

Use of Machine Learning on PMU Data for Transmission System Fault Analysis

Mladen KEZUNOVIC*
Texas A&M University
U.S.A.
m-kezunovic@tamu.edu

Zoran OBRADOVIC
Temple University
U.S.A.
zobrad@gmail.com

Yi HU
Quanta Technology
YHu@Quanta-Technology.com

SUMMARY

Synchrophasor technology has been used for monitoring, control, and protection of bulk power system for over 10 years. Deployment of phasor measurement units (PMUs) in the USA power system has surpassed 3000 units installed in the transmission substations as stand-alone intelligent electronic devices (IEDs) or as a software add-on to other devices such as digital protective relays (DPRs) or digital fault recorders (DFRs). By now, thousands of terabytes of PMU data may have been captured and stored by various transmission system operators (TSOs) and independent system operators (ISOs). This creates an opportunity to deploy advanced machine learning (ML) techniques to detect and classify faults recorded by PMUs automatically to be used by the system operators for rapid, critical decision-making when manual analysis of the past or unfolding events is not feasible.

In this paper we offer a brief background on how the automated fault analysis may be done using DPR and/or DFR data, and compare some of the legacy approaches to the new ML approaches in the context of the system-wide PMU recordings. We then offer insights from developing practical ML solutions that have been applied on field recordings captured by close to 450 PMUs from all three US interconnections (Western, Eastern and ERCOT) over two years (2016-2017). We identify and illustrate ML challenges we addressed: inaccurate data, data with scarce and temporally imprecise fault labels, data recorded by PMUs sparsely located at substations resulting in the fault records taken afar from the ends of the faulted lines, data containing only positive sequence values, and data taken at different voltage levels. We then illustrate the ML model results for fault analysis under different application scenarios.

The novelty of this study is not only in the design, implementation, and performance analysis of the ML algorithms, but also in the use of advanced fault modelling and simulation approaches to improve the training results when developing supervised ML models for fault detection and classification. Extensive simulations of faults were conducted on a 14-bus power system to create a training dataset with over 1400 accurately labelled faults. This dataset was applied to enhance the accuracy of fault detection and classification of machine learning-based models trained with small number of labelled faults in large datasets recorded in the grid interconnections ranging from 5,000 to 70,000 buses.

KEYWORDS

Synchrophasors, Machine Learning, Fault Analysis, Modelling and Simulation.

1. BACKGROUND

As it is well known, power systems experience faults due to weather, deterioration of equipment, animal intrusion and many other related causes [1]. Such faults typically result in a rapid change in the current and quite often in a voltage collapse or swell accompanied by a possible appearance of a DC offset and frequency changes. While the initial faults last for a relatively short time of one to several alternating current cycles (a cycle is 16.67ms in a 60Hz system), the fault clearing event can last much longer if autoreclosing sequences are involved. A single protective relay operation may also change the power system topology and operating condition to a point where some other relays in the vicinity operate unintentionally causing cascading events. As a consequence, the stability limits may be reached, which may result in a system collapse. Hence, the analysis of faults requires details about the local events as well as system-wide impacts. This leads to the analysis requirements that focus on safety and reliability, range across variety of spatiotemporal scales, and aim at different end-users.

Dependability vs Security of Relay Operation

The concept of dependability and security of protection relay operation is defined to differentiate between fast fault clearing requirements in the case the fault occurs (dependability) and restrained operation in the case the event is not a fault (security) [2]. The relay operation performance under such guidelines comes from the requirement to maintain operational safety (fast fault clearing) and system reliability (avoidance of relay overtripping). In order to understand what is happening during the faults, different fault analysis spatiotemporal granularity needs to be applied ranging from insights into the local events to the characterization of the system-wide consequences.

Spatiotemporal Scales of Fault Analysis

Initial fault events may last as short as a fraction of the voltage or current cycle before the protective relays detect the fault and issues a trip signal to operate a circuit breaker or set of breakers to isolate the fault. The modern digital relays can operate as fast as one cycle and advanced circuit breakers can trip as fast as 1-2 cycles bringing the initial fault clearing sequence to a few cycles [3]. After the attempt to clear the faults is made, the automated reclosing sequences lasting several hundreds of milliseconds may be deployed to try to restore the system in the case the faults are of a temporary nature. Should the fault persist, the faulted element is removed and system operation is restored with new system conditions that may affect the voltage, frequency and power flow. All of this is noticeable during and after fault clearing sequence which may last up to several seconds. If a single fault is not causing cascades, the system should be capable of recovering under such N-1 events. However, in perhaps rare circumstances, the new system conditions may cause m failures of other equipment such as circuit breakers, or initiate mis-operation of other relays, which may result in N-m system-wide disturbance that can last for minutes and sometimes longer, and eventually may cause even a system collapse [4]. As observed, the time scale of the fault analysis may range from milliseconds to minutes or even hours, and the spatial scale may include the local relays involved in the fault clearing, or much wider area that may include other affected relays or manifest in a system-wide disturbance.

Users of Automated Fault Analysis

Typically, the users of the automated faults analysis are the protection engineers and operators, former with a goal of understanding the reasons for particular relay operation and related consequences so that future protection system performance can be improved, and later with a goal of assessing the system conditions to determine whether any manual operator action may be used to remedy or prevent the consequences of faults on system reliability and safety. Since the data collected during faults may be overwhelming, the automation of the fault analysis is highly desirable even if no immediate action is taken, and is absolutely necessary if any real-time action is warranted [5].

The next section discussed the analysis solutions that meet different aspect of the of the protective relaying performance assessment, the spatiotemporal scales, and the need for timely results.

2. LEGACY FAULT DETECTION AND CLASSIFICATION APPROACHES

Typically, utilities use three types of systems to perform fault analysis as shown in Figure 1. They are differentiated by the equipment involved in the recording of data during faults. Such solutions are distinctively different when it comes to the level of details the analysis can offer, the spatiotemporal scales it can cover, and the ability to offer a remedial action, all tied to the the inherent design constraints of each system. Next, we cover basic characterises of such systems and their performance limits and benefits. We use the example from Figure 1 to illustrate the options when the voltage is monitored.

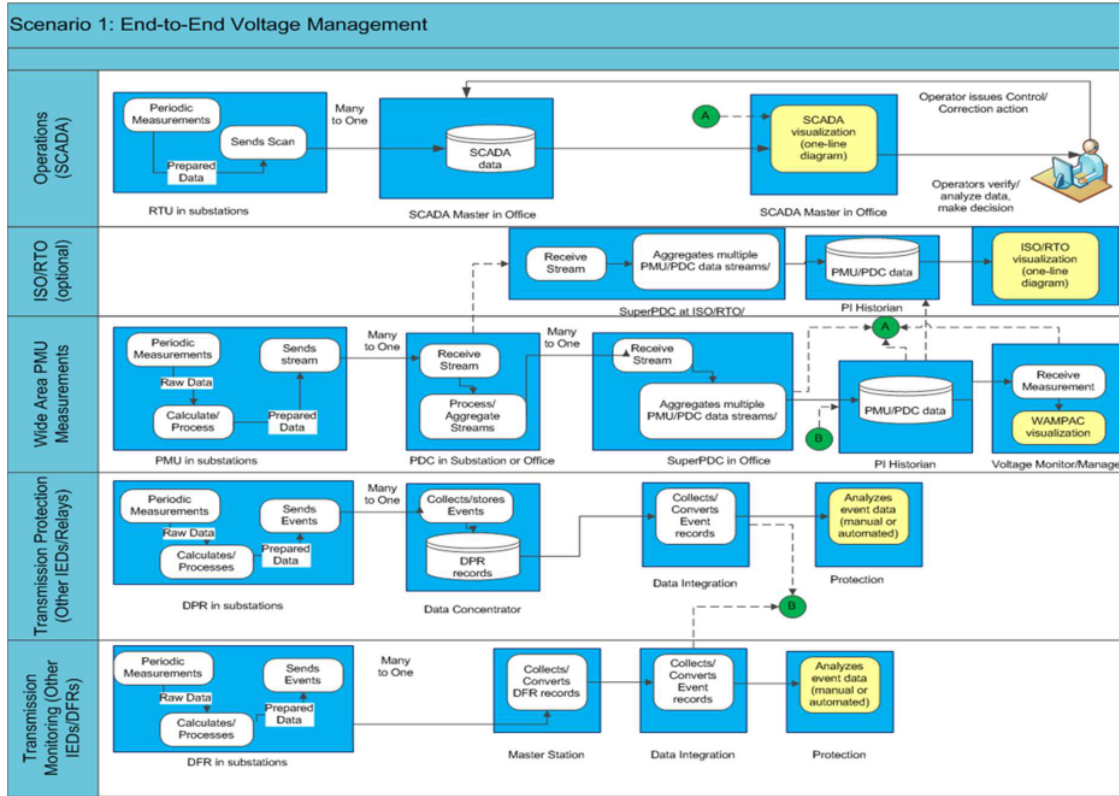


Figure 1 Various substation recording systems for voltage monitoring

Fault Analysis Using Data Captured by Digital Fault Recorders and Protective Relays

Most of the time fault analysis is performed using recording from Digital Protective Relays (DPRs) and Digital Fault Recorders (DFRs) shown at the bottom of Figure 1. Since such an analysis requires determination of phase relationship between three-phase voltages and currents, the sampling on the related recording channels has to be synchronized to the same sampling clock. This is performed at a relatively high sampling rate in the order of 1kHz for DPRs and 5kHz for DFRs. To facilitate an understanding of pre-fault events, a short time window of the pre-fault waveforms is captured. Typically, the sampled values are sent to the centralized place, most often the Protection Engineer's office for off-line analysis. In recent years, automated systems for fault data analysis have been developed and allow such non-operational data to be turned into the operational data that may be used by the operators for mitigating fault impacts within seconds to minutes of the fault occurrence [5].

Several advantages and disadvantages of such analysis systems have been well understood over time. The key advantage is that such systems are capable of tracking transients caused by the faults in all three phases rather accurately, which leads to thorough understanding of the correlation between different faulted phase of voltages and currents. The disadvantage of the DPRs is that they can only capture synchronized recording on the transmission line where they are located, and the files from

multiple relays cannot be easily merged for a system-wide analysis unless their sampling clock is controlled by an accurate timing signal received from the Global Position System (GPS) of satellites, which requires each substation to be equipped with a GPS receiver. The DFRs typically have higher sampling rates than DPRs and can represent transients even better, and their sampling clock is synchronized for all the recording channels across the entire substation. However, merging the files from multiple substations with no synchronized clocks requires elaborate time-alignment techniques or the use of a clock from the GPS receiver. Last but not least, both types of IEDs are triggered instruments, which means that they can only capture waveform samples once an event is detected, so slowly evolving events such as gradual frequency change or voltage sags or swells may not be captured. The reason for their limited recording capability is the design constraint of relatively small local memory for storing the high sampling-rate data.

Fault Analysis Using SCADA Systems

The Supervisory Control and Data Acquisition (SCADA) systems have been developed to collect data from substations using Remote Terminal Units (RTUs) as shown at the top of Figure 1. The data is scanned typically every few seconds, and the sampling of the waveforms by RTUs is not time-synchronized. Hence, the analogue waveforms cannot be used for analysis of fault types because of the low time-fidelity and because the angle between the faulted phases cannot be determined [6].

The key advantage of the SCADA systems is in their ability to collect statuses of circuit breakers, which helps in determining which element of the power system was disconnected due to a fault. However, the RTUs/SCADA may not time-stamp accurately the fast subsequent circuit breaker operations such as the ones involved in the autoreclosing sequences, which may limit the ability of the operators to analyse faults based on the SCADA RTU recordings unless an advanced SCADA with Sequence of Events (SoE) feature is deployed. Also, due to a large number of RTU measurements captured across all the substations, the SCADA systems tend to be designed to only record the channels that experience a large deviation from a pre-set threshold in each scan, so not all the measurements are reported continuously. Since SCADA collects thousands of measurement points, and brings them to the centralized location where the SCADA database resides, measurement errors due to missing or altered data are a common problem. To make such measurement useful for the Energy Management Systems (EMS) applications, a state estimation approach is commonly used for the topology and measurement error detection and mitigation [7].

Fault Analysis Using Synchrophasor Systems

To alleviate shortcoming of the SCADA systems, the synchrophasor systems (shown in the middle of Figure 1) were developed in the USA over 30 years ago, but applied widely only since 2009 when the American Recovery and Reinvestment Act (ARRA) subsidized utilities to deploy PMUs at large scale [8]. Today, synchrophasor systems are widely used to supplement SCADA system by reporting measurements at a high resolution of 30 or 60 frames per second (FPS), which are time-synchronized to a common clock such as UTC (coordinated universal time) derived from the GPS timing source through GPS receivers placed in the substations.

The obvious advantage of the synchrophasor systems is in their ability to provide high-resolution time-synchronized measurements, but at the same time there are multiple disadvantages when it comes to the fault analysis using PMU recordings. First, such systems calculate an estimate of a phasor, which is not a very detailed representation of a transient waveform needed to fully analyse faults. Second, the FPS reporting, while relatively high in comparison to SCADA, it is still not sufficient to capture fast fault transients that may die out in less than 100 milliseconds. Last, due to a relatively high cost of adding PMUs and GPS receivers and communication facilities to transfer streaming PMU data, the PMUs are still sparsely located in limited number of substations not capturing accurately faulted waveforms in the case the event occurs further away from the substations where PMUs are located. Their key advantage still remains the ability to capture evolving events continuously over a wide area, also being a critical information in understanding the fault impacts.

3. SYNCHROPHASOR-BASED APPROACH AND ASSOCIATED DATA ISSUES

In the USA, over 3000 PMUs are installed in the three Interconnections, with over 500 buses in ERCOT, over 20,000 buses in the Western interconnection, and over 75,000 buses in the Eastern Interconnection [8]. It is rather obvious that the sparsity of the PMUs with the respect to the number of substations where PMUs may be installed is significant. Besides, the experiences from the analysis of two-years' worth of recordings from 450 PMUs placed in the three interconnections reveal substantial data quality issues [9]. Last but not least, the natural occurrence of faults results in an uneven distribution of the fault recordings related to different faulty types [10]. All three issues may dramatically affect the ability to build and train ML models to detect and classify faults based on synchrophasor measurements alone. We discuss what we learned about those issues through the analysis of actual field recordings.

Data Recording Limitations for Fault Analysis Caused by PMU Locational Sparsity

The existing PMUs sparsely located across the power systems are often placed near particular power system equipment for certain purposes, such as at large generation stations for generator dynamic performance monitoring and model validation/calibration or at critical substations for situational awareness about the impact of fault clearing topology switching. As a result, sparsely placed PMUs are often far from the locations of various types of faults which makes the recorded data less useful for fault analysis because the voltage drop and current jump due to a fault are the highest at the fault location and they will diminish at the PMUs located further away. In addition, as more PMUs are currently placed at the substations with high voltage levels such as 500kV and 230kV than with lower voltage levels of 130 kV and 69kV, recorded PMU data in voltage levels different from the voltage level of the fault location may not be able to "see" fault correctly if transformers between different voltage levels are wye-delta connected. This may only be an issue for the current situation. As the number of PMUs increases, this will become less of an issue in the future.

PMUs, with a typical 30 or 60 frame per second (FPS) data reporting rate for a 60 Hz system, generate orders of magnitude higher volume of data comparing to SCADA's RTU data scan rate of once several seconds. To reduce the demand for communication network bandwidth and data storage space, often only positive sequence voltage and current phasors from PMUs are transmitted and archived despite the fact that all PMUs are capable of generating and sending the three-phase phasors, which are extremely useful for fault analysis. A non-three-phase type of fault may look less severe in positive sequence than in three-phase phasors (e.g. one phase voltage drops from nominal value down to close zero due to a single-phase-to-ground fault may only show as a 33% voltage drop in the positive sequence voltage phasor). In the dataset we examined, majority of Western interconnection PMUs provide three phase phasors, while less so in the other two interconnections.

Bad Data Impact on Fault Analysis

Two key data quality issues, i.e. missing data and unreasonable data, identified through our analysis have direct impact on fault analysis. Based on the dataset we examined, the results of the analysis are shown in Figure 2 and Figure 3. As it may be observed from Figure 2, several PMUs are missing upwards of 60% of key measurement data while the best PMUs only missing ~2% of the data. From Figure 2, it appears that the PMU scaling was off in many instances, which also affects the fault analysis due to misinterpretation of the scale of voltage and current during faults. No meaningful fault analysis can be performed where a PMU is missing large blocks of data. This issue could not be remedied by performing some types of data pre-processing since that would have to be addressed in the field by identifying the root causes of the issue and taking actions to mitigate them appropriately.

Different types of unreasonable data have been identified in the dataset. Some are outliers, such as phasor angle values have gone beyond the ± 180 degree range, phasor magnitude has negative values that are supposed to be always positive, and extremely large values beyond the reasonable values (e.g., voltage magnitude value is several orders of magnitude higher than the nominal value). Others do not

appear to be actual measurement values, such as flat 60Hz as the measured frequency values while real power system frequency never stays at 60Hz but fluctuate around the 60Hz all the time during normal system operations. Those outliers, if not removed, would not affect the fault analysis.

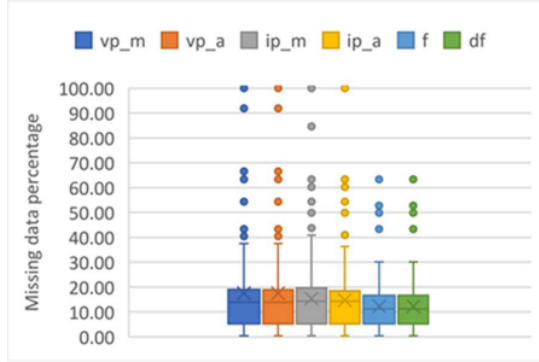


Figure 2 Missing data

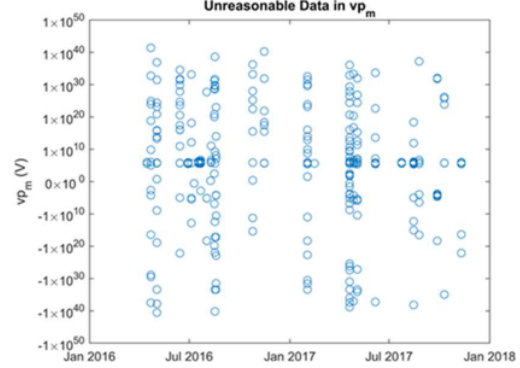


Figure 3 Unreasonable data

Modelling and Simulation to Alleviate Limitations Caused by an Uneven Recorded Fault Types

Fault events are rare occurrences in power systems, which are typically caused by equipment failures, severe weather conditions, and sometimes human errors. Among eleven types of the faults (i.e. three types of single-phase-to-ground as A-G, B-G, and C-G, three types of phase-to-phase faults as AB, BC, and CA, three types of phase-to-phase-to ground as AB-G, BC-G, and CA-G, and the three-phase faults, ABC and ABC-G), the occurrence of the faults for each type is uneven in the recorded data. A majority of recorded faults are single-phase-to-ground faults, followed by faults involving any two phases, and very few are three-phase faults, which presented a real challenge to train ML models effectively, and to test the trained models properly.

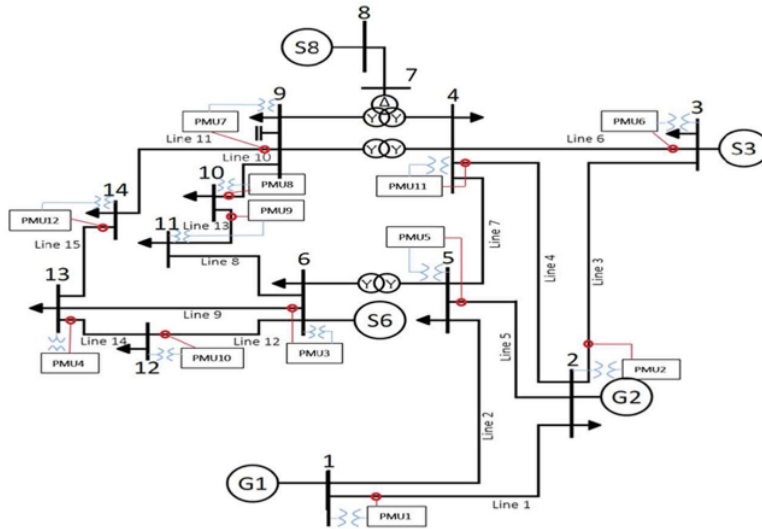


Figure 4 PMU Placement in the synthetic IEEE 14-bus power system

To alleviate the situation created by rare occurrence of the fault events and the uneven distribution of different types of faults recorded in the field, we have built a synthetic 14-bus power system model (Figure 4) on a real-time digital simulator and connected to 4 actual and 8 emulated PMUs to generate a simulated line fault event dataset with 135 fault events for each of the ten types of the faults. Each of the simulated fault events included a typical fault

inception, clearing and reclosing sequences. Following the fast fault clearing, automatic line reclosing is simulated for all single-phase-to-ground faults, and manual line closing by human operators was simulated for all multi-phase faults. The simulated dataset has a separate event log file in which detailed information for each simulated fault event (i.e., time, location, fault resistance, and fault type) is provided. The dataset has been used to train different types of ML models with good results.

4. CHALLENGES IN IMPLEMENTING MACHINE LEARNING APPROACHES

As mentioned earlier, there are many approaches to implementing automated fault detection and classification, and such approaches have been used for the analysis of power system events in the past [5]. ML methods aim to develop a model that discovers patterns in observed temporal data to automatically identify and categorize faults. The ML process consists of (1) choosing an appropriate knowledge representation; (2) selecting an objective or loss function for evaluation of possible models; and (3) using an optimization method to minimize the error on training data by adjusting the parameters of the model.

To capture temporal dynamics for fault detection and classification, ML approaches typically represent data as feature vectors over a sliding time-window, where features could be carefully engineered or extracted more automatically from raw signal. Commonly used objective functions for ML-based detection and classification include maximum likelihood for estimating the parameters of the true probability distribution based on observed training data, and cross entropy that measures how closely the true probability distribution is represented by an approximation obtained by the model with specific parameters. Typically, continuous function optimization algorithms are used to find the model parameters that result in minimal loss. They are based on using derivative information, but there are also alternatives used for non-differential objective functions.

Common ML modelling assumptions are that (a) training and test instances are drawn independently from the same unknown distribution; (b) function mapping explanatory variables (inputs/features) to response variables (outputs) is smooth almost everywhere; (c) signal to noise ratio is large; (d) ignored information has negligible effects; and (e) labels are precise and the set of labelled events is sufficiently large. When it comes to faults detection and classification, some of these assumptions are often violated resulting in various challenges. In particular, we are facing several challenges associated with labelling of faulted events, selecting appropriate data and time-windows for the analysis, and selecting the most suitable ML models. Our choices and implementation constraints discussed next are based on the specific PMU data sets that were available for the implementation of the automated ML solution.

Pre-processing Imprecise and Inaccurate Data Labels

The fault dataset labels are often scarce, inaccurate, or imprecise. Faults that occur far from PMUs might not be assigned labels while labels entered manually are subject to human error, and those obtained from SCADA are imprecise due to lower accuracy of the time-tag information.

To reduce some of these limitations, additional cleansing of faults labels could include rapid, partial, and full visual inspection. Towards this objective we analyzed ML benefits of using (a) handpicked normal operation data; (b) narrower time intervals; (c) reduced time intervals to capture single events; (d) precise centering of events in time-window intervals; and (e) visual inspection of reported events. Our results provide evidence that ML-based faults detection is more accurate when using more preprocessed data. However, data preprocessing requires domain expert's time, which in practice is often extremely limited. Therefore, we analyzed ML cost-benefits of label preprocessing aimed to reduce imprecise and inaccurate annotations. Our extensive experiments conducted on data from the U.S. Western interconnection include recording from 38 PMUs over two years. The most cost-accuracy benefits were obtained when integrating a small fraction of expert-inspected data labels with additional labels that were not manually inspected by a domain expert [11]. When learning fault detection using 12-month data, major accuracy improvements were obtained by full inspection of 2-month data combined with remaining automatically preprocessed data (F1 score 0.86 vs 0.82).

Selecting Knowledge Representation for PMU Time-Windows

Raw PMU recordings require significant transformation to make them suitable for ML training. In our study, PMU signal is down sampled to 30FPS as some PMUs raw recordings are at 60FPS while

others are at 30FPS. Then, each minute of a single PMU recording we represent as a sequence of 180 non-overlapping windows covering 10-frames. Such sequences are concatenated to a high-dimensional tensor over all PMUs. When integrating data from 38 PMUs, N-minute of signal we represent as $N \times 38 \times 180 \times 10$ tensor. We reduce data size by focusing time-window around reported faults as these are considered the most informative for ML training, while the remaining parts of the data are discarded. This representation requires additional transformations to address large data dimensionality, causing a curse-of dimensionality challenge for ML training and inference. In addition, the fault's signature characteristics are not necessarily evident on all PMUs due to their geographical and topological distance of PMUs to the event site.

We addressed both challenges by transforming high-dimensional PMU recordings data tensors to a simplified representation that maintains the essential temporal information, but automatically reduces the data size and noise [12]. This is achieved by adopting Soft Dynamic Time Warping technique to summarize all time windows observed during a single 1-minute interval at all PMUs. The summary time series is a low-dimensional representation of the corresponding group of time series, and it preserves the distortions and sudden jumps, which are essential properties for ML-based fault detection. We found significant benefits of using information observed by several PMUs (e.g. average AUC 0.81 based on 3 PMUs vs 0.77 when relaying on a single PMU) but the variation of performance measured as the standard deviation of AUC decreases as the number of PMUs increases.

5. ML SOLUTION FOR AUTOMATED FAULT DETECTION AND CLASSIFICATION

In this section we discuss our ML automated solutions for fault detection and classification based on the PMU recordings pre-processing and knowledge representation we discussed earlier. We focus on the most important methodological steps that we have pursued as we defined, implemented, and tested the solution.

ML Models for Fault Detection

Traditional 2-class ML approaches can be leveraged for fault detection by learning from positive sequence voltage and current phasors recordings captured by multiple PMUs. Using PMU data pre-processing discussed in the previous section we developed fault detection methods that employ Decision Tree (DT), Multiple Logistic Regression (MLR), Multilayer Neural Networks (MNN), and Multi-Class Support Vector Machine (mcSVM)[11]. We also developed Convolutional Neural Network (CNN) based models where automated de-noising techniques and knowledge representation discussed in the previous section are applied to the measurements and then parallel and concatenation-based convolutional neural networks that include multiple layers performing convolution operations are utilized for fault detection [11].

Our experiments using 2-years PMU recordings at the U.S. Western interconnection provide evidence that CNN-based ML fault detection was more accurate than traditional models. We found that for fault detection voltage was more relevant than current and frequency. However, multi-channel CNN that learn from joint voltage, current and frequency information observed at multiple PMUs were more accurate than alternative ML models [11]. In particular, in our experiments such a model achieved AUPRC of 0.9 and F1 score 0.86, outperforming four traditional machine learning models (DT, MLR, MNN and mcSVM) that achieved AUPRC of 0.62 to 0.82 and F1 score 0.72 to 0.77.

ML for Faults Classification

Once a transmission line fault is detected, another ML method could be applied to classify the fault type to one of 11 categories defined in one of previous sections, which can be generalized as P-G, PP, PP-G, 3P, and 3P-G. A challenge for ML-based fault classification from PMU data is that field recorded data might not be able to distinguish PP from PP-G faults or 3P from 3P-G faults when all PMUs are distant from the fault location. This challenge we addressed by merging AB with AB-G, BC with BC-G, CA with CA-G and ABC with ABC-G fault types. Consequently, the fault classification

problem is reduced from 11 to 7 categories that are easier to separate automatically from three-phase PMU data.

Another challenge for ML-based fault classification is that certain types of line faults are less frequent, causing severe training data imbalance. This is a major problem for many ML algorithms that assume a similar number of cases for each class. ML models trained on such data could have poor sensitivity for infrequent classes. A straightforward solution of subsampling large classes of faults to obtain a balanced training dataset is inappropriate since even more common fault types are quite infrequent as compared to normal operation, and therefore a very long observation time would be required to collect a sufficiently large training data for accurate ML.

To address class imbalance and class size problems, we questioned if these data limitations could be reduced by learning from integrated three-phase field-recorded measurements and line fault simulations [10]. The aim of simulations conducted in our study was to combat class imbalance and to enlarge the size of training dataset needed for more accurate ML. The result of such training data enhancement obtained in our study is shown at Figure 5. Blue bars show the number of recorded line faults for each of 7 categories in year 2016 at the U.S. Western interconnection while red bars show the number of faults simulated at a 14-bus power system model described at Figure 4. Integrated training data is balanced and is five times larger than field-recordings alone. Enhancement is particularly large for three line and three lines to ground faults where a single fault is extended to 86 cases.

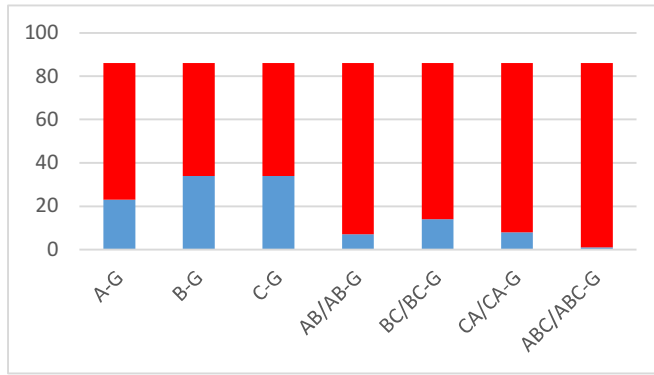


Figure 5 121 field recorded faults (blue) are extended by 481 simulated faults (red) resulting in 86 cases for each of 7 classes

Field-Recorded Data			
Models	Weighted Precision	Weighted Recall	F1-score
SVM	83.25%	91.03%	86.87%
RF	83.31%	91.03%	86.89%
XGBoost	84.13%	91.03%	87.17%
Micro_average_of_precision_recall			94.90%
Integrated Data			
SVM	98.69%	98.62%	98.58%*
RF	98.08%	97.93%	97.83%
XGBoost	98.25%	97.93%	97.88%
Micro_average_of_precision_recall			99.20%

Table 1. Fault classification performance across multiple evaluation metrics

ML using integrated field-recorded data that include information from 38 PMUs and simulated data from 12 PMUs (4 actual and 8 emulated) required for unifying extracted features to the same dimensionality. In our experiments a unified knowledge representation is achieved by measuring the range of voltage of each PMU individually over a data window and computing an aggregated sum normalized by the number of PMUs in each system. Such aggregated range information is computed between each two phases for a data window and finally the ratio between the differences in the voltage range is determined.

In our experiments shown at Table 1 fault classification obtained by SVM, Random Forest (RF) and Extreme Gradient Boosting (XGBoost) models on unseen year 2017 field-recorded PMU data was significantly improved by learning from integrated field-recorded and simulated data [10]. In particular, precision and recall of SVM were 83% and 91% respectively when learning from field-recorded data and both scores improved to 99% when learning from integrated data. F1-score also improved from 87% to 99%, providing additional evidence that it is beneficial to train ML models using integrated field-recorded PMU data and simulations vs. relying on field-recorded data alone when certain types of events are insufficiently represented in field-recorded data in the training period.

6. CONCLUSION

The performed analysis and results lead to the following conclusions:

- ML is an effective approach to detecting and classifying faults if using three-phase recordings of voltage and current, but using only positive sequence voltage makes it challenging.
- Data cleansing and curation to get rid of the bad data and label ambiguities is time well spent since such an additional effort can lead to much better ML algorithm performance.
- If the recordings contain an uneven representation of various fault types, the simulations using a synthetic power system model can generate new fault waveforms to balance the training set.
- Wide-area synchrophasor recordings offers an advantage of capturing a slow change of system voltage, current and frequency that cannot be captured by recording of local IEDs.
- To make a business case for the use of automated ML algorithms for fault analysis, one has to perform an evaluation how much to invest in PMUs to obtain the full system-wide coverage.

ACKNOWLEDGEMENT AND DISCLAIMER

This material is based upon work supported by the Department of Energy under Award Number DE-OE0000913. 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 represents 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 authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

BIBLIOGRAPHY

- [1] "Blackout Tracker," EATON, United States Annual Report, 2018.
- [2] E. A. Udren et al., "Proposed Statistical Performance Measures for Microprocessor-based Transmission-line Protective Relays. Part I. Explanation of the Statistics," IEEE Transactions on Power Delivery, vol. 12, no. 1, pp. 134-143, Jan 1997.
- [3] J. S. Thorp, et. al. "Limits to Impedance Relaying," IEEE Transactions on Power Apparatus and Systems, vol. PAS-98, no. 1, pp. 246-260, Jan. 1979, doi: 10.1109/TPAS.1979.319525.
- [4] US-Canada Power System Outage Task Force, "Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations," DOE OE, April 2004.
- [5] T. Popovic, M. Kezunovic, B. Krstajic, "Implementation Requirements for Automated Fault Data Analytics in Power Systems," International Transactions on Electrical Energy Systems, Published online in Wiley Online Library DOI: 10.1002/etep.1872, January 2014.
- [6] S. A. Boyer, "SCADA: Supervisory Control and Data Acquisition," International Society of Automation, June, 2009.
- [7] A. Abur, A. Gomez Exposito, "Power System State Estimation: Theory and Implementation," Marcel Dekker, ISBN: 0-8247-5570-7, 2004.
- [8] Smart Grid Investment Grant Projects, DOE, (OnLine: <https://www.energy.gov/oe/arra-smart-grid-investment-grant-sgig-projects>)
- [9] Z. Cheng, et. al "Using Synchrophasor Status Word as Data Quality Indicator: What to Expect in the Field?" IEEE Intl. Conf. SGSM2022, Split, Croatia, May, 2022.
- [10] H. Otudi, et. al. "Line Faults Classification Using Machine Learning on Three phases Voltages Extracted from Large Dataset of PMU Measurements," Proc. 54th IEEE HICSS, January 2022.
- [11] M. Pavlovski, et. al. "Hierarchical Convolutional Neural Networks for Event Classification on PMU Measurements," IEEE Transactions on Instrumentation & Measurement. Vol. 70, 2021, DOI: 10.1109/TIM.2021.3115583
- [12] M. Alqudah, et. al. "Convolution-based Fault Detection Utilizing Timeseries Synchrophasor Data from Phasor Measurement Units." IEEE Transactions on Power Systems. Early Access. DOI: 10.1109/TPWRS.2021.3135336