Classification of Fault Location and Prediction of Duration Using Random Forest Regressor (RFR) Model in Power Systems

Zakaria El Mrabet
School of Electrical and Computer Sciences
University of North Dakota
Grand Forks, ND, USA
zakaria.elmrabet@und.edu

Prakash Ranganathan
School of Electrical and Computer Sciences
University of North Dakota
Grand Forks, ND, USA
prakash.ranganathan@und.edu

Shrirang Abhyankar
Electricity Infrastructure and Buildings Division
Pacific Northwest National Laboratory
Richland, WA, USA
shrirang.abhyankar@pnnl.gov

***Abstract*—Accurate detection of faults is critical for the seamless operation of the power grid. Localizing the faults is typically recorded by digital fault recorders (DFRs) and often challenging to capture the fault type, severity (e.g., magnitude), location, and its duration (e.g., periods or number of cycles). Specifically, this work investigated the suitability of multiple machine-learning-based approaches to identify fault location and estimate its length. With the increasing deployment of Phasor Measurement Units (PMUs) and microPMUs in both transmission and distribution network, a large amount of data is being generated, dumped into data storage, and electric utilities facing challenges to mine such Big Data. This paper investigates a Random Forest Regressor (RFR) algorithm, a machine learning model for detecting fault location and predicting its duration. Three cases were studied to evaluate the performance of the model separately to detect: 1. fault location (case 1); 2. fault duration (case 2); and 3. Both location and length of the fault in the real-time streaming environment (case 3). The preliminary results indicate that RFR can detect the fault with an overall accuracy of 65.2% and predict their duration with an error rate of 0.6s. The performance of RFR algorithm is compared against other machine learning models namely, Deep Neural Network (DNN), HAT, Neural Network (NN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayesian (NB), and k-nearest neighborhood (KNN) models.**

***Index Terms*—Fault location classification; Fault duration prediction; Random Forest Regressor; Deep Neural Network; GridPACK; Accuracy; MSE; MAE;**

I. Introduction

Identification of faults are critical for seamless operation of power systems. Utilities are working 24/7 to reduce outage rates that may arise due to contact with natural vegetation (e.g. a tree), animal, or weather events [1], [2], [3]. According to [4], the cost experienced by an "average" consumer for an outage of one hour summer afternoon was estimated to be roughly $3 for a typical customer, $1200 for a small and medium organizations, and $82,000 for a large organization. The cost increases substantially, as outage duration increases from one to eight hours. Furthermore, these costs are usually higher in winter than summer for the same outage duration. Restoring power back to the grid takes time and requires cooperation among line crews, consumer and utility. To protect and prevent the potential damages to people, equipment, and environment, advanced computational algorithms are needed to track, and locate and isolate the faults promptly.

Physical faults in power systems are generally classified into balanced and unbalanced faults [5], [6]. An unbalanced fault, also known as asymmetrical fault, is a commonly occurring fault in the power system and can be of a series or shunt type. In the series fault type, the voltage and frequency values increase, while the current level decreases at the faulty phases. In shunt fault, the current level rises, while the frequency and voltage levels decrease at the faulty phases. There are several shunt fault types: single line to ground (SLG), line to line (LL), double line to ground (DLG), and Three-phase to the ground (LLL). A SLG occurs when a transmission line phase touches a neutral wire or the ground; the DLG or LL fault occurs when two or more phases make a connection with the ground; The cause of this fault is mostly due to natural weather events to storms or high winds or fallen trees. Some times, the cause can be due to an equipment failure, line connecting the remained phases, or a failing tower.

According to [6], [7] the likelihood of occurrence of each fault type is 70% for SLG, 15% for LL, 10% for DL, and 5% for LLL. Although there is a low likelihood of LL fault occurrence, it is considered a severe fault that can cause a rise in fault current magnitude and there by resulting in outages or large damage to grid assets. Hence, necessitates the need for a fault detection and location identification model. This paper focuses on the detection of LL fault and its duration.

There are several fault detection approaches for power systems reported in literature. For example, authors in [5] provide a review of conventional and machine-learning based techniques used for fault location identification. Some conventional approaches include: the traveling wave method, impedance-based method, and Synchronize voltage and current based measurements. The traveling wave approach needs high-speed data acquisition equipment, a Global Positioning System (GPS), sensors, and transient fault recorder to detect the transient waveform for fault location. The location of the

TABLE I
EXISTING LITERATURE ON FAULT LOCALIZATION IN POWER SYSTEM

Category	Approach	Fault types	Advantages	Limitations	References
Conventional	Impedance-based	Physical	Ease of Implementation	The accuracy can be affected in case of a grounded fault where the fault resistance is high The fault duration was not considered	[6], [8]
	Time-wave based	Physical	Large Resistance, load variance, grounding resistance, reflection, and refraction of the traveling wave and series capacitor bank.	The accuracy depends on the correctness of the line parameters'estimated values, including capacitance and line inductance. The fault duration was not considered	[9]
Machine learning	NN + Levenberg-Marquardt	Physical	The detection error is less than 3%. High tolerance to the fault resistance, fault type, fault location, and the embedded remote-end source.	The convergence time for the training process is high. The fault duration was not considered	[10], [11]
	NN to estimate the fault distance of substations	Physical	Optimal results in terms of estimating the fault distance from the sub-stations even under network-topological changes. High tolerant to noise	Inappropriate for detecting fault location in streaming power system data.	[12]
	CNN based on bus voltages	Physical	Optimal localization estimation even under low visibility (7% of buses)	The fault duration was not considered.	[13]
	Random Forest+Decision tree	Physical	Fault location detection accuracy is 91% with minimum number of buses (5-7%)		[14]
	PMU anomaly detection+ MLE + DBSCAN	Physical	The proposed data cleansing approach outperforms Chevyshev and K means and achieve a precision of 95%. Less than 0.9s to classify event for a typical window size of 30 sample data.		[15]
	KNN	Physical	Fault location accuracy reaches 98.70% with an error between 0.61% and 6.5%.	The proposed model was trained/tested on the PV system only.	[16]
	HAT with DDM and ADWIN for classifying traditional and cyber contingencies in real-time	Physical and cyber fault	Classification accuracy is greater than 94% for multiclass and greater than 98% for binary class. Adaptable to the concept of drift events.	The fault location and duration were not considered.	[17], [18]
Hybrid	Wavelet transform and SVM	Physical	The fault classification error is below 1% for all fault types. The overall accuracy is 0.26% for SLG, 0.74% for LLG, 0.20% for LL, and 0.39% for LLLG.	The fault duration was not considered Not suitable for streaming power system data. The accuracy of the SVM depends on selecting and tuning of the appropriate kernel type and hyper-parameters.	[19]
	Wavelet analysis + K-means + ELE	Physical	Fault location accuracy attain 100%	The fault duration was not considered.	[20]
	Wavelet analysis + Fuzzy logic	Physical	The error between the actual fault location and the predicted one is low then 0.002%		[21]
	Discrete wavelet transform + SVM	Physical	Fault location accuracy is 98.27% for IEEE 13-Bus and and 98.29% for the IEEE 34-Bus test systems		[22]

fault is computed by tracking "time-tagging the arrival of the traveling wave at each end of the line and comparing against the time difference to the total propagation time of the line with the help of GPS" [9]. This approach has several advantages, as the approach is not impacted by excessive resistance, load variance, reflection, grounding resistance, refraction of the traveling wave or series capacitor bank [5]. However, the accuracy of the approach relies on capacitance and line inductance. Unlike the time wave method, the impedance-based approaches [6], [8] are simple and easy to implement, as it require only measurement data that include fault voltages and fault currents collected from the digital fault recorder or relays to compute the impedance. The accuracy of this approach can be affected in the case of a grounded fault, where the fault resistances can reach higher values.

There are machine learning (ML) based approaches reported in literature for fault locations. In these ML approaches, training data is generated using inputs such as voltage, current, phase angle, and fault location as output.

For example, authors in [10] proposed a back propagation based neural network (BPN) to estimate fault location in distribution networks. Here, fault current was selected as a key feature to train the NN model. A Levenberg-Marquardt algorithm (also known as damped lease sqaure) is applied to BPN for faster convergence. Then, the BNN model was deployed to run on the DIgSILENT Power Factory 13.2. Similarly, a feed-forward NN (FNN) based approach is proposed in [11]. Here, fault voltages and fault currents are selected as two features to train the model. A sigmoid activation function was used to normalize the data. Their results showed a detection error of less than 3%. Another NN based approach was proposed in [12] to estimate fault distances from substation(s). The selected inputs features includes: three-phase voltage, current, fault conditions, and active power gathered from substation(s). This approach was trained on different fault locations, resistances, and loads. The approach was tested on an IEEE 34-bus system and yielded in promising results, even under dynamic changes in network topology. Additionally, this approach showed more tolerance to noise.

A hybrid method using wavelet transform and Support Vector Machine (SVM) has been published in [19] to locate faults in transmission lines and can be described in three stages. In first stage, voltage and current values emitted by a transmitter were used to locate the fault; the second phase feeds a multi-class SVM model to the training based on selected influential features; and classification of fault location is done using a regression approach. Here, the fault classification error is below 1% for all fault types and specifically 0.26% for SLG, 0.74% for LLG, 0.20% for LL, and 0.39% for LLLG.

Although high accuracy were reported for NN based approaches in the above-mentioned studies, the training time required for NN is longer does not suit for dynamic or real-

time environments. On the contrary, the SVM based approach is faster and relatively accurate, even for larger size data. However, it requires careful selection of appropriate kernel type and hyper-parameters.

In [13], authors proposed a convolutional neural network (CNN) based approach using bus voltages. This method has been trained and tested on an IEEE 39-bus and IEEE 68- bus systems under uncertain conditions for system observability and measurement quality. Their results show that CNN can localize the faulted line even in low visibility (7% of buses) conditions. A KNN based approach for detecting faults in photovoltaic (PV) system is proposed in [16]. This approach has been trained and tested on data generated from a developed PV model. The reported results show a classification accuracy of 98.70% with an error value ranging between 0.61% and 6.5%. Authors in [15] proposed a real-time event classification and fault localization approach for synchrophasor dataset. Their methodology relies on three processes. The first process focus on removing bad data from collected PMU measurements using the Maximum Likelihood Estimation (MLE) approach. In second process, the events are classified using a combination of Density-based spatial clustering of applications with noise (DBSCAN) and logic rules were generated using a physics-based decision tree (PDT) method. This PDT method uses parameters such as active power, reactive power, and fault event types. The third process reports localizing events in real-time using a graph-theory. Finally, a score metric is computed using shannon entropy, and descriptive statistical parameters (e.g., standard deviation, range, mean difference, and crest factor). Three case studies have been considered using metrics such as precision and recall and their reported result show that their proposed data cleansing approach outperforms Chebyshev and K-means methods, with a 95% precision. Additionally, the average run-time taken for their classification algorithm is around 0.09s for a typical window size of 30 samples involving five PMU sensors.

Similarly, authors in [20] have proposed an event location estimation (ELE) algorithm for the wide-area monitoring system for PMU data. Their approach relies on clustering and wavelet analysis to detect and localize events in real-time. In this work, the network is initially divided into several clusters, where each cluster is defined as an electrical zone (EZ) using K-means. Next, a wavelet-based event detection approach is used to detect and localize event occurrences by tracking any large (e.g. event magnitudes) disturbance levels. Once the event is detected, its magnitude is defined using a Modified Wavelet Energy (MWE) value, and its location is estimated at each EZ's. The authors implemented the ELE approach in real-world PMU-setting containing 32 dynamic events with an excellent localizing accuracy values. It is important to note that the authors did not consider data quality issues in the PMU measurements. Some probable causes for data quality issues could be irregular sampling or data rate, bandwidth challenges, and time synchronization errors.

In [21], the authors discusses a wavelet decomposition technique combined with fuzzy logic to identify both faulty line(s) and its locations in a multi-terminal high voltage direct current (MTHVDC) network. In their paper, wavelet coefficients of both positive and negative currents were initially computed and then fed to a fuzzy logic based voting system to identify the faulty line(s). Once the line is identified, a traveling wave-based algorithm is used to determine the exact fault location using the Daubechies wavelets.

A discrete wavelet transform (DWT) combined with SVM for fault detection in distribution networks has been proposed in [22]. Here, features are extracted using SVM and Decision Trees (DT), and then optimized using a genetic algorithm (GA). Their model performance was evaluated on two active distribution networks (e.g., IEEE 13-Bus and IEEE 34-Bus systems), and the authors claim that their model outperforms the probabilistic neural network (PNN).

In [14], authors discusses a fault line identification and localization approach using random forest and decision tree classifiers. Here, the models were trained on an IEEE 68-Bus system. The generated data consists of seven fault types including Three-phase short circuit (TP), Line to ground (LG). The reported experiment results show that classification accuracy of 91%.

Table I provides a summary of the relevant fault detection approaches along with their advantages and their potential limitations. In this paper, we propose a random forest regressor (RFR) based model to detect location of faults and predicting their durations. To the best of the author's knowledge, this is the first attempt that investigates the detection of both fault location and duration using a machine learning model. Also, RFR approach has not been applied to any power system data. We claim that RFR model is ideally suitable for streaming data applications (e.g., SCADA/PMU, demand side management, AMI, and market forecasts), as this model can be applied both as a multi-output classification and regression problems for power system data. Our applied model is unique and novel in a sense that: 1) RFR model includes an ensemble of multiple uncorrelated trees which helps in achieving good generalization by injecting randomness in training the decision trees; 2) PMU streams can process data in batches of varying window sizes, so the number of samples is not significant when looking at streams of 3-5 windows, and hence the interpretation of RFR trees is more natural to classify and predict, and 3) less chance of over-fitting due to bagging and random feature selection, and 4) small number of tuning parameters.

Detecting fault location can be approached as a multi-class classification problem since the output, which is in this case the fault location, is in form of discrete values while predicting the fault duration is a regression problem as the output includes continuous value. In this work, we have combined these two problems into a single regression model by mapping both fault location and its duration values into one single output in form of continuous values. For training the model, GridPACK framework [23] is used to simulate several 3-phase faults scenarios on a 9 bus system to generate the appropriate dataset.

A collection of three experiments were formulated to eval-

uate the performance of the RFR model. In experiment 1, the model is evaluated for fault detection accuracy and compared against seven classifiers, namely, Neural Network (NN), Deep Neural Network (DNN), Support Vector Machine (SVM), k-nearest neighbors (KNN), Naïve Bayes (NB), Decision Tree (DT), and Hoeffding Tree (HT). In experiment 2, the RFR model is evaluated for predicting fault duration and compared against the regression version of the models such as Support Vector Regressor and Decision Tree Regressor. The Mean Squared Error (MSE) and Mean Absolute Error (MAE) were used as evaluation metrics.

In experiment 3, the RFR model is tested in a streaming data environment, where multiple window sizes are considered. The MSE and processing time for the RFR are then compared against those of HT and DNN. The HT and DNN are commonly suggested models for power system streaming data [24], [25], [26]. Additionally, these models are evaluated in terms of dealing with missing and incomplete data which could be due to some equipment failure, data storage issue or unreliable communication.

This paper is organized into the following sections: Section II focus on RFR model description with details on simulated fault scenarios, feature selections, and training/testing process; and section III discusses the analysis of three experiment scenarios for classifying and predicting fault location and duration with off-line/streaming conditions; and Section VI draws conclusions and recommendations for future work.

II. Methodology

A. Random Forest Regressor (RFR) Model

Random forest F is an ensemble approach with several independent and un-correlated decision trees $F = \{t_1, t_2, \ldots, t_t\}$. These uncorrelated trees help the model F in achieving a good generalization by injecting randomness in training the decision trees [27]. Achieving such generalization relies on applying a bagging technique, which combines the concepts such as bootstrapping and aggregation [27]. Given a training set $S = \{X^m, Y^m\}_{(M=1)}^m$, where $X \subset R^D$ and consists of input feature space with parameters such as voltage (v), phase angle (ϕ), current, and frequency (f). Y is a multi-dimensional continuous space $Y \subset R^{D'}$, and includes both the fault location and corresponding fault duration. M is the number of samples, bootstrap is a subset S_t of the entire training set S, where each instance has been randomly sampled using a uniform distribution with/without replacement. The resulting bootstrap data includes the same amount of instances as original data set S, but approximately $1/3$ of these samples are duplicates and approximately $1/3$ of the instances are left out of the bootstrap sample. Multiple passes are performed over input data to create bootstrap for each tree. Once training and testing are completed on the bootstrap data, then the prediction of all the independent trees are averaged as one aggregated value.

Assuming that output variables follow a multi-variate Gaussian distribution with mean and co-variance , the regression posterior can be modeled as:

$$P(y \mid x, P_t) = N_t(y \mid \mu_t, \Sigma_t) \tag{1}$$

Where P_t is a partition built by a random tree t_t and N_t is a multi-variate Gaussian with mean μ_t and co-variance Σ_t predicted in the output space Y from the subsets of the training dataset. The purpose of training the trees is to reduce the uncertainty related to the multi-variate Gaussian model. Specifically, an appropriate splitting function f must be selected to split the subset S_l of the training set. This is done at each arriving node N_l in the tree t_t to reduce any prediction uncertainty caused due to 'splitting'.

Example of function f includes information gain and Gini index. According to [28], [29], the unweighted differential entropy function, which is a continuous version of Shannon's entropy (SE) [16], is considered an optimal function for computing information gain in regression task. The SE function was selected, as it reported satisfactory results in terms of prediction error, and it is given by:

$$f(S_l) = \int_{(y \in Y)} \sum_{i=1}^{n} P(y|S_l) log(P(y|S_l)) dY \tag{2}$$

Where y is an instance of prediction that includes both fault duration and fault location. Since we model the posterior using multi-variate Gaussian, f can be rewritten as [30]:

$$f(S_l) = \frac{1}{2} log((\pi \exp)^{D'} \mid \Sigma^{(S_l)} \mid) \tag{3}$$

where $\Sigma^{(S_l)}$ is the co-variance matrix estimated from the subset S_l. After splitting the subset S_l at node N_l into two subsets nodes S_l^{right} and S_l^{left} using function f, the information gain Δ is calculated using:

$$\Delta = f(S_l) - w_l f(S_l^{left}) - w_r f(S_l^{right}) \tag{4}$$

Where $w_l = \frac{|S_l|}{|S_l^{left}|}$ and $w_r = \frac{|S_l|}{|S_l^{right}|}$. Once the training phase is completed, the prediction phase consists of sending the new received instances through the trees of the forest and the posteriors of all the trees are estimated using following equation:

$$P(y \mid x) = \frac{1}{T} \sum_{t=1}^{T} P(y \mid x, P_t) \tag{5}$$

Where T is the number of trees in the forest and P_t is the partition introduced by t_t, and given any new instance, the model can predict its corresponding fault duration and location by maximizing a posterior:

$$\hat{Y} = argmax_{y \in Y} P(y \mid x) \tag{6}$$

B. Dataset

The simulated fault scenarios were run using GridPACK software, an open source framework designed to support the development and implementation of power grid applications. Examples of these applications include power flow simulation of the electric grid, contingency analysis of the power grid, state estimation based on electric grid measurements, and dynamic simulation of the power grid. These applications are

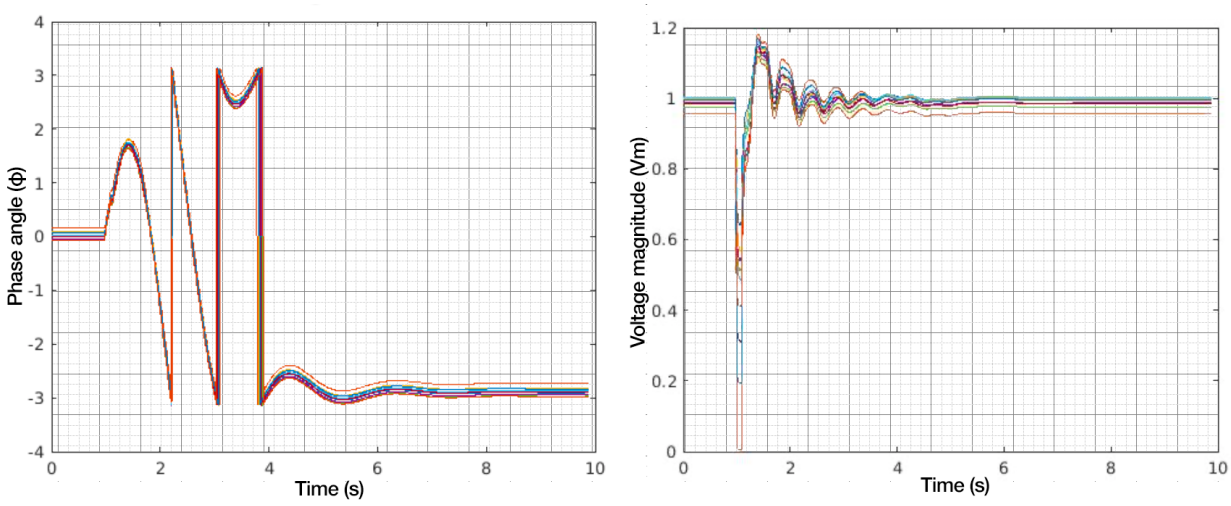

Fig. 1. (a) Phase angle of 9 buses after applying fault, (b) V_m Voltage magnitude of 9 buses after applying fault

TABLE II
COMMON THREE-PHASE FAULT MODELING FOR 9 SCENARIOS WITH DIFFERENT DURATION

Scenario	Fault location	Fault duration	Simulation time	Number of generated sample for each fault duration	Number of generated samples for each scenario
Scenario 1-9	Apply fault at Bus 1-9	0.05s	10s	594 samples	5945 samples/scenario.Total number of samples is 53512
		0.1s	10s	594 samples	
		0.15s	10s	594 samples	
		0.2s	10s	594 samples	
		0.25s	10s	594 samples	
		0.3s	10s	594 samples	
		0.35s	10s	594 samples	
		0.4s	10s	594 samples	
		0.45s	10s	594 samples	
		0.5s	10s	594 samples	

capable of running on high-performance computing architecture (HPC) [12].

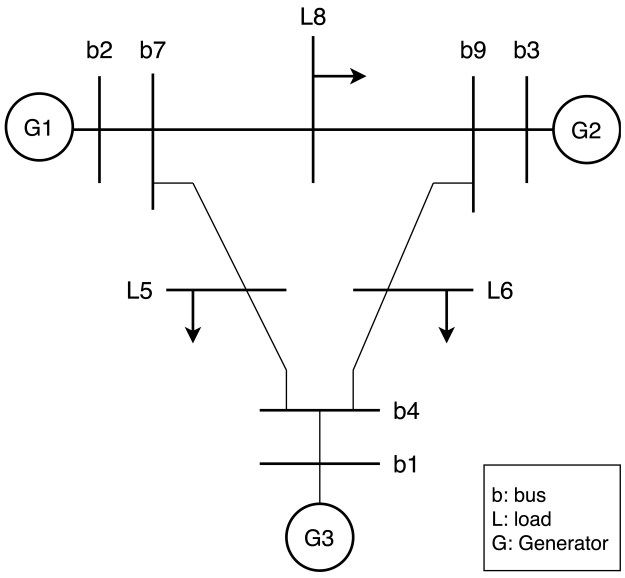

Fig. 2. IEEE 9 bus system with 3 generators

The dynamic simulation application package in GridPACK was selected to simulate a three-phase fault at various bus locations with different fault duration(s) using a nine bus system, as shown in Figure 2. The duration of faults is varied from 0.05 to 0.5 seconds. Besides, fault strength levels (i.e., magnitude) were also varied. An example of scenario 1 is given in Figure 1. Three features were selected to capture both the fault location and duration: the voltage magnitude (V_m) at each bus, the phase angle (ϕ) at each bus, and the frequency (f) of the generators. The timing of the fault applied to each bus is ten seconds. The total number of samples for all simulated scenarios is 53,512 samples. A summary of the training and test data are listed in Table II. Additionally, Figure 3 illustrates the data preparation, training, and testing process using the RFR model.

III. EXPERIMENTAL RESULTS AND DISCUSSION

Three experimental scenarios are considered to evaluate the performance of the RFR model. In experiment 1, the proposed model was assessed based on the accuracy metric, which is the ratio of the correctly classified fault location cases over the total number of cases. In other words, accuracy metric can

be expressed as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{7}$$

Where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative. These values are obtained from the confusion matrix.

Afterward, the RFP model's accuracy is compared against seven classifiers: NN, DNN, SVM, KNN, DT, NB, HT. The tuning parameters for all models are listed in Table III.

The second set of experiments evaluates the model's performance in predicting the fault's duration. Since this feature is continuous value, the accuracy metrics cannot be used for this purpose. Thus, other performance metrics such as MAE, MSE were selected. The MAE is the average of the absolute differences between the actual and predicted fault duration, and it is given by:

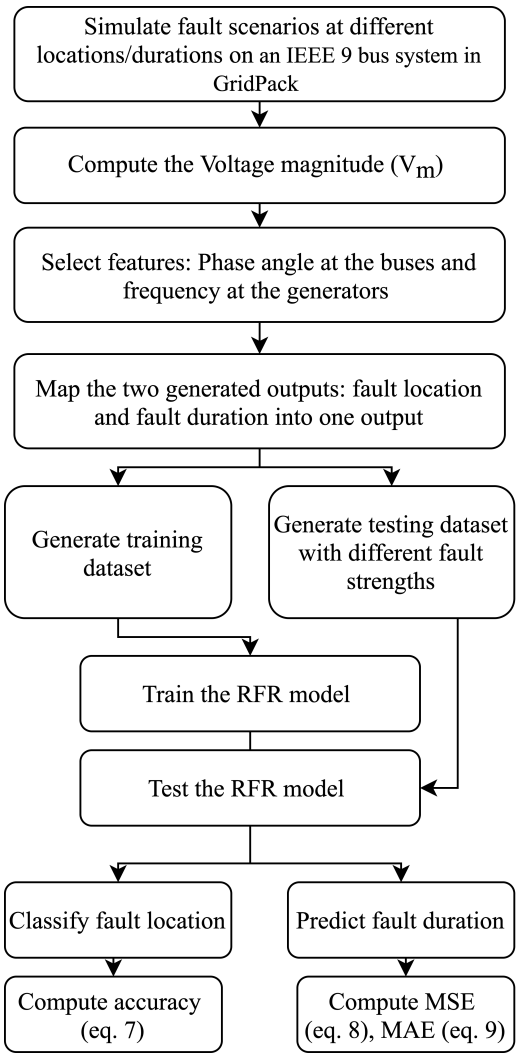

Fig. 3. RFR model for detecting fault location and duration

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid (\hat{y} - y) \mid \tag{8}$$

Where $\hat{y}$ is the predicted fault duration, y is the actual fault duration, and n is the number of instances or cases. Unlike MAE, MSE has the benefit of penalizing for significant errors since it averages the squared differences between the actual fault duration and the predicted one, and it expressed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)^2 \tag{9}$$

The MSE and MAE of the proposed model have been compared to the regression version of the models above. In the third experiment, both fault duration and location were evaluated in a streaming window environment.

A. Experiment 1: Fault location classification

The results of the experiment set 1 are shown in Figure 4. This figure shows the comparison between the proposed model (RFR) and seven models in terms of fault location detection accuracy at nine different fault locations. For instance, Figure 4 (a) illustrates the detection accuracy of a fault occurring at bus 1. RFR approach detects approximately 90% of the fault following by DNN with 78% accuracy. NB reports the poorest performance with an accuracy rate below 1%. Figure 4 (b) shows that DNN outperforms RF in terms of detecting fault on bus 2 with an accuracy difference of 20%. In Figure 4 (c), the RFR reports the highest accuracy rate, which is 90%, followed by DNN with 75%, then KNN with 30%. In Figure 4 (d), SVM has the highest accuracy value, followed by NB, then RFR. In Figure 4 (e), DT reports an accuracy of 100% followed by RFR then DNN with 90% and 70%, respectively. At bus 6, RFR detects 60% of the fault and DNN 55%. At buses 7, 8, and 9, DNN outperforms RFR with an accuracy difference of 20%, 20%, and 10%, respectively.

The performance of the DT, NB, and SVM models on the testing data suggest that they suffer from overfitting since their detection accuracy rates are very high on buses 4 and 5 and

TABLE III
THE SELECTED OPTIMAL HYPER-PARAMETERS FOR EACH MODEL

Model	Hyper-parameters
RFR model	Number of trees=10, splitting criterion=info_gain
NN	1 hidden layer, relu activation function
DNN	1 input layer, 1 output layer, 6 hidden layers with 50 hidden node each, relu activation function, epoch=500
SVM	Kernel: Radial Basis Function , C=1
KNN	Number of neighbors=5
DT	Max Depth=2
HT/HAT	Split confidence=1e-07, drift detector= ADWIN

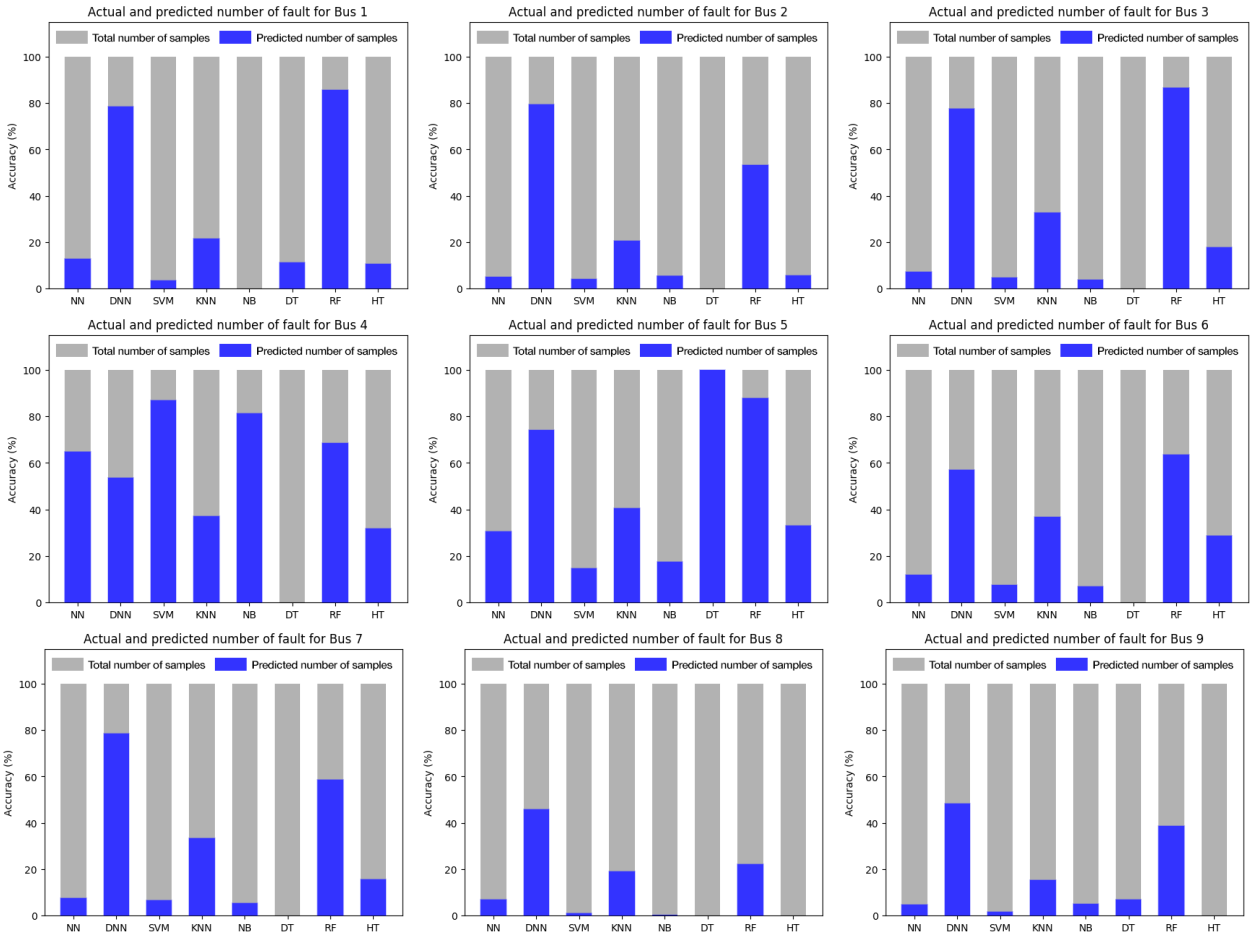

Fig. 4. Comparison between the proposed model (RF/RFR) and NN, DNN, SVM, NB, DT, HT in terms of fault location detection accuracy at various location.

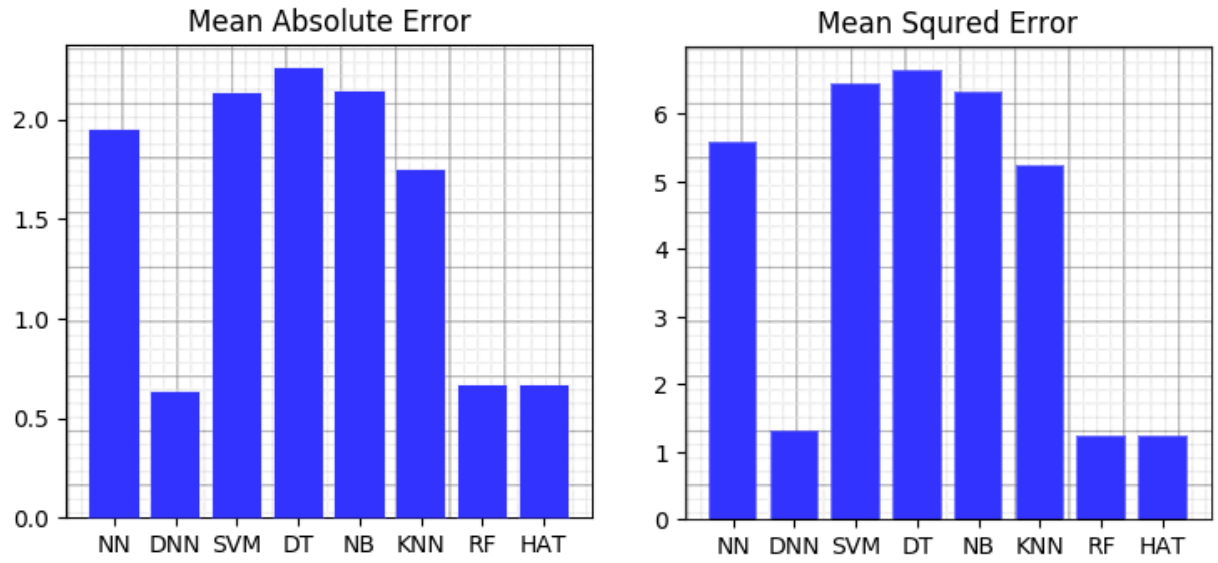

Fig. 5. (a) Comparison between the proposed model (RFR) And NN, DNN, SVM, NB, DT, HT in terms MAE, (b) Comparison between the proposed model (RFR) And NN, DNN, SVM, NB, DT, HT in terms MSE and MAE

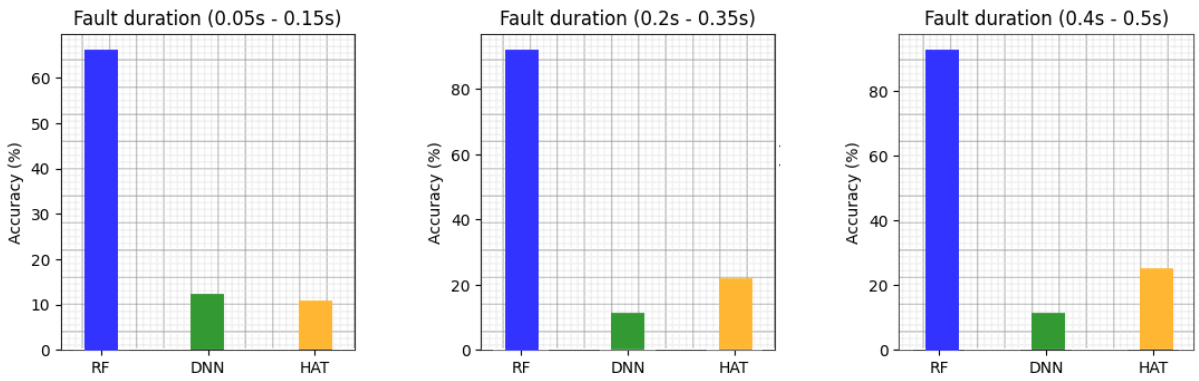

Fig. 6. Accuracy of RF, DNN, and HT in terms of predicting faults duration

poor on the other buses. On the other hand, the consistent DNN and RFR model's overall performances on the nine buses prove that they can achieve good generalization without overfitting or underfitting. Regarding the processing time for training and testing each one of these models, table IV provides the processing time along with the accuracy. Although the testing time for the NN, NB, and, DT are relatively low, these models' accuracy is under 20%. DNN reports an overall accuracy of 65% but the testing time is 0.25%. Regarding the RFR model, it reports an overall accuracy of 65.2% with low testing time, 0.046s.

B. Experiment 2: fault duration prediction

The obtained results for predicting the fault duration is seen in Figure 5 (a). The reported lowest MAE value is by models RFR, HAT, and DNN. On the other hand, the highest MAE value is 2.5 seconds reported by DT. These results suggest that DNN, HAT, RFR are the optimal model in terms of predicting the fault duration since the difference between the actual and predicted fault duration for the entire testing dataset is less than 0.6s. Figure 5 (b) shows the MSE of RFR compared to the other models. As expected, the RFR and HAT report the lowest MSE value, which is close to 1s. However, it is necessary to note that the prediction error for DNN is more than 1.5s. It is important to note the differences between MSE and MAE computations. The MSE measures squared error differences between actual and predicted fault duration. Since RFR, HAT, and DNN report the optimal results in terms of MAE and MSE, these models are selected for the next experiment.

Figure 6 illustrates a comparison between the three optimal performed models, which are DNN, RF, and HAT, in terms of detecting the fault location with three different fault duration: short fault duration ranging between 0.05s and 0.15s, medium fault duration ranging between 0.2s and 0.35s, and long fault duration ranging between 0.4s and 0.5s. As can be seen, the RFR model outperforms the DNN and HAT in terms of detecting the fault with short, medium, and long duration. In Figure 6 (a), the RFR model reports 65% accuracy in terms of detecting the short fault duration, followed by DNN with 12%, then HAT with 10%. In Figure 6 (b), the accuracy RFR increases to 88%, followed by HAT with 22%, then DNN with 10%. In Figure 6 (c), RFR reports 88% fault detection accuracy followed by HAT with 30% then DNN 10%. These results suggest that the RFR model is an appropriate model for detecting short, medium, and long fault duration. The poor results reported by DNN is because it requires more data to achieve its optimal performance. Since we split the dataset into three parts, each of which has a specific fault duration, the first one includes short fault duration with 16212 instances; the second one includes 21500 instances. The third one, for long fault duration, includes 15800 instances, training, and testing DNN on each sub-dataset that is insufficient for achieving its optimal detection accuracy. These results suggest that the RFR model can achieve its highest accuracy with a relatively small number of instances compared to DNN model, which requires a large dataset to achieve its optimal performance.

Figure 7 illustrates the MSE and MAE as a function for the percentage of missing data for the three selected models DNN, RFR, and HT. The purpose of this experiment is to evaluate the models' robustness in dealing with missing data. In a real power system network, the collected measurements, including Voltage magnitude and frequency, can be incomplete due to some equipment failure, data storage issue, or unreliable communication [31]. Thus, it is crucial to evaluate the models' capacity in predicting the fault duration accurately using incomplete data.

As seen from Figure 7, the MSE of the three models increases as the percentage of missing data increase. The RFR's MSE ranges between 2 and 7.5 for a missing data percentage of 10% and 90%, respectively. The DNN reports an MSE value of 3 for 10% of missing data, while HT reports an MSE value of 6 for the same percentage of missing data. The MSE value of the DNN and HT increases as the percentage of missing data increase to reach the highest importance, which is 10. In addition, Figure 7 illustrates the MAE as a function of the percentage of missing data. As can be seen, the evolution of the MAE value of the three models shows similar behavior, as reported in the previous figure. The RFR indicates an MAE value of 0.8, followed by DNN with 1.25, then HT with 2.1 for 10% of the missing data. The MSE values of the three

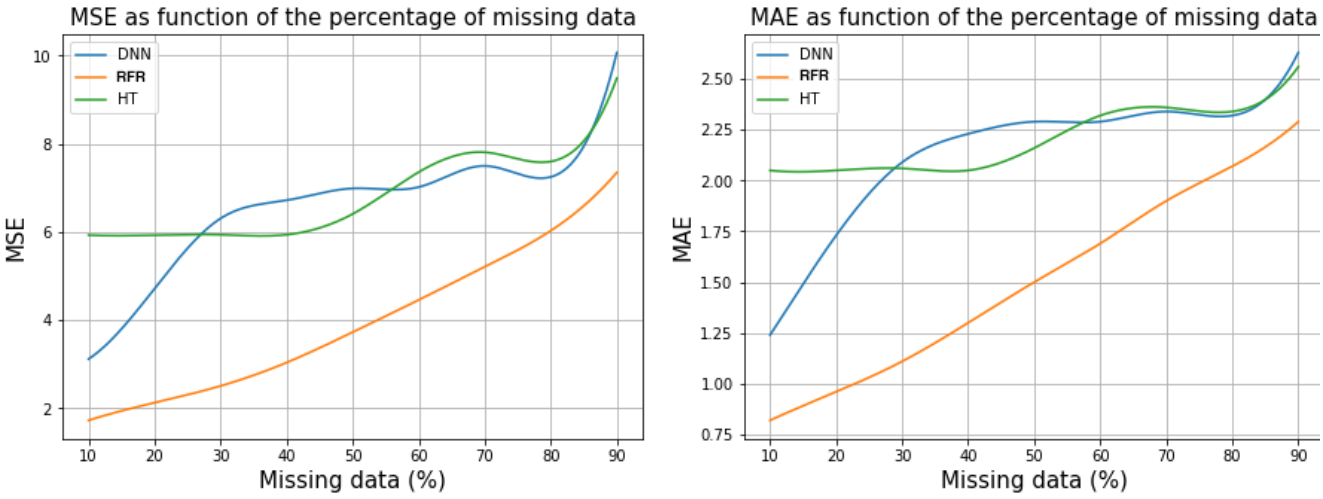

Fig. 7. MSE and MAE as a function of the percentage of missing data for the three models: DNN, HT, and RFR

models increase as the percentage of missing data increase to reach their highest values, which are 2.25 for RFR, 2.30 for HT, and DNN with 90% of missing data. The results reported in these figures suggest that the RFR model is more resilient and tolerant of the missing data. Thus, it is an optimal model for predicting the fault duration even with incomplete data.

C. Experiment 3: Fault Duration Prediction

In this experiment, the RFR, DNN, and HAT, selected from the previous experiment sets, are evaluated with streaming data. For this purpose, the models have been trained incrementally. In other words, they have not been trained and tested on the entire dataset, but they have been incrementally trained by providing one sample at a time. Then, the MSE and the processing time of each model are evaluated, as seen in Figure 8. As seen from that Figure, the MSE of RFR values is holding constant values and below 0.05s, as the number of samples increases. For HT, the MSE drops from 28 seconds to 5 seconds for samples between 0 and 2, then stagnates at 5s, before it drops to 0.05s for more than 30 samples. For DNN, the MSE value drops sharply from 30 seconds to 2s, as the number of samples increases to 30 samples, then decrease slowly to achieve its lowest value and stabilizes at 0.05s. In terms of processing time per sample, RFR reports the lowest value, which is 0.0028 ms, followed by DNN with 0.0032 ms, then HT with 0.7 ms. The results obtained in this third experiment set suggest that RFR is a potential model for detecting fault location and its duration in a near real-time streaming environment. A summary of the obtained results in three experiments are provided in Table IV. After careful examination of the studied models' performances, especially the overall accuracy, MAE, MSE, and the processing time, the overall ranking can be categorized as: high for RFR, medium for DNN, and low for the other models.

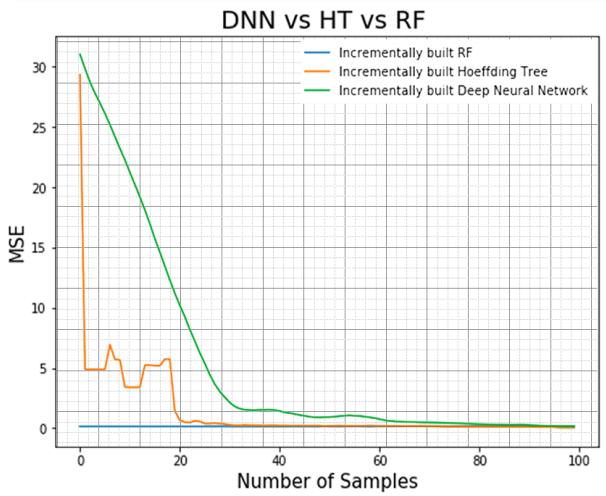

Fig. 8. Comparison between DNN, HT, and RFR in terms of MSE

IV. Conclusion

The application of a Random Forest Regression (RFR) model to power system data is investigated to detect both the fault location and predict its length. The model is trained and extensively tested on PNNL's GridPACK software, an open-source framework for power grid applications, simulating multiple fault scenarios. For each scenario, a fault is generated on a specific bus for different durations. Several machine learning models were compared with RFR for three experiment cases. The preliminary results indicate that both RFR and DNN models can detect faults with an accuracy of approximately 70%. For predicting fault durations, both the RFR and HT models yield better results. Specifically, the RFR model outperforms DNN and HT models and hence suitable to be deployed for real-time situational awareness to capture both fault location and its length.

TABLE IV
SUMMARY OF THE RFR'S PERFORMANCES COMPARED TO THOSE OF DNN, HAT, NN, SVM, DT, NB, AND KNN, OBTAINED IN THE THREE EXPERIMENTS

Experiment	Performance metrics	RFR	DNN	HAT	NN	SVM	DT	NB	KNN
1. Fault location detection	Overall accuracy	65.2%	65%	14.77%	17%	16%	13.8%	14.5%	27.6%
2. Fault duration prediction	MSE	1.1s	1.2s	1.1s	5.6s	6.5s	6.6s	6.2s	5.1s
	MAE	0.6s	0.6s	0.6s	1.9s	2.2s	2.5s	2.2s	1.8s
3. Fault duration prediction in streaming data	Processing time	0.0028 ms	0.0032 ms	0.7 ms	-	-	-	-	-
Overall ranking		High	Medium	Low	Low	Low	Low	Low	Low

V. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) Award 1537565.

REFERENCES

[1] H. Haes Alhelou, M. E. Hamedani-Golshan, T. C. Njenda, and P. Siano, "A survey on power system blackout and cascading events: Research motivations and challenges," *Energies*, vol. 12, no. 4, p. 682, 2019.

[2] M. R. Salimian and M. R. Aghamohammadi, "A three stages decision tree-based intelligent blackout predictor for power systems using brittleness indices," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5123–5131, 2017.

[3] Y. Zhang, Y. Xu, and Z. Y. Dong, "Robust ensemble data analytics for incomplete pmu measurements-based power system stability assessment," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1124–1126, 2017.

[4] L. Lawton, M. Sullivan, K. Van Liere, A. Katz, and J. Eto, "A framework and review of customer outage costs: Integration and analysis of electric utility outage cost surveys," Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), Tech. Rep., 2003.

[5] S. S. Gururajapathy, H. Mokhlis, and H. A. Illias, "Fault location and detection techniques in power distribution systems with distributed generation: A review," *Renewable and sustainable energy reviews*, vol. 74, pp. 949–958, 2017, publisher: Elsevier.

[6] G. A. Ajenikoko and S. O. Sangotola, "An overview of impedance-based fault location techniques in electrical power-transmission network," *International Journal of Advanced Engineering Research and Applications (IJA-ERA)*, vol. 2, no. 3, pp. 2454–2377, 2016.

[7] P. K. Lim and D. S. Dorr, "Understanding and resolving voltage sag related problems for sensitive industrial customers," in *2000 IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No. 00CH37077)*, vol. 4. IEEE, 2000, pp. 2886–2890.

[8] S. F. Alwash, V. K. Ramachandaramurthy, and N. Mithulananthan, "Fault-location scheme for power distribution system with distributed generation," *IEEE Transactions on Power Delivery*, vol. 30, no. 3, pp. 1187–1195, 2014, publisher: IEEE.

[9] G. Ma, L. Jiang, K. Zhou, and G. Xu, "A Method of line fault location based on traveling wave theory," *International Journal of Control & Automation*, vol. 9, 2016.

[10] S. A. M. Javadian, A. M. Nasrabadi, M.-R. Haghifam, and J. Rezvantalab, "Determining fault's type and accurate location in distribution systems with DG using MLP Neural networks," in *2009 International Conference on Clean Electrical Power*. IEEE, 2009, pp. 284–289.

[11] Y. Aslan, "An alternative approach to fault location on power distribution feeders with embedded remote-end power generation using artificial neural networks," *Electrical Engineering*, vol. 94, no. 3, pp. 125–134, 2012, publisher: Springer.

[12] F. Dehghani and H. Nezami, "A new fault location technique on radial distribution systems using artificial neural network," 2013, publisher: IET.

[13] W. Li, D. Deka, M. Chertkov, and M. Wang, "Real-time faulted line localization and pmu placement in power systems through convolutional neural networks," *IEEE Transactions on Power Systems*, vol. 34, no. 6, pp. 4640–4651, 2019.

[14] A. Zainab, S. S. Refaat, D. Syed, A. Ghrayeb, and H. Abu-Rub, "Faulted line identification and localization in power system using machine learning techniques," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 2975–2981.

[15] S. Pandey, A. Srivastava, and B. Amidan, "A real time event detection, classification and localization using synchrophasor data," *IEEE Transactions on Power Systems*, 2020.

[16] S. R. Madeti and S. Singh, "Modeling of pv system based on experimental data for fault detection using knn method," *Solar Energy*, vol. 173, pp. 139–151, 2018.

[17] N. Dahal, O. Abuomar, R. King, and V. Madani, "Event stream processing for improved situational awareness in the smart grid," *Expert Systems with Applications*, vol. 42, no. 20, pp. 6853–6863, Nov. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095741741500322X

[18] U. Adhikari, T. H. Morris, and S. Pan, "Applying hoeffding adaptive trees for real-time cyber-power event and intrusion classification," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4049–4060, 2017.

[19] S. Ekici, "Support Vector Machines for classification and locating faults on transmission lines," *Applied soft computing*, vol. 12, no. 6, pp. 1650–1658, 2012, publisher: Elsevier.

[20] D.-I. Kim, A. White, and Y.-J. Shin, "Pmu-based event localization technique for wide-area power system," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 5875–5883, 2018.

[21] A. Hossam-Eldin, A. Lotfy, M. Elgamal, and M. Ebeed, "Combined traveling wave and fuzzy logic based fault location in multi-terminal hvdc systems," in *2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC)*. IEEE, 2016, pp. 1–6.

[22] Y. Mohammadnian, T. Amraee, and A. Soroudi, "Fault detection in distribution networks in presence of distributed generations using a data mining–driven wavelet transform," *IET Smart Grid*, vol. 2, no. 2, pp. 163–171, 2019.

[23] B. Palmer, W. Perkins, Y. Chen, S. Jin, D. Callahan, K. Glass, R. Diao, M. Rice, S. Elbert, and M. Vallem, "GridPACKTM: A framework for developing power grid simulations on high-performance computing platforms," *The International Journal of High Performance Computing Applications*, vol. 30, no. 2, pp. 223–240, 2016, publisher: SAGE Publications Sage UK: London, England.

[24] A. Muallem, S. Shetty, J. W. Pan, J. Zhao, and B. Biswal, "Hoeffding Tree Algorithms for Anomaly Detection in Streaming Datasets: A Survey," *Journal of Information Security*, vol. 8, no. 4, pp. 720–726, Oct. 2017. [Online]. Available: https://m.scirp.org/papers/abstract/79818

[25] Y. He, G. J. Mendis, and J. Wei, "Real-Time Detection of False Data Injection Attacks in Smart Grid: A Deep Learning-Based Intelligent Mechanism," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2505–2516, Sep. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7926429/

[26] Z. E. Mrabet, D. F. Selvaraj, and P. Ranganathan, "Adaptive hoeffding tree with transfer learning for streaming synchrophasor data sets," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 5697–5704.

[27] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://link.springer.com/10.1023/A:1010933404324

[28] B. Glocker, O. Pauly, E. Konukoglu, and A. Criminisi, "Joint classification-regression forests for spatially structured multi-object segmentation," in *European conference on computer vision*. Springer, 2012, pp. 870–881.

[29] H. Linusson, *Multi-output random forests*. University of Bor\aas/School of Business and IT, 2013.

[30] O. Pauly, "Random forests for medical applications," PhD Thesis, Technische Universität München, 2012.

[31] C. Genes, I. Esnaola, S. M. Perlaza, L. F. Ochoa, and D. Coca, "Robust

recovery of missing data in electricity distribution systems," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 4057–4067, 2018.