient stability study (600 cycles or 10 s after the fault). If the relay operates during this period, the case is labeled as 1; otherwise, it is labeled as zero. We performed a similar process on the WECC system. A large number of contingencies are studied, and the required features and labels are obtained from the distance relay all over the system to build a comprehensive dataset.

Figure 2. Conceptual representation of the process of extracting the features and labels for training an ML model.

Note that the performance of the model is also tested with using longer periods of impedance trajectories observed by the distance relays as the features of the model. It is observed that using longer periods of the data does not significantly influence the performance of the model. Thus, relay impedance trajectory for 1 cycle after the fault is considered as the input to satisfy the desired computational speed during the application phase (the orange rectangle in Figure 1). Additionally, note that due to the software limitations, to perform the transient stability studies of the contingencies, only the dynamic models of the distance relays of the transmission lines with the voltage level equal to or above 345 kV are included in the dynamic file of the system. If available, the results of the offline studies can be combined with the historical records of the impedance trajectories observed by the distance relays in the system during prior contingencies along with the records of their operations. Note that after the ML model is trained, it is used to identify CDRs for any contingency. Thus, the process of building a comprehensive dataset is only performed once at the training stage.
As there are two outcomes for predicting distance relays operations, the problem of predicting distance relay operations is a binary classification problem (with two classes of 0 and 1). Throughout this paper, the following definitions for Class 0 and Class 1 of the classification problem hold:

- **Class 0**: This class represents the samples of no-operation of distance relays, i.e., the samples where a distance relay does not operate for the entire simulation time, in the dataset.
- **Class 1**: This class represents the samples of distance relay operations in the dataset.

Different types of ML algorithms can be used for classification problems. As it can train fast, highly accurate, and robust models, an RF classifier is used here to train the ML model. To achieve the best performance, the hyperparameters of the RF model needs to be tuned. For this purpose, a grid search method with two stages is used. At each iteration of the grid search, different ML models are trained with different combinations of hyperparameter values. The performances of each of these models should be tested to find the hyperparameter values that create a model with the best performance. For this purpose, the K-fold cross-validation method is used. The grid search and the K-fold cross-validation methods are explained in Sections 5.1 and 5.2, respectively.

Note that while the RF is used as the algorithm of choice, any other classification algorithm can be used with the proposed method without need for any significant change in the overall approach.

In the next stage (the stage of dynamic analysis with CDRs), the trained model is used to identify the CDRs for any contingency under study. Note that this contingency might be a new contingency that has not been studied during building the dataset; as the results of this paper confirm, the proposed approach performs well for such out-of-sample scenarios.

In this stage, to extract the required data for the RF model, first, the proposed method performs an initial transient stability study of the contingency under study. No distance relay models are included in this initial study. Since the trained model only needs the impedance trajectories observed by distance relays up to 1 cycle after the fault as its input features, the initial transient stability study can be terminated 1 cycle after the fault, which can considerably reduce the simulation time. Note that at each time step of the transient stability study, the impedance trajectories observed by distance relays can be easily obtained without any change in the current practice of performing transient stability studies and without any need to model the distance relays. Bus voltages and line currents can be used for this purpose. Figure 3 shows the conceptual representation of the process for extracting the required features of the RF model at this stage. As seen in this figure, the bus voltages from bus 1 and the line current flowing from bus 1 to bus 2 are sampled at the time of fault \( V_{1,1} \) and \( I_{1,1} \), along with \( \frac{1}{4} \) of a cycle \( V_{1,2} \) and \( I_{1,2} \), \( \frac{1}{2} \) of a cycle \( V_{1,3} \) and \( I_{1,3} \), \( \frac{3}{4} \) of a cycle \( V_{1,4} \) and \( I_{1,4} \), and one cycle after the fault \( V_{1,5} \) and \( I_{1,5} \). The impedance trajectory observed at the location of the distance relay (the required features for the RF model), then, can be computed using Equation (1).

\[
Z_{1,i} = \frac{V_{1,i}}{I_{1,i}} \quad \forall i \in \{1, 2, 3, 4, 5\} 
\]
After obtaining the last sample from the results of the simulation, the transient stability study can be terminated, and the obtained features are fed into the trained model. We followed the same procedure on the WECC system. The bus voltages and line currents at the locations of distance relays all over the system are sampled. Then, the impedance trajectories observed by distance relays are calculated, and are fed to the trained RF model.

Consequently, the trained model predicts whether any of the distance relays in the system operate for the entire simulation time or not. The distance relays that are predicted to operate are identified as the CDRs for the contingency. Thus, their dynamic models are included in the dynamic file of the system. In this stage, the updated relay settings for the identified CDRs can be obtained (from the protection groups of utilities) and be included in the dynamic models of the distance relays to achieve a more precise assessment of the dynamic response of the system. Finally, using the stability software, the method captures the dynamic response of the system to the contingency by performing a transient stability study with only modeling the CDRs. Note that the small number of the CDRs identified by the method can easily be included in the studies without any of the challenges of modeling all the distance relays in the system.

3. The Random Forest Algorithm

The RF algorithm is a well-known and successful ML algorithm that has widely been used for various classification and regression problems. Composed of multiple single decision trees (DTs), the RF operates as an ensemble ML algorithm.
Figure 4 shows a conceptual RF that is composed of $N$ decision trees. As illustrated in this figure, each DT of the RF is trained using a subset of the dataset. Each individual DT might use a specific subset of the features (the impedance trajectories observed by distance relays) and learn a series of explicit rules to decide if a distance relay operates during the transient stability study period or not. The rules learnt by each DT are different from those learnt by the other DTs. As seen in the figure, each DT of the RF is composed of a root (the first circle in each DT), internal nodes (colored circles in the figure), branches (the lines connected to the circles), leaves (the final nodes), and classification rules [32]. Each internal node is an attribute on a feature (e.g., whether the real part of impedance trajectory observed by distance relay at the time of fault is greater than a particular number or not). Each branch shows the outcome of a test. Each leaf represents a class label (0 or 1), and each classification rule is the path from the root to leaf. In classification problems (such as the problem of this paper), Gini impurity or entropy are commonly used to decide how nodes of a DT branch. In this paper, Gini impurity is used for decision making at each node of each individual DT of the RF. Gini impurity is a number between 0 and 0.5 and can be calculated using Equation (2) [33].

$$Gini = \sum_{i=1}^{C} f_i(1 - f_i) = 1 - \sum_{i=1}^{C} f_i^2$$  \hspace{1cm} (2)
classification of the sample at the related note is wrong (the sample belongs to another class) if a random class label is given to a random out-of-pocket sample in the dataset (e.g., a particular distance relay operation during a specific contingency) based on the class distribution in the dataset. In this regard, if an attribute on a node divides a dataset \( D \) with size of \( n \) into two subsets \( D_1 \) and \( D_2 \) with sizes of \( n_1 \) and \( n_2 \), the Gini impurity for that attribute can be calculated using Equation (3). Consequently, the attribute that yields the minimum value for the Gini impurity is selected for splitting the node [34].

\[
Gini_A(D) = \frac{n_1}{n} Gini(D_1) + \frac{n_2}{n} Gini(D_2)
\]

(3)

As seen in Figure 4, in classification problems, such as the problem of this paper, after each DT makes its decision and classifies distance relays into Class 1 (operation) and Class 0 (no-operation), the RF methods select the majority of the votes from the DTs to make the final prediction. In the regression problems, the output of the RF is selected as the average of the predictions of its DTs.

As the depth of the DTs in RF growth, they can learn very complex rules. Therefore, tracking the flow of data within all the trees of the RF and interpreting the classification rules they develop can be really challenging (even impossible). More details about the RF algorithm can be found in [35,36].

Numerous advantages of the RF have made it a popular ML algorithm to be widely deployed in many problems, including the problem of this paper [35–37]. Major advantages of the RF are as follows:

- One of the widely known problems of a single DT is overfitting. This causes DT to have a poor performance on new datasets. To overcome this problem, in the RF algorithm, multiple DTs are trained using bootstrap aggregating or bagging techniques. Each of these DTs is trained on different part of the training set. This makes RF robust against overfitting, and significantly enhances its accuracy on new datasets [38].
- The performance of the RF in handling unbalanced and non-linear datasets is outstanding [38]. For the application of this paper, as the dataset is unbalanced and has more cases of Class 0 than Class 1, this feature of the RF algorithm is of great importance.
- Owing to its high speed (both for training and prediction), the RF can easily handle the datasets with high dimensionality. The reason is that the RF only uses a subset of features to train each DT. Additionally, the computational process of the RF algorithm can be divided between multiple computers, i.e., the algorithm is parallelizable, which can significantly increase the speed of the algorithm. The high speed of the RF algorithm is favorable for the problem of this paper since it enables the trained model to predict the operation of all the distance relays that exist in bulk power systems in a short time [37,38].

4. Metrics for Testing the Trained Model

Two metrics of Recall and Precision are used to test the performance of the trained model. These metrics are introduced in this section. Note that in the formulations of Recall and Precision, \( TP \) is the number of true-positive cases, i.e., the cases that belong to Class 1 (distance relay operations cases), and the trained model correctly predicts that they belong to Class 1. \( FN \) is the number of false-negative cases, i.e., the cases that belong to Class 1; however, the trained model falsely predicts that they belong to Class 0 (distance relay no-operation cases). Finally, \( FP \) is the number of false-positive cases, i.e., the cases that belong to Class 0; however, the trained model falsely predicts that they belong to Class 1.

Recall is the ratio of the number of cases that a model correctly predicts to belong to Class 1 to the number of all the cases in the test set that belong to Class 1 (all the distance relay operation cases in the test set). Recall can be formulated using Equation (4) [39].

\[
Recall = \frac{TP}{TP + FN}
\]

(4)
In the application of identifying CDRs, it is of the highest importance that the trained model does not miss any distance relay operation. In other words, ideally, the trained model should correctly predict all the distance relay operations. Otherwise, the missed distance relay operations are not listed as CDRs and hence are not included in the final transient stability study of the contingency. To guarantee that no CDRs is missed, FN should be very low, and the Recall value should be very high (close to 1). Therefore, in this paper, maximizing the Recall value is considered the objective of the grid search on the hyperparameters of the RF.

Precision is the ratio of the number of cases that a model correctly predicts to belong to Class 1 to the number of all the predictions of the model in the test set indicating that a case belongs to Class 1 (total case 1 predictions). Precision can be formulated using Equation (5) [39].

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

In this paper, a high number of false-positive cases show that the trained model is prone to misclassify many of the no-operation cases of distance relays (Class 0) as operation cases (Class 1) and identify those distance relays as critical. Therefore, many distance relays might be included in the dynamic file of the system, which is undesirable; modeling many distance relays might cause the same challenges as modeling all the distance relays. However, modeling additional distance relays in the final stability analysis will not impact the accuracy of the analysis as long as the number of these relays does not exceed the limitation of transient stability analysis software tools and the relay settings are accurate. Thus, although maximizing the Precision value is not considered the main objective, it is tried to achieve a suitable value for Precision that does not result in the modeling of many distance relays in transient stability studies.

In order to visualize the definition of the Recall and Precision, Figure 5 is provided. In this figure, cases that belong to Class 1 (relay operation cases) are shown with plus signs, and the cases that belong to Class 0 (no operation cases of the relays) are shown with circles. The black line in the middle of the figure is the hypothesis learned by the trained model, in a way that the model predicts all the cases that are above the line as the cases belonging to Class 1 and all other cases as the cases belonging to Class 0. In this condition, the cases that belong to Class 1 and are above the line are true-positive prediction (as the trained model correctly predicts them to belong to Class 1) and are shown by green plus signs. The cases that belong to Class 0 and are below the line are true-negative prediction (as the trained model correctly predicts them to belong to Class 0) and are shown by green circles. All the cases that belong to Class 0 and are above the line are the false-positive predictions (as the trained model erroneously predict them to belong to Class 1) and are shown by red circles. All the cases that belong to Class 1 and are below the line are the false-negative predictions (as the trained model erroneously predict them to belong to Class 0) and are shown by red plus signs. Therefore, in this case, the Recall is the number of green plus signs divided by the sum of the numbers of green plus signs and red plus signs. On the other hand, the Precision is the number of green plus signs divided by the sum of the numbers of green plus signs and red circles.
5. The Grid Search and the K-Fold Cross-Validation Methods

5.1. The Grid Search Method

The grid search, also known as parameter sweep, is a traditional method to optimize the hyperparameters of an ML algorithm. It includes searching within a defined range of the hyperparameter space of the ML algorithm. A performance measure should be considered to direct the grid search in optimizing the hyperparameters. In this paper, as mentioned earlier, maximizing Recall value is considered the performance measure. Therefore, during the grid search, the Recall is measured by using the cross-validation method [40,41].

The flowchart of the grid search method is provided in Figure 6. As seen in this figure, the method starts after providing the list of hyperparameters that needs to be tuned, their range, and number of steps of the grid search. Then, the method builds its search space for different hyperparameters and train an RF model with each combination of the hyperparameter values. Using K-fold cross-validation, the trained model is tested, and the related Recall value is recorded. At this stage, the method checks whether all the possible combination of hyperparameter values is tried. If so, the method trains the final RF model with the best hyperparameter values found in the iterations. Otherwise, the method tries other combinations of hyperparameter values to train the RF model.

The grid search method can be easily used to optimize the hyperparameters of an ML model. Its biggest disadvantage is that when the number of hyperparameters to be tuned increases, this method can be computationally heavy and time-consuming. However, because the hyperparameters to be tuned are usually independent of each other, the processing task of the grid search can be divided into several parallel tasks with little or no effort, and each task can be performed with a separate computer [42].
The grid search method can be easily used to optimize the hyperparameters of an ML model. Its biggest disadvantage is that when the number of hyperparameters to be tuned increases, this method can be computationally heavy and time-consuming. However, because the hyperparameters to be tuned are usually independent of each other, the processing task of the grid search can be divided into several parallel tasks with little or no effort, and each task can be performed with a separate computer [42].

Building the search space of the grid search

- Are all the combination tried?
  - No
  - Yes

- Training the RF model with specific combination of hyperparameter values
- Evaluating the performance of the model using the K-fold cross-validation method
- Train the final model with the best hyperparameter values found

Figure 6. The flowchart of the grid search method.

5.2. The K-Fold Cross-Validation

For testing the models trained during the grid search method on the new cases that are not included in the training set, the K-fold cross-validation method is used. The K-fold cross-validation method is a resampling procedure with a single parameter called K that shows the number of parts that the dataset needs to divide into. The flowchart of this method is provided in Figure 7.

As seen in Figure 7, the method first randomly shuffles the dataset. Then, it divides the dataset into K parts. These parts are referred to as folds. At each iteration of the method, one of the folds is excluded from the dataset (hold out fold), and the other K-1 folds comprise the training set. Then, a model is trained on the training set. The excluded fold serves as the test set for testing the performance of the trained model. The Recall value of the model trained at each iteration is recorded and added to the sum of the Recall values obtained at different iterations of the method. The stopping criteria for the iterations is that all the folds are used as a test set. Therefore, at each iteration, the method checks whether all the folds are used as the test set. If so, the method calculates the average of the recorded Recall values in all the iterations and reports it as the performance metric of the model [43,44].
6. Case Studies

The proposed method is tested on the 2018 summer peak load data of Western Electricity Coordinating Council (WECC) system. The details of the test system are provided in Table 2. The maximum capacity of the system to generate active power is 281.38 GW and the load of the system is 174.3 GW. The PSLF is used for performing stability studies. Note that as the PSLF normally used in the industry for performing transient stability studies on the WECC system, it is used in this paper for the same purpose. However, the method can be implemented using any other stability software tool.

Overall, to build the dataset, 929 contingencies under 5 different topologies and the operating conditions of the system are studied. To observe the behavior of distance relays in these studies, a reference dynamic file is created for the system that includes two widely used distance relay models in the PSLF model library, namely, “Zlin1” and “Zlinw” [45].

Figure 7. The flowchart of the K-fold cross-validation method.
Table 2. The details of the test system.

| Assets          | The Number of Assets in the System |
|-----------------|------------------------------------|
| Buses           | 23,297                             |
| Transmission lines | 18,347                           |
| Transformers    | 9050                               |
| Generators      | 4224                               |

Zlin1 is a distance relay model in the PSLF model library with three operation zones [45]. Zlin1 needs to be modeled on each line with its related settings. Whereas Zlinw is a generic distance relay model in the PSLF library with two operation zones [45]. Zlinw does not need to be specifically modeled for every transmission line; rather, it monitors all the lines within a predefined voltage range. Depending on their voltage level, each of the lines in the system is modeled with one of these two distance relay models. Table 3 shows the voltage level of the lines whose distance relays are modeled with the Zlin1 and Zlinw models.

Table 3. The voltage level of the lines modeled with Zlin1 and Zlinw models.

| Model | Transmission Lines Voltage Level |
|-------|----------------------------------|
| Zlin1 | >345 kV                          |
| Zlinw | 100-345 kV                       |

Note that since the Zlin1 needs to be included for each line, separately, due to the software limits, it is only modeled on the lines with a voltage level equal to or higher than 345 kV as these lines are the most critical lines of the system. For the lines with a lower voltage level, the Zlinw is used, as this model does not need to be included on each line separately.

To train the RF model, the “scikit-learn” library [46], which is a widely used library in Python, is used. For hyperparameter tuning, a two-stage grid search method with the objective of maximizing the Recall is used. The K parameter in the K-fold cross-validation method is considered to be equal to 5. Therefore, at each iteration of the K-fold cross-validation method, 80% of the dataset compose the training set, and the remaining 20% compose the test set.

The most important hyperparameters of the RF classifier are class weights, the maximum depth of the trees (MDT), the number of trees in the forest (NTF), the minimum sample split, the minimum sample leaf, the number of features considered when looking for the best split, and whether bootstrap samples are used when building trees [46]. Precise tuning of these hyperparameters is critical for enhancing the performance of the model. Explaining the role of each of these hyperparameters in the performance of the RF model is not in the scope of this paper, and more details in this regard are provided in the “scikit-learn” library [46]. A wide range of values can be selected for these hyperparameters. Hence, an approximate best value for each hyperparameter is obtained at the first stage of the grid search. Then, at the second stage, the grid search method searches around the approximate value obtained from the first stage to find the exact value for each hyperparameter that yields the highest Recall. To visualize how the Recall value of the trained model changes with the change in each hyperparameter of the RF, the change in the Recall value for the various values of NTF and MDT is provided in Figure 8a. Additionally, the change in the Recall value for the various values of class weights and NTF is provided in Figure 8b. Note that in Figure 8b, the set of the numbers on the X-axis (the class weight axis) shows the weight of each class. For example, “0:1, 1:10” means the weight of 1 for Class 0 and the weight of 10 for Class 1. The other sets of numbers are interpreted likewise. The “balanced” weight on this axis means a weight for a class that is inversely proportional to the class frequency in the input dataset [46]. For example, if the “balanced” option is used for the class weight and the number of
samples in the dataset belonging to Class 1 is 3 times of those belonging to Class 0, the weight of Class 1 will be 1/3 of the weight of Class 0.

Figure 8. The change in the Recall value for different values of MDT and NTF (a), as well as for the different values of class weights and NTF (b).

Figure 8a shows that in comparison to NTF, MDT has slightly more impact on the performance of the models. It also shows that the model has a better performance when MDT is set to 10. This indicates that increasing the depth of the trees does not lead to training models with better performances and it can show that each individual DT does not need to develop a very sophisticated sets of rules for decision making. Figure 8b shows that the impact of class weights on the performance of the trained model is the highest among all the hyperparameters, and as the weight of Class 1 increases, the Recall value increases as well. The reason is that with the increase in the class weight of Class 1, the number of the cases that belong to Class 1 (relay operation cases) increases in the dataset. Therefore, the algorithm trains models that are more biased toward correctly classifying the cases that belong to Class 1. This increase in the Recall value comes at the price of deteriorating the Precision value, because as the model become more biased toward correctly classifying the cases that belong to Class 1, the chance of misclassifying the cases that belong to Class 0 (no-operation cases of the relays) as class 1 increases. Therefore, although the weight of 100 for Class 1 gives the highest Recall value in the first stage of the grid search (above 0.99), the Precision value is very poor for this weight of
Class 1 (below 0.5). Thus, for the first stage, the weight of 10 is selected for Class 1, which yields a high Recall value (around 0.979), while it keeps the Precision at a reasonably high value (above 0.7).

Table 4 shows the best values obtained for each hyperparameter from the grid search method. Using the values provided in Table 4 for training the RF model yields the best results in terms of the Recall value (while maintaining a high Precision value). The performance of the model is evaluated on the entire dataset, including over 900 contingencies, using the K-fold cross-validation method. The results reveal that the model has Recall and Precision values of 0.981 and 0.737, respectively. This high Recall value ensures that the model can identify all the CDRs and does not miss any distance relay operation, even with different topologies and operating conditions of the system, whereas the reasonably high value of Precision shows that the model does not identify many distance relays as critical. The final RF model is trained on the whole dataset and implemented in the application phase of the method to identify the CDRs for any contingency under study.

Table 4. The selected hyperparameters of the RF model.

| Hyperparameter                      | Tuned Value                                                                 |
|-------------------------------------|-----------------------------------------------------------------------------|
| Class weights                       | 1 for Class 0 and 20 for Class 1                                            |
| Number of trees in the forest       | 15                                                                          |
| Maximum depth of trees              | 15                                                                          |
| Minimum sample split                | 2                                                                           |
| Minimum sample leaf                 | 2                                                                           |
| The number of features              | “auto”—equal to the square root of the number of all the features         |
| Bootstrap                           | “False”—meaning the whole dataset is used to build each tree              |

To further illustrate the performance of the proposed method in identifying the CDRs and capturing the precise response of the system during various contingencies, three different contingencies are considered for more detailed studies. Various topologies and operating conditions of the system are considered in these studies. For each contingency, three different cases are analyzed:

- **Case 1:** To show how including distance relays in transient stability studies impacts the results of the studies, the first case considered in studying each contingency performs the studies without modeling any distance relays.

- **Case 2:** In this case, the reference dynamic file, which was used to conduct the studies during creating the dataset stage, is used for performing stability studies of each contingency. As mentioned earlier, this dynamic file includes Zlin1 models for the lines with a voltage level equal to or higher than 345 kV and a Zlinw model for the lines with a voltage range of 100 kV to 345 kV. This case is referred to as the reference case throughout the paper.

- **Case 3:** In this case, the method of this paper is used to identify the CDRs for the contingencies under study. Then, the stability studies are conducted with only including Zlin1 models for the identified CDRs. Similar to the second case, the Zlinw model is used in the third case to monitor the lines with a voltage range of 100 kV to 345 kV. Note that, the method is able to find all the CDRs on the lines regardless of the voltage level of the line. However, to be able to compare the results of this case (using the proposed method) with those of Case 2 (the reference case), the settings of the method are changed to only identify the CDRs on the lines with a voltage level equal to or higher than 345 kV.

In both the second and the third cases, the reach of Zlin1 operation zones are considered 80%, 120%, and 220% of the line impedance, for zone 1, 2, and 3, respectively. Additionally, time delays of 0, 0.2, and 0.3 s are considered for zone-1, -2, and -3 operations, respectively. The same zone-1 and -2 reaches and time delays are used for the Zlinw model. Circuit breaker delay time is set to 0.05 s, which means that it takes 0.05 s for the circuit breaker to open.
breaker to open the line after receiving the tripping signal from its distance relay. It needs to be mentioned that although, for simplicity, generic values are chosen for relay settings, the method can be implemented with precise relay settings obtained from the protection groups of electric utility companies to provide a more precise assessment of the dynamic response of the system.

Note that these contingencies are only provided to show in more details the performance of the proposed method and the response of the system for three simulated out-of-sample contingencies. The values of 0.981 and 0.737 reported earlier for the Recall and Precision metrics are obtained from testing the trained model on the entire dataset (over 900 contingencies). To protect the proprietary data, arbitrary numbers are used throughout the paper to represent power system assets.

Contingency 1: In this contingency, a bus fault occurs on Bus 3 of the system and is cleared after 4 cycles by removing California Oregon Intertie (COI) including three 500 kV tie lines. During the summer peak load, COI transfers 4,113 MW of active power from its north to its south. As these three lines transfer a considerably high amount of power, they are very important tie lines of the WECC system, and their outage is a known critical emergency for the WECC system, which has the potential to jeopardize the system stability. Therefore, this critical N-3 outage has been considered for a more detailed analysis. To evaluate the performance of the method under new pre-fault operating conditions, a new pre-fault operating condition of the WECC system is considered for this contingency. In the new operating condition, a uniform increase of 2 percent in the large loads (above 100 MW) of the areas that import power through COI, which leads to a 55.71 MW increase in the net load of these areas, is considered. This increase in the load is compensated by three of the largest generators existing in the areas that send power through COI, without violating the limitations of any generation unit in the system. The results of the studies conducted for Case 1 (modeling no distance relay), Case 2 (the reference case), and Case 3 (the proposed method) of this contingency are illustrated in Figure 9a,b,c, respectively.

By comparing Figure 9a,b, it is seen that, compared to the reference case, when distance relays are not modeled in the stability studies, the results of these studies reveal a different dynamic response of the system in terms of the relative rotor angles (RRA) of a group of generators. Additionally, the similarity of Figure 9b,c illustrate that even this major disturbance under the new pre-fault operating condition of the system using the proposed method and the reference case leads to a similar assessment of the dynamic response of the system. Note that the results of the stability study reveal the similarity of the RRA of all the generators in the reference case and the proposed method. For the sake of clarity, however, only the RRA of a group of selected generators are shown.

Table 5 shows the list of the relays that operate in the reference case (Case 2) and the proposed method (Case 3). Lines 1, 2, and 3 are in the area that imports power through COI, and line 4 is in the area that exports power through COI (Lines 1–4 are close to the location of fault). Therefore, the distance relays of these lines operate due to the initial impacts of the disturbance. Correctly capturing the operations of these distance relays shows that the method can identify the distance relay operations due to the initial impacts of disturbances. All other lines listed in Table 5 are in the areas far from the fault location. Lines 8, 10, 15, and 17 are the tie-lines that connect two areas far from the fault location. The ability to correctly predict the operations of the distance relays of these lines and identify them as critical shows that the method can capture the distance relay operations that occur as a result of system-wide impacts of disturbances. Therefore, it is observed that the method can identify all the CDRs for this severe contingency under the new pre-fault operating condition of the system.
protection groups of electric utility companies to provide a more precise assessment of the dynamic response of the system.

Note that these contingencies are only provided to show in more details the performance of the proposed method and the response of the system for three simulated out-of-sample contingencies. The values of 0.981 and 0.737 reported earlier for the Recall and Precision metrics are obtained from testing the trained model on the entire dataset (over 900 contingencies). To protect the proprietary data, arbitrary numbers are used throughout the paper to represent power system assets.

**Contingency 1:**
In this contingency, a bus fault occurs on Bus 3 of the system and is cleared after 4 cycles by removing California Oregon Intertie (COI) including three 500 kV tie lines. During the summer peak load, COI transfers 4,113 MW of active power from its north to its south. As these three lines transfer a considerably high amount of power, they are very important tie lines of the WECC system, and their outage is a known critical emergency for the WECC system, which has the potential to jeopardize the system stability. Therefore, this critical N-3 outage has been considered for a more detailed analysis.

To evaluate the performance of the method under new pre-fault operating conditions, a new pre-fault operating condition of the WECC system is considered for this contingency. In the new operating condition, a uniform increase of 2 percent in the large loads (above 100 MW) of the areas that import power through the COI, which leads to a 55.71 MW increase in the net load of these areas, is considered. This increase in the load is compensated by three of the largest generators existing in the areas that send power through COI, without violating the limitations of any generation unit in the system. The results of the studies conducted for Case 1 (modeling no distance relay), Case 2 (the reference case), and Case 3 (the proposed method) of this contingency are illustrated in Figure 9 (a), (b), and (c), respectively.

![Figure 9](image_url)

**Figure 9.** The RRA of a set of generators in Contingency 1: (a) Case 1, (b) Case 2, and (c) Case 3.
Table 5. The list of all distance relay operations in Contingency 1.

| Time (s) | Relay                  | Time (s) | Relay                  |
|---------|------------------------|----------|------------------------|
| 1.050   | Lines 1, 2, 3          | 1.050    | Lines 1, 2, 3          |
| 1.888   | Line 4                 | 1.888    | Line 4                 |
| 2.629   | Line 5                 | 2.629    | Line 5                 |
| 2.671   | Line 6                 | 2.671    | Line 6                 |
| 2.683   | Lines 7                | 2.683    | Lines 7                |
| 2.767   | Line 8                 | 2.767    | Line 8                 |
| 2.771   | Line 9                 | 2.771    | Line 9                 |
| 2.950   | Line 10                | 2.950    | Line 10                |
| 2.979   | Line 11                | 2.979    | Line 11                |
| 3.009   | Line 12                | 3.009    | Line 12                |
| 3.025   | Line 13                | 3.025    | Line 13                |
| 3.179   | Line 14                | 3.179    | Line 14                |
| 5.921   | Line 15                | 5.921    | Line 15                |
| 6.321   | Line 16, 17            | 6.321    | Line 16, 17            |

Having a closer look into Table 5 and Figure 9 shows that following the distance relay operations on the lines 8 and 10 (two of the critical tie lines of the system) at 2.767 (s) and 2.95 (s) the RRA of the generators starts to deviate more from those in Case 1. Additionally, it is seen that immediately after distance relay operation on line 17 (which is another tie line of the system) at 6.321 (s) the divergence in the network solution occurs in cases 2 and 3. This further highlights the importance of identifying these CDRs in the system and modeling them in stability studies. It needs to be mentioned that the severe disturbance of COI outage has system-wide effects and leads to many distance relay operations. Proper remedial actions, such as load shedding and controlled islanding, are designed for this contingency by the industry. However, in this research, as we are not provided with the remedial actions employed by the industry, these actions are not included in the stability studies. However, if the remedial actions are provided, the method is expected to be able to find the CDRs and capture the behavior of the system.

Note that improving the reliability and resiliency of power systems by designing proper preventive/remedial actions requires a detailed assessment of the responses of different assets and protection schemes in the system, and only analyzing if the system maintains its stability or not during a contingency is not sufficient. Thus, although in this contingency, the system becomes unstable in either cases of modeling or not modeling distance relays, to have a precise assessment of the system response, which is crucial for devising proper preventive and remedial actions, modeling distance relays is necessary.

Contingency 2: The same N-3 COI outage that is analyzed in Contingency 1 is also considered for Contingency 2. However, to test the performance of the method in the case of a different pre-fault topology of the system, the pre-fault topology of the system is modified by taking two critical 500 kV transmission lines of the system out of service.

The results of stability studies performed for Case1, Case 2, and Case 3 of this contingency are shown in Figure 10a,b,c, respectively. Figure 10a shows that for this contingency, if the distance relays are not modeled in stability studies, the dynamic response of the system in terms of the RRA of a set of generators is different from the reference case. The similarity of Figure 10b,c demonstrates that the system dynamic response is similar in both the reference case and the proposed method, showing the accuracy of the method in capturing the response of the system. Additionally, in this contingency, with modeling distance relays, a divergence occurs in the network solution at 6.259 (s), which leads to the termination of the transient stability study.
Figure 10. The RRA of a set of generators in Contingency 2: (a) Case 1, (b) Case 2, and (c) Case 3.

Table 6 provides the list of all the distance relay operations in Cases 2 and 3. Similar to Contingency 1, distance relay operations occur both on the lines close to the fault location and the lines far from it. As seen in Table 6, the method can correctly predict all the distance relay operations in this contingency, which shows that the method can detect distance relay operations as a result of the initial and system-wide impacts of this severe contingency under the new pre-fault topology of the system. Additionally, similar to Contingency 1, a
significant change in the RRA of the generators in Case 1 with respect to those in Case 2 are observed after the distance relay operations on lines 8 and 10 at 2.863 (s) and 3.046 (s), which further shows the importance of identifying these CDRs and including them in transient stability studies.

Table 6. The list of all distance relay operations in Contingency 2.

| Time (s) | Relay       | Time (s) | Relay       |
|---------|-------------|----------|-------------|
| 1.050   | Lines 1, 2, 3 | 1.050    | Lines 1, 2, 3 |
| 1.917   | Line 4      | 1.917    | Line 4      |
| 2.692   | Line 5      | 2.692    | Line 5      |
| 2.854   | Line 9      | 2.854    | Line 9      |
| 2.863   | Lines 6, 8  | 2.863    | Lines 6, 8  |
| 2.884   | Line 7      | 2.884    | Line 7      |
| 3.046   | Line 10     | 3.046    | Line 10     |
| 3.063   | Line 11     | 3.063    | Line 11     |
| 3.092   | Line 12     | 3.092    | Line 12     |
| 3.096   | Line 13     | 3.096    | Line 13     |
| 3.275   | Line 14     | 3.275    | Line 14     |
| 5.925   | Line 15     | 5.925    | Line 15     |
| 6.242   | Line 17     | 6.242    | Line 17     |
| 6.259   | Line 16     | 6.259    | Line 16     |

Contingency 3: In this contingency, the proposed method is tested in another type of disturbance. The disturbance is an N-3 generator outage. The disconnected generators are among the generators with the highest active power generation in the system. These three generators produce a total active power of around 1268 MW. The pre-fault operating condition of Contingency 1 is considered in this contingency as well. This N-3 generator outage has system-wide impacts, causing several distance relay misoperations due to unstable power swings. Therefore, identifying these CDRs by the proposed method further shows its ability to identify the operations of distance relays due to unstable power swings at locations far from the initial event.

The results of stability studies performed for Case 1, Case 2, and Case 3 are shown in Figure 11a,b,c, respectively. Comparing Figure 11a,b reveals that if the distance relays are not modeled in stability studies, the studies cannot capture the actual response of the system. Additionally, as seen in Figure 11c, the response of the system in the case of using the proposed method is similar to that in Case 2 shown in Figure 11b. This further shows the effectiveness of the method in capturing the dynamic response of the system.

Table 7 provides the list of distance relay operations in both the reference case and the proposed method. As seen in this table, the method can correctly detect all the distance relays operations in the reference case. This further shows the effectiveness of the method in predicting distance relay operation due to the system-wide impacts of a disturbance.
The Reference Case The Proposed Method

| Time (s) | Relay |
|---------|-------|
| 1.050   | Lines 1, 2, 3 |
| 1.917   | Line 4 |
| 2.692   | Line 5 |
| 2.854   | Line 9 |
| 2.863   | Lines 6, 8 |
| 2.884   | Line 7 |
| 3.046   | Line 10 |
| 3.063   | Line 11 |
| 3.092   | Line 12 |
| 3.096   | Line 13 |
| 3.275   | Line 14 |
| 5.925   | Line 15 |
| 6.242   | Line 17 |
| 6.259   | Line 16 |

Contingency 3: In this contingency, the proposed method is tested in another type of disturbance. The disturbance is an N-3 generator outage. The disconnected generators are among the generators with the highest active power generation in the system. These three generators produce a total active power of around 1268 MW. The pre-fault operating condition of Contingency 1 is considered in this contingency as well. This N-3 generator outage has system-wide impacts, causing several distance relay misoperations due to unstable power swings. Therefore, identifying these CDRs by the proposed method further shows its ability to identify the operations of distance relays due to unstable power swings at locations far from the initial event.

The results of stability studies performed for Case 1, Case 2, and Case 3 are shown in Figure 11 (a), (b), and (c), respectively. Comparing Figure 11a,b reveals that if the distance relays are not modeled in stability studies, the studies cannot capture the actual response of the system. Additionally, as seen in Figure 11c, the response of the system in the case of using the proposed method is similar to that in Case 2 shown in Figure 11b. This further shows the effectiveness of the method in capturing the dynamic response of the system.

Figure 11. The RRA of a set of generators in Contingency 3: (a) Case 1, (b) Case 2, and (c) Case 3.
Table 7. The list of all distance relay operations in Contingency 3.

|                         | The Reference Case | The Proposed Method |
|-------------------------|--------------------|--------------------|
|                         | Time (s)           | Relay              | Time (s)           | Relay              |
|                         | 3.075              | Lines 15           | 3.075              | Lines 15           |
|                         | 3.283              | Line 17            | 3.283              | Line 17            |

Analyzing these contingencies shows that even under different topologies and operating conditions of the system, the method can correctly find the CDRs and capture the precise response of the system for any type of contingency. To further show that to what extent the method reduces the number of dynamic models included in transient stability studies to represent distance relays, Table 8 is provided. Table 8 shows the number of CDRs in each contingency. It also shows that what portion of the distance relays in the reference case are identified as CDRs in each contingency. Table 8 shows that less than 3.23% of the distance relays in the reference case are detected as CDRs in each contingency. Therefore, the dynamic models of this small number of CDRs with their updated settings can easily be obtained and included in the dynamic file of the system without violating the current limitation of stability software tools. This also significantly decreases the maintenance burden of keeping the distance relay setting information updated in the dynamic file and the computational burden of the computing system.

Table 8. The total number of the identified CDRs.

| Contingency | Total Number of the Identified CDRs | Percentage of the Total Number of Distance Relays |
|-------------|------------------------------------|-----------------------------------------------|
| 1           | 37                                 | 3.23%                                         |
| 2           | 35                                 | 3.05%                                         |
| 3           | 26                                 | 2.27%                                         |

For the contingencies studied in this section, the processing time of the three major processes performed in the proposed method to identify the CDRs is provided in Table 9. These three major processes are as follows:

Table 9. The processing time of the proposed method.

| Contingency | Process 1 (s) | Process 2 (s) | Process 3 (s) | Total (s)   |
|-------------|---------------|---------------|---------------|-------------|
| 1           | 108.95        | 1.2           | 0.008         | 110.16      |
| 2           | 108.70        | 1.2           | 0.008         | 109.91      |
| 3           | 22.46         | 1.2           | 0.008         | 23.67       |

- Process 1: Performing the initial early-terminated transient stability analysis of the contingency under study.
- Process 2: Extracting the required features of the ML model from the results of the stability study and pre-processing them to be fed into the trained ML model.
- Process 3: Predicting the operation (or no-operation) of all the distance relays in the system.

The processing time of performing stability studies depends on the type of the disturbance being studied as well as the operating condition and topology of the system. Therefore, as seen in Table 9, the processing time for performing the initial transient stability study (Process 1) is different for each contingency, whereas the processing time for extracting the features and the prediction time of the ML model (Processes 2 and 3) are the same for all the contingencies. The total processing time for any contingency studied is less than 111 s. The three contingencies studied in this section are among the most severe disturbances of the WECC system. Hence, solving transient stability studies of these contingencies requires more processing time than other contingencies. The fact that the
processing time of the method for these contingencies is less than 111 s guarantees that the processing time for any other contingency is also very small. This shows that the method can be used in planning studies to promptly identify the CDRs for a large number of the contingencies that are required to be studied.

7. Conclusions

This paper proposes a fast ML-based method to find the CDRs that should be modeled in transient stability studies of different types of contingencies to achieve a precise assessment of the dynamic response of the system during the contingency. The method is based on training an ML model to learn the latent pattern between the impedance trajectories observed by distance relays at the time of fault and 1 cycle after that during a contingency and their operations for up to 10 s later. The RF classifier is used as the machine learning algorithm. After being trained using the dataset created from extensive offline transient stability studies and the records of historical outages, the RF model can predict the operations of distance relays during any contingency under study. To do so, the model uses the dynamic response of the system captured from the early-terminated transient stability study of that contingency. The test system is the 2018 summer peak load data of the WECC system. Testing the performance of the trained model on the entire dataset, including over 900 contingencies, reveals that the model has a great performance in terms of the Recall (above 0.98) and Precision (above 0.73) metrics. To show in more detail how the method can identify the CDRs for any contingency under study and accurately represent the system dynamic response, three contingencies are studied under new topologies and operating conditions of the system.

The results show that not including the dynamic models of distance relays in stability studies can result in an inaccurate assessment of the system behavior. Additionally, the results show the effectiveness of the method in detecting the CDRs and capturing the precise behavior of the system. It is observed that compared to Case 2 (the reference case), the method requires modeling a very small number of distance relays in transient stability studies (less than 3.23% of all the distance relays in Case 2 for any of the contingencies studied). Additionally, it is observed that the total processing time for any of the studied contingencies is below 111 s, which illustrates the high speed of the method in finding the CDRs in the system.

The small number of CDRs identified by the method can be easily modeled in stability studies without exceeding the limitations of stability software. Additionally, using this method, only CDRs are required to be accurately tracked for changes in their settings, which considerably reduces the maintenance burden. Finally, by reducing the number of distance relays models included in the dynamic file, the method significantly reduces the computational burden of the computing system. Thus, more dynamic models can be included in the analysis of various contingencies.

One drawback of the proposed method is that with the passing of time and as the power system evolves (e.g., the load of the system increases, the topology of the systems changes, new distributed generations are added to the system, etc.), the system dynamic response to the contingencies might change. In this case, the training set that is used for training the RF model might not correctly reflect the behavior of the system and the protective relays during the contingencies. Therefore, the number of wrong predictions of the trained model might increases. However, the proposed method can be implemented with reinforcement learning algorithms, where the trained model continuously updated with adding new samples to the dataset (new transient stability studies performed on the updated system condition).

Similar strategies can be applied to identify other critical protective relays, such as generator protective relays, that play significant roles in defining the system dynamic response during a contingency and are needed to be modeled in transient stability studies. Hence, developing similar methods for finding the critical protective relays other than distance relays can be a proper continuation of this research.
Author Contributions: Methodology, R.V. and M.K.; Software, R.V.; Validation, M.K.; Formal analysis, R.V.; Investigation, R.V.; Writing—original draft, R.V.; Writing—review & editing, R.V. and M.K.; Visualization, R.V.; Supervision, M.K.; Project administration, M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Salt River Project (SRP) utility and are available from Khorsand with the permission of SRP.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

| Abbreviation | Description |
|--------------|-------------|
| CAPE         | Computer-Aided Protection Engineering |
| CDR          | Critical Distance Relay |
| COI          | California Oregon Intertie |
| DT           | Decision Tree |
| ISO          | Independent System Operator |
| LOS          | Loss Of Synchronism |
| MDT          | Maximum Depth of the Trees |
| ML           | Machine Learning |
| NTF          | Number of Trees in the Forest |
| PSLF         | Positive Sequence Load Flow analysis |
| PSS/E        | Power System Simulator for Engineering |
| RF           | Random Forest |
| RRA          | Relative rotor angle |
| UFLS         | Underfrequency Load Shedding |
| UVLS         | Undervoltage Load Shedding |
| WECC         | Western Electricity Coordinating Council |

References

1. Perez, L.; Flechsig, A.; Venkatasubramanian, V. Modeling the protective system for power system dynamic analysis. *IEEE Trans. Power Syst.* 1994, 9, 1963–1973. [CrossRef]
2. Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations; U.S.–Canada Power System Outage Task Force: Ottawa, ON, Canada, 2004.
3. Elizondo, D.; de La Rue, J.; Phadke, A.; Horowitz, S. Hidden failures in protection systems and their impact on wide-area disturbances. In Proceedings of the IEEE Power Engineering Society Winter Meeting, Columbus, OH, USA, 28 January–1 February 2001; pp. 710–714. [CrossRef]
4. Abdi-Khorsand, M.; Vittal, V. Identification of Critical Protection Functions for Transient Stability Studies. *IEEE Trans. Power Syst.* 2017, 33, 2940–2948. [CrossRef]
5. Vakili, R.; Khorsand, M.; Vittal, V.; Robertson, B.; Augustin, P. An Algorithmic Approach for Identifying Critical Distance Relays for Transient Stability Studies. *IEEE Open Access J. Power Energy* 2021, 8, 107–117. [CrossRef]
6. Samaan, N.A.; Dagle, J.E.; Makarov, Y.V.; Diao, R.; Vallem, M.R.; Nguyen, T.B.; Kang, S.W. Modeling of protection in dynamic simulation using generic relay models and settings. In Proceedings of the 2016 IEEE Power and Energy Society General Meeting (PESGM), Boston, MA, USA, 17–21 July 2016; pp. 1–5. [CrossRef]
7. Chatterjee, P.; Khorsand, M.; Hedman, K.W. Enhanced Assessment of Power System Behavior during Multiple Contingencies. In Proceedings of the 2018 North American Power Symposium (NAPS), Fargo, ND, USA, 9–11 September 2018; pp. 1–6. [CrossRef]
8. Nedic, D.P. Simulation of Large System Disturbances. Ph.D. Dissertation, Department of Electrical and Electronic Engineering, University of Manchester Institute of Science and Technology, Manchester, UK, 2003.
9. Pourbeik, P.; Kundur, P.S.; Taylor, C.W. The anatomy of a power grid blackout. *IEEE Power Energy Mag.* 2006, 5. [CrossRef]
10. ISO New England Operating Procedure No. 24—Protection Outages, Settings and Coordination. ISO New England, February 2019. Available online: https://www.iso-ne.com/static-assets/documents/2019/02/op24_rto_final.pdf (accessed on 30 September 2022).
11. Lavand, S.A.; Soman, S.A. Predictive Analytic to Supervise Zone 1 of Distance Relay Using Synchrophasors. *IEEE Trans. Power Deliv.* 2016, 31, 1844–1854. [CrossRef]
12. Abdi-Khorsand, M.; Vittal, V. Modeling Protection Systems in Time-Domain Simulations: A New Method to Detect Mis-Operating Relays for Unstable Power Swings. *IEEE Trans. Power Syst.* 2016, 32, 2790–2798. [CrossRef]

13. Vittal, V.; Lolliard, S.; Bose, A.; Khorsand, M.; Klaei, I. Evaluation of Protective Relay Dynamic Response via a Co-Simulation Platform; PSERC: Chandigarh, India, 2017.

14. Kirsch, D.S.; Nedic, D.P. Consideration of hidden failures in security analysis. In Proceedings of the 14th Power Systems Conference, Seville, Spain, 24–28 June 2002; pp. 24–28.

15. Tamronglak, S. Analysis of Power System Disturbances Due to Relay Hidden Failures. Ph.D. Dissertation, Department of Electrical Engineering, Virginia Tech, Blacksburg, VA, USA, 1994.

16. Salehimehr, S.; Taheri, B.; Faghihlou, M. Detection of power swing and blocking the distance relay using the variance calculation of the current sampled data. *Prot. Control Mod. Power Syst.* 2021, 104, 913–927. [CrossRef]

17. Hosseini, S./A.; Taheri, B.; Abyaneh, H.A.; Razavi, F. Comprehensive power swing detection by current signal modeling and prediction using the GMDH method. *Prot. Control Mod. Power Syst.* 2021, 6, 15. [CrossRef]

18. Taheri, B.; Hosseini, S./A.; Askarian-Abyaneh, H.; Razavi, F. Power swing detection and blocking of the third zone of distance relays by the combined use of empirical-mode decomposition and Hilbert transform. *IET Gener. Transm. Distrib.* 2020, 14, 1062–1076. [CrossRef]

19. Jafari, R.; Moaddabi, N.; Esfandari-Nasab, M.; Gharehpetian, G.B.; Naderi, M.S. A Novel Power Swing Detection Scheme Independent of the Rate of Change of Power System Parameters. *IEEE Trans. Power Deliv.* 2014, 29, 1192–1202. [CrossRef]

20. Shrestha, B.; Gokaraju, R.; Sachdev, M. Out-of-step protection using state-plane trajectories analysis. *IEEE Trans. Power Del.* 2013, 28, 1083–1093. [CrossRef]

21. Paula, H.; Pereira, C.S.; De Conti, A.; Morais, A.P.; Silveira, E.G.; Andrade, J.S. Rotating negative-sequence phasors for blocking and unblocking the distance protection during power swings. *Electr. Power Syst. Res.* 2021, 202, 107554. [CrossRef]

22. Tekdemir, I.G.; Alboyaci, B. A novel approach for improvement of power swing blocking and deblocking functions in distance relays. *IEEE Trans. Power Deliv.* 2017, 32, 1986–1994. [CrossRef]

23. Desai, J.; Makwana, V. Power Swing Blocking Algorithm based on Real and Reactive Power Transient Stability. *Electr. Power Components Syst.* 2021, 48, 1673–1683. [CrossRef]

24. Jannati, M.; Mohammad, M. A novel fast power swing blocking strategy for distance relay based on ADALINE and moving window averaging technique. *IET Gener. Transm. Distrib.* 2020, 15, 97–107. [CrossRef]

25. Ghalesefidi, M.M.; Ghaffarzadeh, N. A new phaselet-based method for detecting the power swing in order to prevent the malfunction of distance relays in transmission lines. *Energy Syst.* 2019, 12, 491–515. [CrossRef]

26. Eltamaly, A.M.; Elgharfa, A.N.A. Modeling of distance protection logic for out-of-step condition in power system. *Electr. Eng.* 2017, 100, 1891–1899. [CrossRef]

27. Yang, B.; Vittal, V.; Heydt, G.T. Slow coherency based controlled islanding: A demonstration of the approach on the August 14, 2003 blackout scenario. *IEEE Trans. Power Deliv.* 2006, 21, 1840–1847. [CrossRef]

28. Xu, G.; Vittal, V. Slow Coherency Based Cutset Determination Algorithm for Large Power Systems. *IEEE Trans. Power Syst.* 2009, 25, 877–884. [CrossRef]

29. Xu, G.; Vittal, V.; Meklin, A.; Thalman, J.E. Controlled Islanding Demonstrations on the WECC System. *IEEE Trans. Power Syst.* 2010, 26, 334–343. [CrossRef]

30. Kamwa, I.; Pradhan, A.K.; Joos, G. Automatic Segmentation of Large Power Systems Into Fuzzy Coherent Areas for Dynamic Vulnerability Assessment. *IEEE Trans. Power Syst.* 2007, 22, 1974–1985. [CrossRef]

31. Kamwa, I.; Pradhan, A.K.; Joos, G.; Samantaray, S.R. Fuzzy Partitioning of a Real Power System for Dynamic Vulnerability Assessment. *IEEE Trans. Power Syst.* 2009, 24, 1356–1365. [CrossRef]

32. Sun, K.; Likhat, B.; Vittal, V.; Kolluri, V.S.; Mandal, S. An online dynamic security assessment scheme using phasor measurements and decision trees. *IEEE Trans. on Power Syst.* 2007, 22, 1935–1943. [CrossRef]

33. Suthaharan, S. Support Vector Machine. In *Machine Learning Models and Algorithms for Big Data Classification*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 25, pp. 197–227.

34. Carabine, F. “Gini Impurity.” Learn Data Science. Available online: https://www.learnrdatasci.com/glossary/gini-impurity/ (accessed on 30 September 2022).

35. Blau, G.; Scornet, E. A Random Forest Guided Tour; Springer: Berlin/Heidelberg, Germany, 2016; Volume 25, pp. 197–227.

36. Gao, W.; Zhou, Z. Toward convergence rate analysis of random forest for classification. In Proceedings of the 34th conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 6–12 December 2020.

37. Vakili, R.; Khorsand, M. Enhancing Situational Awareness: Predicting Under Frequency and Under Voltage Load Shedding Relay Operations. In Proceedings of the 2021 North American Power Symposium (NAPS), College Station, TX, USA, 14–16 November 2021; pp. 1–6. [CrossRef]

38. Vakili, R.; Khorsand, M. Machine-Learning-based Advanced Dynamic Security Assessment: Prediction of Loss of Synchronism in Generators. In Proceedings of the 2020 52nd North American Power Symposium (NAPS), Tempe, AZ, USA, 11–13 April 2020; pp. 1–6. [CrossRef]
39. Olsun, D.; Delen, D. Advanced Data Mining Techniques; Springer: Berlin/Heidelberg, Germany, 2008; p. 138.

40. Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. A Practical Guide to Support Vector Classification; Tech. Rep.; Department of Computer Science, National Taiwan University: Taipei, Taiwan, 2003; Available online: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (accessed on 30 September 2022).

41. Chicco, D. Ten quick tips for machine learning in computational biology. BioData Min. 2017, 10, 35. [CrossRef] [PubMed]

42. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 2012, 13, 281–305.

43. Fushiki, T. Estimation of prediction error by using K-fold cross-validation. Stat. Comput. 2009, 21, 137–146. [CrossRef]

44. Mosteller, F.; Tukey, J.W. Data analysis, including statistics. In Handbook of Social Psychology; Addison-Wesley: Reading, MA, USA, 1968.

45. PSLF User’s Manual, PSLF Version 18.1-01; General Electric: Boston, MA, USA.

46. Random Forest Classifier, Scikit Learn Developers. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed on 30 September 2022).