Training Machine Learning Models with Simulated Data for Improved Line Fault Events Classification From 3-Phase PMU Field Recordings

Hussain Otudi[1], Taif Mohamed[2], Mladen Kezunovic[2], Yi Hu[3], Zoran Obradovic[1]

[1] Computer and Information Sciences Department, Temple University Philadelphia, PA, U.S.A.

[2] Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, U.S.A.

[3] Quanta Technology Raleigh, NC, U.S.A.

Abstract

Utilizing data analytics and machine learning (ML) on phasor measurement units (PMUs) data to analyze faults automatically is the focus of this paper. Insufficient labels and natural uneven distribution of different types of line fault events found in field-recorded PMU data make supervised ML model development challenging. To address this issue, we train off-the-shelf Support Vector Machine (SVM) ML models for line fault classification using simulated PMU data obtained from a combination of 12 physical and virtual PMUs placed on a synthetic IEEE 14-bus system as well as using this simulated data unified with field recordings. A conducted sensitivity study is focused on three factors, 1) the number of PMUs used to train the ML model, 2) the voltage level at which the model is trained, and 3) the vicinity of PMUs to transmission line faults. The ML models trained with simulated and field data are evaluated on one-year field-recorded data collected from 38 PMUs sparsely located in the US Western interconnection. We demonstrate that when training ML models with only simulated data, the performance varies significantly with different number of PMUs, voltage level, and PMU placement in separate areas of the synthetic grid (F1 score of 0.78 to 0.92). We obtained an F1 score of 0.94 using the simulated dataset integrated with field recordings. The performance of a ML model developed using simulated data is also evaluated on the three-phase voltage signals extracted from 188 PMUs in the Eastern interconnection accompanied by imprecise labels, where the majority of the labels do not identify the fault type. On this extremely challenging task, we achieved 77% accuracy solely using synthetic data for ML training.

Keywords: Transmission Line, Fault Classification, Phasor Measurement Units, Machine learning

1. Introduction

In recent years, synchrophasor technology has been deployed to supplement supervisory control and data acquisition (SCADA) systems to improve electric grid monitoring, control, and protection by providing a precise and comprehensive "view" of an entire interconnection. Reporting rate for PMU data is over 100 times faster than SCADA's, which improves situational awareness, wide-area monitoring, power system planning, event analysis, and real-time operations (Kezunovic et al., 2014; Patel et al., 2010).

Manual fault classification of historical or real-time streaming PMU data is infeasible due to high data reporting rates and an increasing number of PMUs placed in the power grid. Since some field recorded data may have insufficient or inaccurate labels or some line has limitations in captured fault types, a reasonable approach might be to adopt simulated data to train ML models to automatically process and evaluate collected PMU data for more efficient and accurate fault classification.

Multiple feature extraction and ML methods were considered to automatically analyze large amounts of power system data in the past. For the early discovery of abnormal events, principal component analysis (PCA) was used to reduce the dimensionality of data gathered from PMUs (Xie et al., 2014). Complex cascading events were detected using the same method. For the extraction of discriminative features, the minimum volume closing ellipsoid (MVEE) approach was used (Rafferty et al., 2017; Dahal et al., 2014), whereas, for event classification, the agglomerative hierarchical clustering method was applied (Khan et al., 2015). Some of the approaches used for automated event detection (Biswal et al., 2016; Brahma et al., 2017) include fast variation of the discrete S-transform (Biswal et al., 2016) and signal energy transform (Yadav et al., 2019). The normalized value of the wavelet coefficient energy was used as a feature engineering technique in (Kim et al., 2017). Swinging door endings based on dynamic programming have been used to identify event start times (Cui et al., 2019). Micro-PMU data was used to examine the applicability of event detection in distribution networks (Shahsavari et al., 2019), which compared the performance of support vector machines, decision trees, and K-Nearest Neighborhood for event detection. Inputs for the Convolutional Neural Network classification model were generated using a wavelet transform-based feature engineering method (Wang et al., 2019).

This material is based upon work supported by the Department of Energy under Award Number DE-OE0000913.

The true performance of any trained ML model can be evaluated with the field recorded data. However, the field-recorded PMU data may not be a good candidate for supervised training of ML models for line fault event classification due to insufficient labels and/or uneven distribution of different types of line faults (Otudi et al., 2022).

In this paper, we demonstrate that off-the-shelf SVM ML models can be successfully trained with simulated PMU data from a synthetic grid exhibiting exact labels and an even distribution of all types of line faults to classify line faults in field recorded data. The conducted study can be carried out using a variety of ML techniques as simulated data is shown to help when certain types of line faults are insufficiently present in the observed field recordings data. However, the performance is highly dependent on the selection of the PMUs with related simulation waveforms used for training.

Our contribution is in demonstrating the sensitivity of choosing simulated data captured at various PMU locations and voltage levels to train the ML models for improved classification accuracy. Using PMU data from simulations allows us to remove the biases that come from insufficient and inaccurate labels found in the field recorded data, as reported in (Otudi et al., 2022). Our method outperforms the previous one by capturing data at the voltage level where PMUs are clustered closer to the simulated fault location.

Our findings are presented in more detail as follows: Section 2 discusses the problem background and how this paper's contribution expands previous work. Section 3 demonstrates how both field-recorded and simulated PMU data are utilized to extract features. Section 4 explains the methodology used in developing and training the ML model. The findings are presented in Section 5, followed by conclusions.

2. ML Model Development Challenges

A. *The rationale for the use of ML for line fault classification*

The practice in the power grid has long been to use the SCADA system, which cannot capture fault waveforms due to the low scan rate. To overcome this shortcoming, several intelligent electronic devices (IEDs), such as digital fault recorders (DFRs) and digital protection relays (DPRs) located in substations are triggered to capture fault waveforms and breaker operations at the time and in close vicinity of the fault. The limitations of using such high-precision triggered data by the system operator are: a) slowly evolving system-wide events that may eventually result in relay tripping may not cause the recording instruments to trigger, and b) the field-recorded fault data is not transferred to the operators in real-time. In the case of Independent System Operators (ISOs), who typically do not have access to the DFR or DPR recordings but use only SCADA RTU scans, characterizing fault conditions in real-time is almost impossible.

PMUs provide streaming data in real-time and can capture waveform distortion that characterizes the occurrence of a fault. This leads to an opportunity for differentiating and detecting not only the fault transients but also the slowly evolving events that may contribute to the relay tripping or may be caused by it. Recognizing such events automatically and associating them with the occurrence of a fault or relay operation can be achieved with the help of data analytics and machine learning on PMU recordings of voltage, current, and frequency. The higher reporting data rates and sampling time-synchronization of PMUs have made online real-time analysis of fault events feasible. However, the large amount of data from field PMUs requires advanced data analytics to automate fault detection and classification.

One of the main challenges when working with the field-recorded PMU data is that the event logs often have inaccurate or incomplete labels describing faults. The natural uneven distribution of different types of faults in field-recorded data also creates a bias in the ML model's training. The goal behind developing ML models is to use an unbiased sample set of accurately labeled events to train models to recognize and characterize such events properly. Since supervised ML approaches for line fault classification have serious limitations when labels are unavailable or imprecise, and event types are biased, it becomes critical to have an alternative unbiased training dataset with precise and complete labels, such as a simulated PMU dataset with known fault events. We hypothesize that it might be feasible to rely on a simulated PMU dataset only to train ML models for classifying line faults in the field rather than doing it on field recordings that are biased and have insufficient or incomplete labels. The simulated data can be used to facilitate unbiased training by generating an even distribution of different types of events.

B. *Challenges addressed in prior work on line fault classification using ML*

Classifying line faults using ML models trained on integrated simulated and field-recorded data is reported in (Otudi et al., 2022). In this previous study, the IEEE 14-bus system is used to simulate fault cases rarely seen in the field, and the simulated PMU data is used to complement the field-recorded dataset for training ML models. It was demonstrated that this approach successfully enhances the accuracy of several ML models used for fault classification. In (Otudi et al., 2022), field-recorded data is labeled in two steps: a) Automatic labels are assigned based on six features extracted from three-phase voltage data recorded by PMUs, and b) These features are used by supervised machine learning models to improve the fault characterization accuracy based on the provided labels. The six features are in a binary form (+1/-1), which for ten fault types resulted in 2^6=64 possible classification outcomes. Any fault that does not meet the criteria of the first nine classes is assigned to the last class, namely a three-phase (3P) line fault. Because some PMUs may be located at a distance from the fault, some voltage signals experience an indistinguishable change, which causes classification biases toward 3P and 3P-G faults, all lumped under the 3P category.

Beyond the scarcity of the three-phase fault occurrences, the original raw labels provided with the field-recorded data had several other issues. The fault events corresponding to raw labels had to be visually inspected by a domain expert to improve the accuracy of the start and end times of an event. In some instances, the fault was labeled with the wrong type. Moreover, some labels had to be split to reflect several fault events of different types happening within a short time window. The improved labels did not solve all the issues found in the training data, such as the bias in the distribution of fault types, and it would be impractical to perform visual inspection of recorded PMU data for an extended time period. To make the training data set less biased, the field recordings with wrong labels are supplemented with data from simulated recordings with accurate labels. The use of simulations from a synthetic grid balances the number of line faults of each type and improves the training of various machine learning models. When using the IEEE 14-bus synthetic grid in the previous study, no sensitivity studies were done on PMU placement, PMU association with the faulted line, or voltage levels. Just an arbitrary combination of possible fault scenarios is used to get a more reliable training set (Otudi et al., 2022).

C. Additional challenges that need to be addressed with new ML methods for line fault classification

1. Eliminating Bias. What made the fault classification feasible in (Otudi et al., 2022) is the availability of labels that specify the simulated fault type, even if the labels provided along with the data are inaccurate or incomplete. The presence of simulated data with accurate labels helped mitigate the bias mentioned in (Otudi et al., 2022). However, we also faced the situation when inspecting year-long PMU recordings from the US Eastern interconnect where the number of events labeled as faults is much lower, and the quality of a few labels that are provided is not very good. For example, the line events are rarely specified in terms of phase or general fault type, and only a few fault events are labeled. As a result, the process followed in (Otudi et al., 2022) would be infeasible if applied to the recorded data from the Eastern interconnection.

2. Selecting optimal number, location, and voltage-level placement of PMUs in the simulated network. None of these factors were investigated in our prior research (Otudi et al., 2022). As these factors could impact an ML model both during training and particularly when tested on unseen field-collected data, this has prompted us to do a sensitivity analysis for these three conditions. Accordingly, we conducted a sensitivity study to address the following:

<u>Step 1</u>: Sensitivity to the optimal number of PMUs used for training.

As simulated fault data from twelve PMUs is available to use for training purposes in this study, an interesting research question was whether measurements from all PMUs are needed for training. How different the results would be if we only use one PMU, and which PMU's data will yield the best results? To answer this question, the same ML models were trained separately using data captured by each of the twelve PMUs, and their overall performance is compared in the results section.

<u>Step 2</u>: Impact of number of PMUs at a given voltage level and their proximity to faults

Field-recorded data at the time of the fault revealed that PMUs at different voltage levels record different signal patterns. The fault signature captured by PMUs at a particular voltage level may be weak, while some PMUs are entirely "blind" to a fault. This sparked the investigation into whether the voltage level at which the PMU records in the synthetic grid affects training of the ML model. By separating the synthetic grid into a high-voltage and low-voltage area, the separation of PMUs by numbers was inevitable due to the different number of lines, each monitored by a PMU. By design, the system has more PMUs placed closer together on the low-voltage side, which also affected the results.

3. Dealing with an extreme use case of imprecise or incomplete event labels. The main goal of this study is to assess if an algorithm that is resilient to inaccurate or incomplete event labels from field recorded data can be developed to optimize the ML algorithm. Final testing of the developed models was done on the US Eastern interconnection since its field-recorded data was accompanied by imprecise and incomplete labels. The existing labels rarely specify the fault type or duration accurately. The only reliable way to test this dataset is to use the models developed using simulated data.

3. Data Description

Both simulated and field-recorded datasets are used in this study. Field-recorded data used for this research covers two years period (2016–2017) and is obtained from two sources. The first data source comes from the

38 PMUs in the Western Interconnection, and the second data source comes from the 188 PMUs in the Eastern Interconnection of the US. The simulated dataset is used to train ML models, whereas the field-recorded dataset is used to test these models. Further properties of the two datasets utilized in this investigation and their impacts are described next.

A. Properties of the Field Recorded Data

For confidentiality reasons, the field data was stripped of geographic information and of any physical location or technological characteristics of the PMUs, and no information was provided about the electric grid to which they are connected. For this experiment, the PMUs provide data on voltage, current, and phase angle, for each phase, as well as positive sequence voltage and current, frequency, and rate of change of frequency. After visual inspection of field-recorded current magnitude measurements, it was recognized that the current magnitude change was not as prominent as the voltage magnitude change for most occurrences in the dataset due to a substantial distance of PMUs from the fault location. Hence, the three-phase voltage magnitudes are used to preprocess data from the event log.

By examining the Western interconnect data, we observed that event logs are inaccurate. For example, the start and end times of the events were recorded at the time resolution of minutes, while the fault events typically have a range of 100 – 300 milliseconds. Fault durations thus are often much shorter than one minute, and so our model takes fault signals as input and classify line faults within a 2-second window. Furthermore, improvement of the labeling of events through visual inspection by domain experts to create more accurate labels becomes highly impractical when the amount of PMU data is very large. The Eastern interconnect dataset with 188 PMUs and poorer data quality presented more difficult challenges than the Western interconnection dataset of <40 PMUs and better data quality. For the Eastern Interconnection dataset, there are only 40 out of 1059-line events that have proper fault type descriptors and most of the PMUs do not have three phase measurement.

The Western Interconnection dataset was used for testing new ML models because this data has already been well-studied in (Otudi et al., 2022) with both raw and cleansed fault type labels. In addition, 40 events with fault type descriptors from the Eastern interconnection dataset are used to compare the results with the ones from the Western interconnection dataset when only simulated data is used for ML model training. The ML model accuracy from (Otudi et al., 2022) also served as a reference against which we compared the performances of our new ML algorithms.

B. Properties of the Simulated Data

To overcome the aforementioned issues with recorded data, in this paper, we develop and train a line fault classification model using simulated PMU data alone. New models are developed using the same IEEE

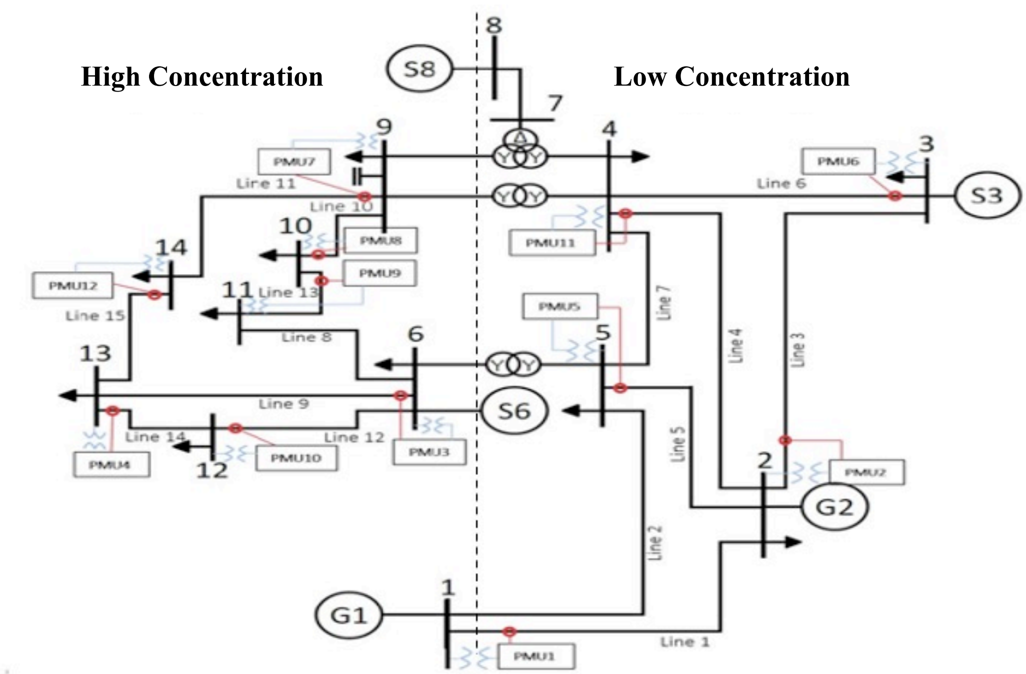

Figure 1. Simulated IEEE 14-bus power system with PMU placement

14-bus system used in (Otudi et al., 2022). Since line fault event signatures are sometimes partially distinguishable in scarcely located PMU data, our study shows how to implement automatic means of identifying these occurrences from such measurements. The study in this paper eliminates the automatic labeling stage used in (Otudi et al., 2022) and explores the use of individual PMU recordings from various locations in the 14-bus system to train ML models and test them on field-recorded data. The analysis of results of such supervised learning using higher-precision simulated PMU data demonstrates that the new approach is capable of classifying line faults captured by field PMUs with high accuracy without the need for accurately labeled field-recorded data. The synthetic IEEE 14-bus system (Figure 1) implemented on a real-time digital simulator RTDS is used to generate 1,350 simulated faults in various locations with different fault resistances and types. For every line in the system, a fault was simulated at locations of 10%, 50%, and 90% of the length of the line. All single-phase-to-ground faults have automatic line reclosing simulated, while all other fault types have manual line switching by operators simulated following the fast fault clearing. Each simulated fault event is logged in a separate event log file with precise labels for fault start and end time, fault location, fault type, and fault resistance value.

A simulated PMU dataset was generated using electromagnetic transient program implemented on a real time simulator with some PMUs connected to the simulator as hardware in the loop and some represented as a virtual high precision model. The analog output of RTDS of simulated fault events were fed into the analog input of 12 PMUs (Figure 1), among which PMU1 to PMU4 are physical PMUs, and PMU5 to PMU12 are software-emulated (virtual). The output data from PMUs is transmitted in the IEEE C37.118 format to a phasor data concentrator (PDC), where it is stored and combined to generate the simulated PMU dataset for training. No noise was added in the fault events simulations, the generated simulated PMU data have shown excellent resemblance of the field recorded fault events as seen in the field recorded PMU data. This may be attributed to the analog signal filters and the phasor calculation algorithms of actual PMUs that have the ability of filtering out most of such noises in the field recorded and simulated PMU data.

The simulated IEEE 14-bus power system has two voltage levels (low voltage on the left and high voltage on the right) connected through three transformers, as shown in Figure 1. The low-voltage side of the system has more PMU as there are more lines on the low-voltage than on the high-voltage side. This means that if and when data from these two sides are used as separate training datasets, the low-voltage side contains more training data, and the PMUs are closer to the simulated faults.

C. Summary of simulated and field-recorded data

Table 1 summarizes the properties of simulated and field-recorded data used in this study as described in this section.

Table 1. Statistics of 3-phase PMU data

Properties	**Simulated data**	**Field-recorded data**
No. of PMUs	12	38
No. of faults	1,350	266

4. Methodology

Data preprocessing, feature extraction, and data labeling are the three essential steps that were taken to train and test the ML models in this study. These three steps were done for both the simulated and field-recorded data. Figure 2 illustrates this process.

A. Data preprocessing

The simulated data from each PMU is preprocessed to allow for the training of the ML models. The magnitude of the A, B, and C voltage phasors is

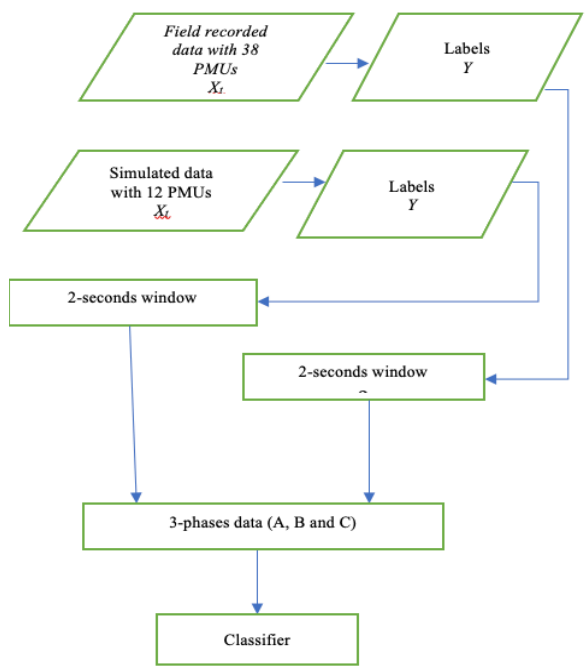

Figure 2. Unified representation of simulated and field-recorded data

retrieved from the simulated dataset for each PMU. Then, the time series data is divided into two-second intervals and windowed for discretization and quantification. we preprocess the data independently as three cases. In the first two cases, the simulated data was preprocessed from a synthetic grid after split into a subsystem (lines 1–7) that has fewer PMUs and a subsystem (lines 8–15) that has more PMUs. In the third case, simulated data was preprocessed for all 12 PMUs. Field-recorded datasets are used from 38 out of 43 PMUs placed in the Western interconnection, with the remaining five PMUs being removed due to the unavailability of the three-phase voltage measurement data and excess bad data issues. The data quality issues include missing or duplicated data, many outliers, and wrong time tags (PARTF: PMU Data Quality 2017). Data from these remaining PMUs were also divided into two-second windows.

In a previous study the impact of time window selection on the fault detection accuracy is analyzed and it is found that shorter 2-sec windows result in the increased accuracy as compared to longer windows of up to 3-minute duration (Dokic et al., 2022). This can be explained by the smaller fluctuation of normal operation within a shorter time window and exclusion of multiple events having a start time within a single window. Windows shorter than 2-sec are not considered since this results in large uncertainty in the distance between the event log start time and actual event start time due to imprecise event logs in the field recorded data.

Because the developed models can only classify what is already declared as a line fault, only the data windows that were originally labeled in the event log as a "line" event were used. As mentioned earlier, for the Eastern interconnection dataset, there are only 40 out of 1059-line events that have proper fault type descriptors.

B. Feature Extraction

Data from a synthetic grid as well as field recordings include phasor measurements related to line faults that are labeled in the event log. The impact of PMU concentration and the voltage level on fault detection is considered when using an ML model trained on simulated data. The proposed feature extraction from PMU measurements is described as follows.

Step 1: Three-phase voltage values in the range va_m(t_{min},t_{max}), vb_m(t_{min},t_{max}), vc_m(t_{min},t_{max}) in 2-second intervals are monitored to identify sudden voltage change in each phase due to a line fault.

Step 2: Next, the voltage range $\Delta\emptyset$ is computed for each of the PMUs placed in the simulated and actual network, respectively. Then, the lowest magnitude (MAG) of voltage value is subtracted from the highest voltage measurement in a 2-second window, as shown in

Table 2. Distribution of transmission line fault types in field-recorded data

Line fault type	Fault description	Faults frequency
1	Fault (P-G{A})	22
2	Fault(P-G{B})	34
3	Fault (P-G{C})	36
4	Fault (PP {AB})/PP-G{AB})	13
5	Fault (PP {BC}/PP-G{BC})	8
6	Fault (PP {CA}//PP-G{CA})	1
7	Fault (3P{ABC}/3P-G{ABC})	7

Equation 2. The sum of these ranges is determined for all PMUs in a network.

$$\Delta(V\emptyset) = \max(V(\emptyset)\ MAG) - \min\ (V(\emptyset)\ MAG) \quad (1)$$

$$SUM(V\emptyset) = \sum_{i=1}^{\#PMUs} \frac{\Delta(V(\emptyset))}{Number\ of\ PMUs} \quad (2)$$

Step 3: The difference between each pair of phase scores, $\emptyset(A, B\ and\ C)$ is then measured, and the signs of these three numbers are used as the first three features as shown in Equation 3 for phases A and B

$$\emptyset(AB) = SUM(VA) - SUM(VB) \quad (3)$$

Step 4: In addition, the ratio between the voltage range is computed. As shown in Equation 3 for phases A and B, the higher value is always divided by the smaller, and these values are employed as extra characteristics for training a classifier.

$$AB\_diff = sign\ (A, B) \quad (4)$$
$$X = ratio\ (A, B) \quad (5)$$
$$Y = ratio\ (B, C) \quad (6)$$
$$Z = ratio\ (C, A) \quad (7)$$

Step 5: Lastly, the absolute value of the Phase difference on each phase pair is computed. The above steps are applied to all combinations of phases (AB, *BC), (BC, CA), AND (CA, AB).*

C. Data Labeling

We formulate the problem as a multiclass classification task addressed by a support vector machine (SVM) (Scikit Learn, n.d.). An SVM is a supervised machine learning algorithm that transforms data to a high-dimensional feature space to maximize a gap among categories.

Generally, there are 11 types of line faults that can occur in the field. There are three possible line-to-ground faults (A-G, B-C, and C-G), three types of line-to-line faults (AB, BC, and CA), three types of double-line-to-ground faults (AB-G, BC-G, and CA-G), and three-phase and three-phase to ground faults (ABC and ABC-G). It is very difficult to differentiate between the line-to-line faults and double-line-to-ground faults in the field-recorded data, however. The same is true for three-phase and three-phase to-ground faults. Due to the labeling issues associated with the field recorded data, visual inspection of originally labeled fault events in field recorded data based on expert knowledge was applied to generate cleaned labels for those events that can be positively confirmed with precise start and stop time and remove those that couldn't. Most of the labeling issues are the results of using imprecise methods (e.g. using imprecise time at a 1-minute resolution) and/or manual processes without the use of the field recorded PMU data. Sparsity of PMUs may have also played a role as PMUs far away from a fault location or in different voltage levels may not be able to "see" the fault event clearly.

As a result, the problem is reduced to a seven-class problem, as shown in Table 2 for the Western interconnection data. This requires the development of binary models for seven different types of line faults. Each binary classifier model assigns a class label to a specific fault type, with the class receiving the most votes being chosen.

To optimize SVM parameters, 75% of the data selected completely at random is used to train the models, while a disjoint 25% is utilized to validate the accuracy of the trained model. The trained model is evaluated on the year 2017 field recordings. The overfit was analyzed by evaluating and validating the performance of SVM using five cross-validations (5-CV) before testing the model on an unknown set.

5. Results and Discussion

Apache Parquet files are utilized to store field-recorded data, while Apache Spark (Apache spark, n.d.) was used to retrieve the data. Figure 2 illustrates the unified representation of the simulated grid and field recordings of line fault classification before applying the ML technique.

We developed several ML models, each of which helps us to discuss how to approach various challenges:

Use Case 1: We trained an ML model on simulated data only and checked whether such model would be sufficient to classify line faults without relying on field-recorded data. We compared the performance of the ML model trained on simulated data to the model developed in the previous study trained on simulated and field data to demonstrate that the new model does not have the bias that existed in the prior method.

Use Case 2: We performed a sensitivity study of the location of PMUs with respect to the simulated faults in the synthetic system and illustrated how the PMU location/concentration affects the model performance.

Use Case 3: We evaluated the use of the proposed approach on the PMU data set obtained from PMUs in the Eastern interconnection of the US, where labels are imprecise and indicate neither the fault type nor a line where the fault occurred.

A. ML model trained using simulated data is sufficient

Three ML models are trained using only simulated data observed at higher voltage and lower concentration of PMUs (model 1), lower voltage and higher concentration of PMUs (model 2), and both voltage levels (model 3). The results of applying these models to field-recorded data from the year 2017 shown in Table 3 provide clear evidence that using simulated data alone (models 1, 2, and 3) outperforms a model trained on field-recordings only (model 4). In particular, model 4 achieved an F1 score of 0.74, while the F1 score for models 1 to 3 was 0.88 to 0.94, demonstrating that simulation is an excellent choice when learning to classify line faults from sparse streaming data. F1 score is reported in conducted experiments since it provides a measure balance between precision and recall. It is computed as (2*precision*recall)/(precision+recall), where the precision is the number of true positive results divided by the number of all positive and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive.

Table 3: F1 score on 2017 field-recorded test data for models 1-4

Model/ Training Dataset	PMU Concentration	Voltage-level and Number of PMUs	Weighted F1-score
1/simulated	low	5 PMUs at high-voltage	0.88
2/simulated	high	7 PMUs at low-voltage	0.93*
3/simulated	both	All 12 PMUs	0.88
4/field-recorded (2016)	N/A	all 38 field-recorded PMUs	0.74

In models 5, 6, and 7 presented in Table 4, where simulated data is used to enhance field recordings, simulated data from the area with higher PMU concentration was also more helpful than simulated data from the area with lower PMU concentration or using all simulated data (F1 score of 0.94 is obtained by model 6 vs. 0.88 obtained by model 6 and 0.92 by model 7).

Table 4: F1 score on the 2017 field-recorded test data for models 5-7

Model/ Training Dataset	PMU Concentration	Voltage-level and Number of PMUs	Weighted F1-score
5/simulated & field-recorded (2016)	low	5 PMUs at high-voltage and 38 PMUs of field-recorded	0.88
6/simulated & field-recorded (2016)	high	7 PMUs at low-voltage and 38 PMUs of field-recorded	0.94*
7/simulated & field-recorded (2016)	both	50 PMUs (12 in simulated and 38 in field-recorded)	0.92

Our experiments provide evidence that the proposed method eliminates the biases that are found in (Otudi et al., 2022), as shown in Figure 3.

In the previous method, if none of the labels aligned with predefined combinations, the fault is classified as 3 phases or 3 phases to ground. This is a serious issue when the classifier is applied to small datasets. In contrast, the proposed method eliminates that bias by not relying on the labeling technique and manually centering the line fault in a training window

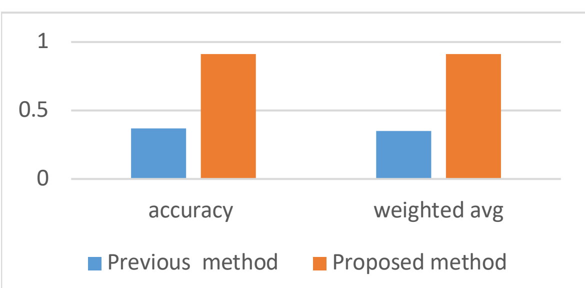

Figure 3. Comparing the previous method (i.e., rule based labeling technique) vs the proposed method

B. PMUs concentration/voltage level impacts the model performance

Placement of PMUs in the synthetic grid affect the accuracy of the ML model classification. Figure 4 depicts and compares performance when using a lower concentration of PMUs (orange bar) with the case when a higher concentration is used (blue bar). Two criteria are examined: accuracy and F1 weight average. In this experiment, two subsystems are used to train the ML models. Model 1 is trained on a subsystem with lower PMU concentration, and Model 2 is trained on a subsystem with higher PMU concentration, as shown in Figure 1. Both models are tested using field-recorded data from 2017. Then, the suggested approach is further evaluated by domain expert visual inspection. We found that the misclassified are the same 16 events when tested on the field-recorded events from 2017. These 16 events constitute 48.5 %, 69.5 %, and 61.5% of the errors made by models 1, 2, and 3 trained on simulated data of subsystems where the PMU concentration was low, high, and from both subsystems, respectively.

Visual inspection revealed that these 16 events are difficult to classify even by a domain expert. The reason is that time windows either contained more than one event, or the events themselves are difficult to distinguish because of their very close resemblance. For example, Figure 5 shows one such event. The model classified this event as a CA-G fault, but it is logged as a C-G fault. It is difficult to classify this fault because a significant voltage dip is seen in phases A and C. However, the first dip is clearly in phase C, while the second is in phase A. This is an example of an event where the fault may have evolved into a different type during the auto-reclosing sequence.

Next, we extend our experiment by training classifiers using an equal number of line faults on lower and higher PMU concentration subsystems of the

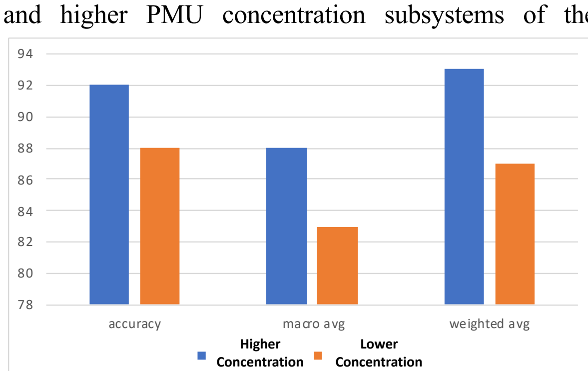

Figure 4. Fault classifier comparison for higher vs lower PMU concentration using three metrics

simulated network. Table 4 illustrates how the difference in PMU concentration affects the classifier accuracy. We observe that even if the number of PMUs is intentionally made equal, training an ML model on the subsystem with a higher number of lines resulted in better accuracy, as reported in Table 5.

In addition, training data is split into an equal number of PMUs selected from the high (right) and low (left) voltage subsystems. Resulting models 8 and 9 are trained on PMU's measurement from the synthetic grid and evaluated on the field recorded data from the year 2017.

Model 8 outperformed model 9 with a low standard deviation (see Table 5). Moreover, models 10 and 11 are trained on the mixture of simulated and field- recorded data resulting in high performance and low value of std.

Table 5. Weighted F1 score for models 8-11 when tested on field recordings from year 2017

Model/ Training Dataset	Event location	No. of lines	STD	AVG
8/simulated	low	4	0.001	0.92
9/simulated	high	4	0.002	0.78
10/simulated & field-recorded (2016)	low	4	0.010	0.93
11/simulated & field-recorded (2016)	high	4	0.036	0.87

C. Model evaluation on Eastern interconnection

As previously mentioned, the data received from PMUs installed in the US Eastern interconnection exhibits imprecise and incomplete labels for line faults. The majority of the labels do not identify the fault type. We wanted to see how the model trained using only simulated data and precise corresponding labels for such data will perform when tested on the mentioned field data. The model used in this evaluation is model 2. The results are evaluated by visual inspection of the three-phase voltage signals extracted from 188 PMUs in the Eastern interconnection. Because missing PMU data presented a real challenge to view some of the faults, only the fault instances that are identifiable are included in this evaluation. The issue of missing PMU data was addressed during data preprocessing by imputing the missing data with the last available value. The resulting performance of the classifier model resulted in an accuracy of 77 %, which is not high but acceptable for the case with almost no usable labels.

6. Conclusion

Three tasks considered in this work are:

- Confirm if the usage of only simulated data that have high precision labels captured by PMUs in a synthetic system will yield high-performance models,
- Check the sensitivity of the models when PMU location and voltage level are varied, and whether the concentration of PMUs in the simulated subsystem affects the accuracy of the ML model trained using simulated data.
- Evaluate performance of the model trained using simulated data and tested on field-recorded data from the US Eastern interconnection.

The results suggest the following:

- If event log labels are inaccurate or insufficient, a useful line fault classification model could be trained with simulated PMU data.
- Using simulated data alone to train an ML line faults classification model is preferable to relying only on improperly labeled PMU field recordings.
- Compared to the previous method (Otudi et al., 2022), the proposed method eliminates the bias that was introduced by the automatic labeling method.
- The use of higher-concentration PMUs in the synthetic grid is beneficial, which in our case meant using PMUs from the left-hand side of Figure 1.
- A model trained using simulated data from a network with higher PMU concentration and field-recorded data resulted in higher prediction accuracy.

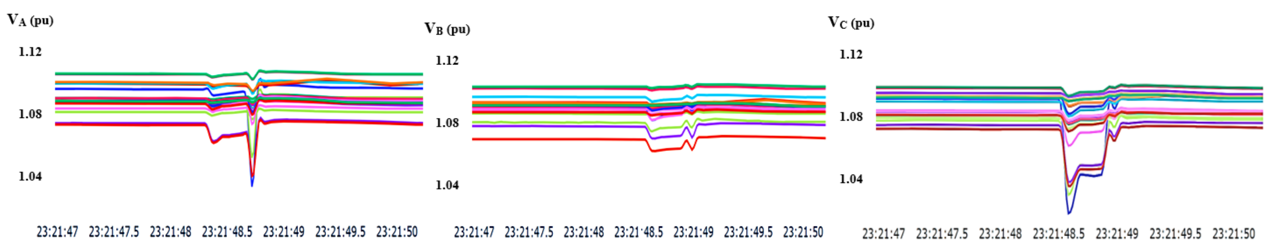

Figure 5. Misclassified event on phases A, B, and C (from left to right)

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed or represented that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Acknowledgment

Mr. H. Otudi has been funded by the College of Computer Science and Information Technology at Jazan University in Saudi Arabia.

7. References

Apache spark™ - unified engine for large-scale data analytics. Apache Spark™ - Unified Engine for large-scale data analytics. (n.d.). Retrieved from https://spark.apache.org/.

Biswal, M., Brahma, S. M., & Cao, H. (2016). Supervisory protection and automated event diagnosis using PMU Data. IEEE Transactions on Power Delivery, 31(4), 1855–1863. https://doi.org/10.1109/tpwrd.2016.2520958

Biswal, M., Yifan Hao, Chen, P., Brahma, S., Cao, H., & De Leon, P. (2016). Signal features for classification of power system disturbances using PMU Data. 2016 Power Systems Computation Conference (PSCC). https://doi.org/10.1109/pscc.2016.7540867

Brahma, S., Kavasseri, R., Cao, H., Chaudhuri, N. R., Alexopoulos, T., & Cui, Y. (2017). Real-time identification of dynamic events in power systems using PMU data, and potential applications—models, promises, and challenges. IEEE Transactions on Power Delivery, 32(1), 294–301. https://doi.org/10.1109/tpwrd.2016.2590961

Cui, M., Wang, J., Tan, J., Florita, A. R., & Zhang, Y. (2019). A novel event detection method using PMU data with high precision. IEEE Transactions on Power Systems, 34(1), 454–466. https://doi.org/10.1109/tpwrs.2018.2859323

Dahal, O. P., Brahma, S. M., & Cao, H. (2014). Comprehensive clustering of disturbance events recorded by phasor measurement units. IEEE Transactions on Power Delivery, 29(3), 1390–1397. https://doi.org/10.1109/tpwrd.2013.2285097

Dokic, T., Baembitov, R., Hai, A. A., Cheng, Z., Hu, Y., Kezunovic, M., & Obradovic, Z. (2022). Machine learning using a simple feature for detecting multiple types of events from PMU Data. 2022 International Conference on Smart Grid Synchronized Measurements and Analytics (SGSMA). https://doi.org/10.1109/sgsma51733.2022.9806000

Kezunovic, M., Meliopoulos, S., Venkatasubramanian, V., & Vittal, V. (2014). Application of time-synchronized measurements in power system transmission networks. Springer International Publishing.

Khan, M., Ashton, P. M., Li, M., Taylor, G. A., Pisica, I., & Liu, J. (2015). Parallel detrended fluctuation analysis for fast event detection on massive PMU Data. IEEE Transactions on Smart Grid, 6(1), 360–368. https://doi.org/10.1109/tsg.2014.2340446

Kim, D.-I., Chun, T. Y., Yoon, S.-H., Lee, G., & Shin, Y.-J. (2017). Wavelet-based event detection method using PMU Data. IEEE Transactions on Smart Grid, 8(3), 1154–1162. https://doi.org/10.1109/tsg.2015.2478421

Otudi, H., Dokic, T., Mohamed, T., Kezunovic, M., Hu, Y., & Obradovic, Z. (2022). Line faults classification using machine learning on three phase voltages extracted from large dataset of PMU measurements. Proceedings of the Annual Hawaii International Conference on System Sciences. https://doi.org/10.24251/hicss.2022.425

PARTF: PMU Data Quality: A framework for the attributes of PMU data quality and a methodology for examining data quality impacts to synchrophasor applications. North American SynchroPhasor Initiative. (2017). Retrieved from https://www.naspi.org/node/352

Patel, M., Aivalitiotis, S., & Ellen, E. (2010). Real-time application of synchrophasors for improving reliability. Reference Documents | North American SynchroPhasor Initiative. Retrieved from https://www.naspi.org/reference-documents

Rafferty, M., Liu, X. A., Laverty, D., & McLoone, S. (2017). Real-time multiple event detection and classification using moving window PCA. 2017 IEEE Power & Energy Society General Meeting. https://doi.org/10.1109/pesgm.2017.8274000

Scikit Learn: Support Vector Machines. scikit. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/classes.html?highlight=svm#module-sklearn.svm

Shahsavari, A., Farajollahi, M., Stewart, E. M., Cortez, E., & Mohsenian-Rad, H. (2019). Situational awareness in distribution grid using Micro-PMU Data: A machine learning approach. IEEE Transactions on Smart Grid, 10(6), 6167–6177. https://doi.org/10.1109/tsg.2019.2898676

Wang, S., Li, L., & Dehghanian, P. (2019). Power Grid online surveillance through PMU-embedded convolutional neural networks. 2019 IEEE Industry Applications Society Annual Meeting. https://doi.org/10.1109/ias.2019.8912368

Xie, L., Chen, Y., & Kumar, P. R. (2014). Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis. IEEE Transactions on Power Systems, 29(6), 2784–2794. https://doi.org/10.1109/tpwrs.2014.2316476

Yadav, R., Pradhan, A. K., & Kamwa, I. (2019). Real-time multiple event detection and classification in power system using Signal Energy Transformations. IEEE Transactions on Industrial Informatics, 15(3), 1521–1531. https://doi.org/10.1109/tii.2018.2855428