DISTURBANCE AND OUTAGE ROOT CAUSE ANALYSIS IN DISTRIBUTION SYSTEM USING PHYSICS-AWARE DATA-DRIVEN APPROACHES

By

AMIR GHOLAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
School of Electrical Engineering and Computer Science

DECEMBER 2022

© Copyright by AMIR GHOLAMI, 2022
All Rights Reserved

© Copyright by AMIR GHOLAMI, 2022
All Rights Reserved

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of AMIR GHOLAMI find it satisfactory and recommend that it be accepted.

Anurag K. Srivastava, Ph.D., Chair

Noel Schulz (Co-chair), Ph.D.

Anamika Dubey, Ph.D.

ACKNOWLEDGMENTS

The journey of my dissertation would not have been possible without the continuous support and guidance by my advisor, committee members, family and friends.

First, I would like to express my gratitude to my advisor, Dr. Anurag Srivastava, for giving me an opportunity to work under his supervision for my PhD. I am forever grateful for your constant motivation, inspiration, and support to me to excel in my field. I am obliged to him for critical inspection at each step of my work and providing constructive feedback on my work.

I would like to sincerely thank my co-chair Dr. Noel Schulz, and my committee member Dr. Anamika Dubey for their time, and encouraging and valuable comments. I am indebted for their constructive criticism, informative suggestions and critical questions which helped me in achieving my desired goal in this dissertation.

I would like to thank my labmates and friends: Dr. Sanjeev, Dr. Sajan, Dr. Venkatesh, Linli, Srayashi, Chuan for their line of thoughts and discussions and useful feedback, and for their friendly support.

I am very grateful to my family: my parents, my sister and my brother-in-law for their support in my entire life. I am always grateful to my mother, for her continuous

support in my interest for higher studies. Thank you for your sacrifices and believing in me to pursue my dreams. I am grateful to Almighty for blessing me and my family.

DISTURBANCE AND OUTAGE ROOT CAUSE ANALYSIS IN DISTRIBUTION SYSTEM USING PHYSICS-AWARE DATA-DRIVEN APPROACHES

Abstract

by Amir Gholami, Ph.D.
Washington State University
December 2022

Chair: Anurag K. Srivastava

Securing cyber-power distribution systems (DS) against malicious events is critical with integration of distributed energy resources (DERs), supporting digital automation and increasing vulnerabilities. Situational awareness utilizing power data (e.g., data from distribution phasor measurement units (D-PMUs)) and cyber data (e.g., network packets data) is the key objective of this dissertation for enhanced real-time monitoring and decision support.

To enable the reliable and resilient DS operation, this work aims to develop an automated tool consisting of multiple modules including, a) data aggregation to synchronize the resolution and time stamp of multiple metering sources throughout

the DS, b) analyze data anomalies, c) state estimation, d) cyber-physical event detection and classification, and e) Outage Root Cause Analysis (ORCA). These modules utilize range of theoretical approach including the Ensemble Extended Kalman Filter (EEKF), data fusion, data analytics, unsupervised machine learning, Hierarchical Clustering and FP-Growth Rule Mining by exploiting the real-time measurements as well as the system physics. Developed approaches have been validated using the IEEE and OPAL-RT test systems as well as measurements from actual hardware sensors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEDICATION

Dedicated to my Mother

CHAPTER 1. INTRODUCTION

1.1 Background and Motivation

The U.S. power grid is going through a massive transition from fossil-fuel based to renewable energy based power systems. As of Jan 12, 2018, more than 94% of newly integrated electrical energy capacity has been renewable energy based sources which has cut down carbon emissions by 1% annually [1]. In 2020, renewable energy sources constituted of 21% of all electricity generated in the U.S. Renewable sources and surpassed both nuclear and coal based energy generation for the first time in record [2]. Fig. 1.1 shows the prevalence of renewable in the U.S. power grid. The inclusion of renewable is going to increase rapidly in the next few decades as proposed by the 80×30 U.S. Clean Energy Standard [3]. The US government aims to generate 80% of the electricity from renewable and nuclear energy sources. To reach the proposed goal, generation units between 60 GW and 80 GW of new clean energy sources need to be integrated in the power grid. In this context, the power grid both in transmission and distribution levels needs to be able to plan and withstand such aggressive clean energy integration. A major part of renewable energy integration has happened in

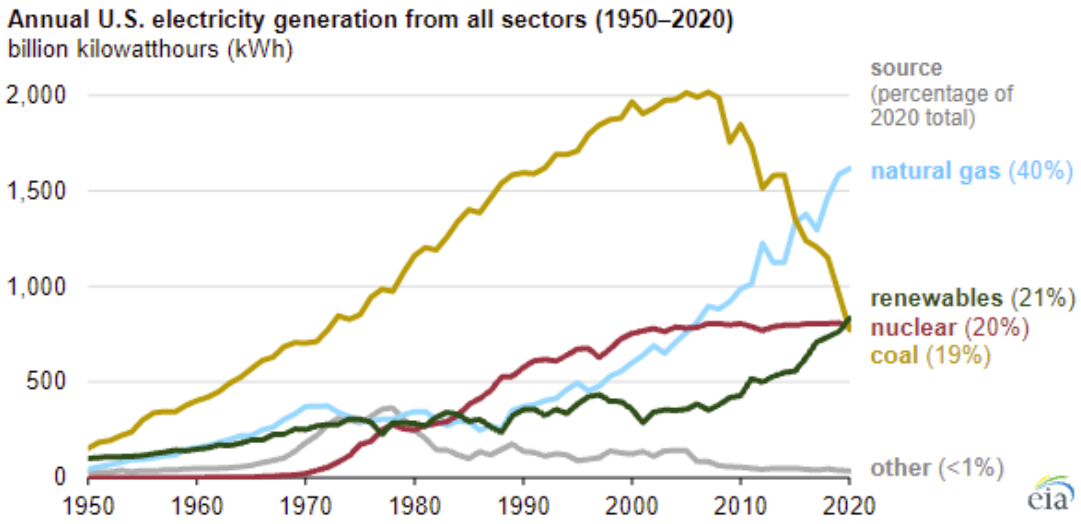


Figure 1.1: Annual electricity generation since 1950 in the US

distribution power systems in form of rooftop photovoltaic (PV) panels and battery energy storage systems (BESS). It is thus essential to look into distribution power system operation and reliability aspects due to massive penetration of renewable energy sources.

Electrical power system operation problems are broadly classified into (a) planning problems and (b) control of operation problems. Planning problems deal with problems that the grid may face in the next 10 and 20 years. Some examples of planning problems are unit commitment (UC), expansion planning, hosting capacity prediction etc. Control of operation problems deal with a smaller time scale problems possibly the next 24 hours, next hour or next minute. Optimal power flow problem is a control of operation problem which talks about how the power grid can be run op-

timally while satisfying operational constraints. Traditionally, power system utilities have used centralized schemes or local or a combination of the two schemes for optimal power system operation in distribution power grids. Centralized control schemes involve accumulation of measurement data in a centralized unit and thereby running an optimization problem in the Advanced Distribution Management System (ADMS) to generate optimal setpoints or commands for controllers in the distribution feeder. Hence, centralized decision making is based on centralized computation and network-wide measurements & communication among the physical controllers and the ADMS. Local schemes, on the other hand, involve only local measurement, communication and computation. Both centralized and local schemes are not suitable for optimal operation of present-day distribution grids. The reasons are described below.

Distributed energy resources (DER) are renewable energy driven generation units. Hence, power injections from DERs are intermittent and variable in nature. Intermittency in power injections of DERs cause voltage transients and other power quality issues in distribution power feeders. With fast-acting power electronic converter interfaced DERs acting as control agents, the number of controllers in the centralized optimization problem is significantly high. This makes centralized optimization computationally-heavy which may not provide any feasible solution to the distribution system operator (DSO) to tackle fast transients. Moreover, centralized decision making involves network-wide measurements and communication which make

them prone to single-point failures. Local solutions, on the other hand, though are simple to implement do not give optimal solutions and hence do not provide network-wide optimally. To avoid these disadvantages of centralized and local schemes, the control architecture of distribution grids is transitioning to a more distributed architecture. Distributed architecture implies distributed sensing, distributed computation and distributed decision making while having communication among neighbor controllers.

1.2 Problem Statement

With the increasing penetration of distributed energy resources (DERs), the diversification of load demand, and the intensification of random behaviors in distribution network (such as the interaction between electric vehicles (EVs), distribution grid as well as the characteristics of intermittent power generation, etc.), the operational complexity of the distribution network is also growing. These difficulties pose new challenges for distribution system operators (DSOs) to comprehensively and accurately perceive the state of the system in their area of responsibility. **Situational awareness in the power system, consisting of such components as automation, root cause analysis, and decision making, defines an exceptional capability for operations and self-improvements, in which deficiencies may**

cause distribution systems to fail to operate effectively under certain conditions. The lack of SA has been identified as a cause of several recent significant electrical disturbances worldwide. Therefore, improving SA is critical prior to establishing the system abilities to withstand and reduce the magnitude and duration of disruptive events. It is a prerequisite and strategy for enhancing the grid resilience.

Improving SA can be achieved through sensor data acquiring, big data mining, measuring data processing, and information exploring. Current state-of-the-art grid data sources and monitoring approaches are typically stovepiped into repositories of operational data (distribution phasor measurement unit (D-PMU) data, supervisory control, and data acquisition (SCADA) data), customer or billing data (advanced metering infrastructure (AMI) meter data), and third-party data (such as local news, forecasting weather data). This segmentation limits the planning and analysis of solar photovoltaic (PV) power generation in real-time operations, thereby reducing the data values. In addition, data measured by sensors are usually not corrected for gaps and errors or archived, which further limits their utility for SA. The primary goal of my study is how to coordinate the segmented data, thrifty pre-processing, thus reducing the data burden of power facilities and achieving a comprehensive state of SA of the grid with high penetration of DER. The other goal of this study is the application of power system information. When the power grid has enough information infrastructure to support SA, how to made better use of it to improve

the power grid resilience.

In order to enable the power grids to achieve a higher level of SA and improve their resilience in the future, my Ph.D. work primarily focused on the research of power system data analysis, the application of machine learning and artificial intelligence in smart grids, PMU signal processing, and the control of distribution network. The first half of my research was mainly concerned with the improving power system SA that enhances the system monitoring and perception ability. In the next step, I will focus on how to effectively use the comprehensive information and SA to control the power grid under certain environments to improve the resilience of the power grid.

1.3 Objectives and Tasks

The main objectives and tasks of the this Ph.D. research work is determined based on the detailed identified gaps in the existing literature. An integrated tool consisting of data analytic and artificial intelligence advancements along with power system expertise knowledge base has been targeted in this dissertation. Effort is made towards establishment of a Distribution System-based architecture to process the streaming and non-streaming data with the goal of enhancement of situational awareness and easing the required post-event restoration processes.

The following bullet points summarize the specific objectives of this Ph.D. research

work:

- Real-time perception of Distribution System operation by investigation of sensor measurements as well as historical data
  - Development of a mechanism for multi-sensor measurement accumulation
  - Sensor resolution enhancement by Data-In-Data-Out measurement fusion from multiple devices
  - Data synchronization and denoising
  - Detection of bad data followed by mitigation
  - Development of State Estimation technique tailored for Distribution Systems with high noise levels and low observability with presence of bad data
- Cyber-physical Event Investigation
  - Detection of Events with broader classification
  - Detailed labeling of event type with nuanced classification
  - Accurate identification of the event location with respect to the neighboring sensors
- Outage Management System improvement by integration of automated outage detection and root cause analysis tools

1.4 Dissertation Organization

The dissertation consists of five chapters. Chapter 2 presents the enhancement of SA by development of various modules to accomplish the measurement preprocessing and denoising in Distribution Systems (DS). In this chapter detailed novelties in this Ph.D. research work are discussed under the title of Data Fusion and online and offline data preprocessing techniques. The developments are validated using IEEE standard system.

Chapter 3 is focused on disturbance analysis in DS by detection and classification of anomalies. In this portion of the dissertation, the definition of anomaly is to target bad data, noise, missing values, and any actual power system events. State Estimation with proposed OF-WLS technique is elaborated in this chapter as well as presentation of performance validation results.

Chapter 4 is aiming towards specification of multiple cyber-physical events and the response of the DS towards each of the events. Data-driven and model-based techniques are integrated in this chapter and variety of events have been introduced as events of interest in this portion of the research.

The subject of outage management and root cause analysis is addressed in Chapter 5, targeting load outage events in case of scenarios with large-scale events with cascading nature. A hierarchical clustering approach is introduced to identify a broader

level of the outage root cause and in the next step, using the proposed data-mining technique, nuanced outage root cause scenario is generated. Upon an occurrence of an outage event, the developments in this chapter, would be beneficial for utilities to be able to properly and rapidly follow the power restoration process,

Lastly, in chapter 6, a conclusion of the entire sections of this PhD work is discussed and potential lines of research for future continuation of the under-the-study area are proposed.

1.5 Publications

Journals

- **Gholami Amir**, A. Vosughi and Anurag K. Srivastava, "Denoising and Detection of Bad Data in Distribution Phasor Measurements Using lFiltering, Clustering, and Koopman Mode Analysis," *IEEE Transactions on Industry Applications*, vol. 58, no. 2, pp. 1602-1610,.

- S. Som, R. Dutta, **A. Gholami**, A. Srivastava and S. Chakrabarti "DPMU-based multiple event detection in a microgrid considering measurement anomalies," *Applied Energy*, vol. 308, 2022, ISSN 0306-2619.

- **Gholami Amir**, Anurag K. Srivastava and S. Pandey "Data-driven failure di-

agnosis in transmission protection system with multiple events and data anomalies," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 4, pp. 767-778, July 2019, doi: 10.1007/s40565-019-0541-6.

- **A. Gholami**, and A. Srivastava "ORCA: Outage Root Cause Analysis in DER-rich Power Distribution System Using Data Fusion, Hierarchical Clustering and FP-Growth Rule Mining," *IEEE Transactions on Smart Grid*, Submitted

- A. Arman, **A. Gholami**, M. Namaki, A. Srivastava and Y. Wu "Spatio-Temporal Deep Graph Network for Event Detection, Classification and Localization in Cyber-Physical Electric Distribution System," *IEEE Transactions on Idustrial Informatics*, Submitted

Conferences

- **A. Gholami**, and A. Srivastava "Comparative Analysis of ML Techniques for Data-Driven Anomaly Detection, Classification and Localization in Distribution System ," *NAPS 2020*,

- **A. Gholami**, M. Musavi, A. Srivastava and A. Mehrizi-Sani "Cyber-Physical Vulnerability and Security Analysis of Power Grid with HVDC Line ," *NAPS 2019*,

- **A. Gholami**, Chuan Qin, S. Pannala, Anurag K. Srivastava, F. Rahmatian, R. Sharma and S. Pandey. "Anomaly Detection in D-PMU Data using Statistical, and Clustering Techniques," *IEEE MSCPES* 2022.

- A. Vosughi, **Gholami Amir** and Anurag K. Srivastava, "Denoising and Bad Data Detection in Distribution Phasor Measurements Using lFiltering, Clustering, and Koopman Mode Analysis," *IEEE 2021 IAS Annual Meeting.*

REFERENCES

[1] "MS Windows NT kernel description." https://electrek.co/2018/01/12/94-percent-new-electricity-capacity-usa-from-renewables/. Accessed: 2020-10-20.

[2] "MS Windows NT kernel description." https://www.eia.gov/todayinenergy/detail.php?id=48896. Accessed: 2020-10-20.

[3] "MS Windows NT kernel description." https://spectrum.ieee.org/clean-energy-standard. Accessed: 2020-10-20.

CHAPTER 2. ENHANCE SITUATIONAL AWARENESS BY DATA PREPROCESSING, DENOISING AND FUSION

2.1 Introduction

Distribution-level phasor measurement units (D-PMU) data are prone to different types of anomalies given complex data flow and processing infrastructure in an active power distribution system with enhanced digital automation. It is essential to pre-process the data before being used by critical applications for situational awareness and control. In this work, two approaches for detection of data anomalies are introduced for offline (larger data processing window) and online (shorter data processing window) applications. A smoothing wavelet denoising method is used to remove high-frequency noises. An ensemble approach built upon the margin-based maximum likelihood estimator (MB-MLE) method is developed to detect anomalies in denoised data by integration of the results from different base detectors including Hampel filter, Quartile detector and DBSCAN. The processed data with offline analysis is used to fit a model to the underlying dynamics of synchrophasor data using Koopman Mode Analysis, which is subsequently employed for online denoising and

bad data detection (BDD) using Kalman Filter (KF). The parameters of the KF are adjusted adaptively based on similarity to the training data set for model fitting purposes. Developed techniques have been validated for the modified IEEE test system with multiple D-PMUs, modeled and simulated in real-time for different case scenarios using the OPAL-RT Hardware-In-the-Loop (HIL) Simulator. Distribution grid is transforming from a passive to an active distribution network with the integration of Distributed Energy Resources (DERs) and the emergence of Microgrid concepts. Monitoring and control becomes critical with increased number of components in active distribution network using new sensors data such as distribution phasor measurement units (D-PMUs). Synchrophasor data provides a wealth of information that enables the grid operator to capture fast transient dynamic events. However, D-PMU data needs to be pre-processed in order to denoise and detect and mitigate the bad data before being fed to any other critical applications.

2.1.1 Related Works

Power system is a multi-layer distributed cyber-physical system. Synchrophasor data Anomalies may originate in different cyber-physical system layers. Fig. 2.1 illustrates data flow of synchrophasor data in different layers of smart distibution grid. Spatial-temporal correlated anomaly originated from physical layer and related to

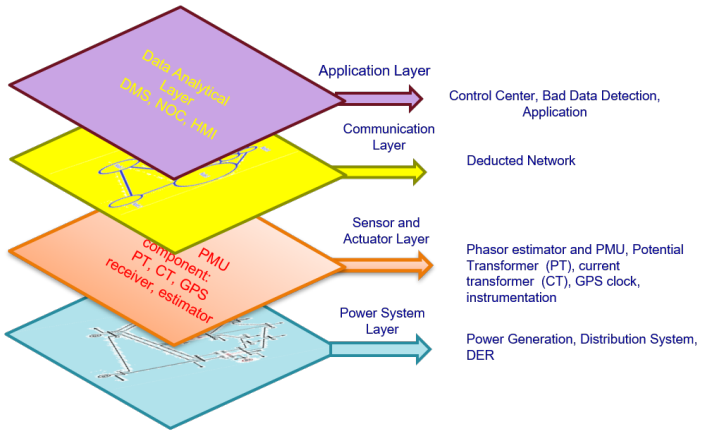


Figure 2.1: Synchrophasor data in different layers of power distribution system

physical events [1]. In sensor layer, field D-PMU data is captured with contamination of Bad Data and noise from base sensors (voltage/current/GPS) [2]. Also, this data needs to be sent over the communication network and it is prone to communication delay, failure, and cyber attacks. Anomalies in sensor and communication layer, are not usually temporal and spatial correlated. Available data in application layer is contaminated with anomalies of previous layers and distinguishing between the source of the anomaly is challenging. [3] Fig. 2.2 determines different types of events that can originate anomaly in D-PMU data.

Mis-classification between Bad Data (BD) and spatial-temporal correlated events can result in detection of actual events as bad data or noise. False-negatives and false-

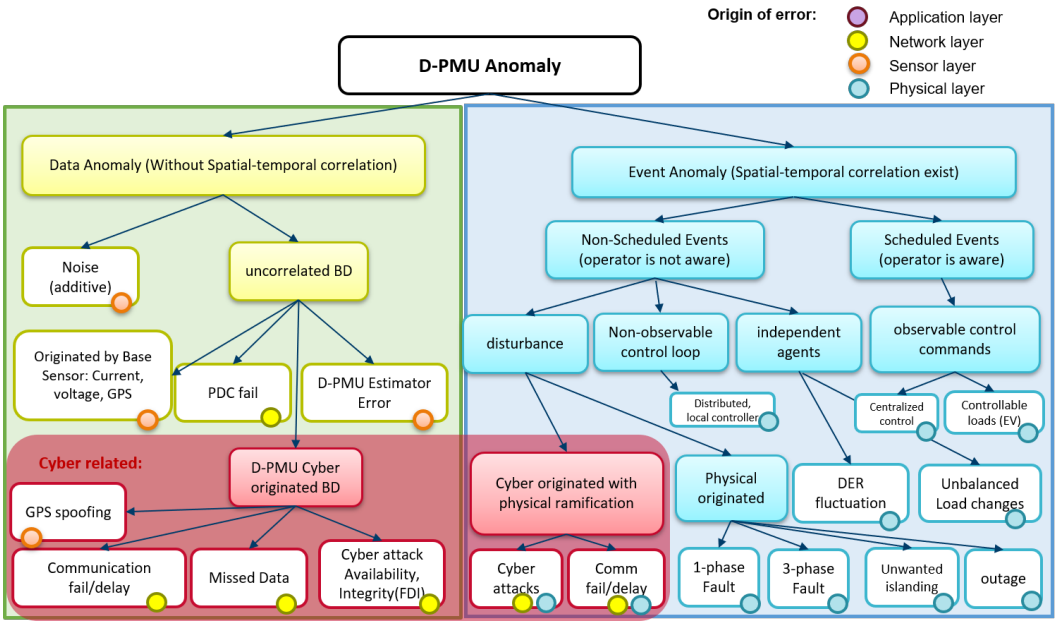


Figure 2.2: Possible Causes for D-PMU Data Anomalies

positives in detecting non-spatial-temporal BD, result in BD in measurements. BD may misleads the event classification/localization and can impact control decision. [4] Data pre-processing and measurement denoising is a significant step which needs to be accomplished prior to investigation of any physical or cyber-induced events throughout the system [5, 6]. A proper BDD technique is of enormous concern in such Distribution System (DS) applications as State Estimation (SE) [7], fault classification [8], fault location identification [9], and situational awareness [10]. [11] presents a comprehensive study on different ML-based approaches for distribution level disturbance analysis with regard to efficacy examination and performance eval-

uation of each technique. Assuming a normal distribution for load profiles, authors of [12] propose a novel BDD approach based on the Weighted Least Squate (WLS) method. Literature is mainly focusing on either BDD or denoising separately accompanied with detection of sensor faults in applications with renewable energies [13], while in the current research work, the focus is on development of two tools for online and offline BDD and denoising in an integrated manner. Preliminary work for phasor measurement denoising in distribution system is report in [14], and further expanded and comprehensively elaborated in this work with bad data detection.

Wavelet transform has been used widely for transient analysis in the power system [15, 16] and is employed for signal smoothing in this work. The authors of the current manuscript combine the results from a variety of base anomaly detectors such as Hampel filter [17], Quartile technique, and DBSCAN using MB-MLE to accomplish the BDD and mitigation. For online signal processing, the recursive Bayesian Kalman filter (KF) has been used [18]. For the prediction step of KF, rather than using a physical model, we utilize Koopman Mode Analysis (KMA) [19] for fitting the model for prediction of behaviour of the system with a data-driven approach. Multiple KMA applications in power system has been developed including but not limited to power system stability and dynamic analysis [20], control and estimation [21], fault line isolation [22], and cyber attack detection [23]. There has been other studies of voltage stability-based analysis incorporating HVDC lines, including cyber and physical co-

simulation in [24] which are also of high potential fits for KMA applications.

The focus of this work is to detect anomalies in D-PMU data and denoising of the measurements based on the applications in offline and online manners. As far as the offline applications are concerned, the computational time cost is relaxed and bad data detection and denoising can be done with smoothing methods. MB-MLE is used to integrate the base detector scores to detect the bad data. Different base anomaly detectors are employed with various detection nature and tailored mathematics to detect the anomalies and this diversity enhances the likelihood of isolating outliers. The wavelet smoothing technique is employed to denoising the signal. For online mode, the KF has been used to denoise the signal and detect the bad data based on the residual analysis. Given the complexity of DS non-linear model extraction with high resolution synchrophasor data, data-driven Koopman mode analysis has been used to fit a model that describes the underlying dynamic interactions between the D-PMU measurements through the prediction step of Kalman Filter (KF). Since the training data set is not necessarily the most comprehensive inventory of events with BD, the online evaluation of the performance of the fitted model is essential and the parameters of the KF must be adjusted adaptively based on the credence of the fitted model to rely more on measurement as necessary. This is done by observing the deviation of the mean of the estimated state from measurements. Fig. 2.3 illustrates the connection of proposed approaches in an integrated coherent manner.

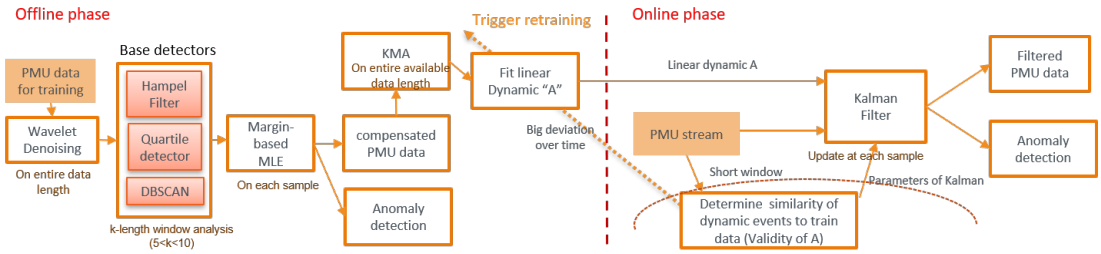


Figure 2.3: The scheme of proposed online and offline approaches for D-PMU non spatial temporal anomaly detection

To summarize, the main contributions of this research can be enumerated as follows:

1. Developed ensemble-based integration of multiple bad data detection (BDD) techniques for processing larger time window of data

2. Development of an adaptive KF for online denoising and BDD with shorter time window of data, tailored for closed-loop controls and time-sensitive applications

3. Developed algorithim for linearization of the DS model during dynamic transients using Koopman Mode Analysis (KMA), resulting in reduced BDD processing time with high accuracy

2.2 Offline Tool

2.2.1 Wavelet Filter

Wavelet transform is an extension of Fourier transform where frequency and temporal analysis are performed simultaneously to capture signal frequency evolution over time. For that, wavelet transformation localizes features in the raw captured measurements with different scales. The fundamental principle of wavelet transform is the sparse representation of the temporal signal that means the signal can be presented in the limited large-magnitude coefficient and the small value wavelet coefficient are related to noise which can be removed without affecting the original signal of interest. After setting the thresholds, the signal is reconstructed with inverse wavelet transform. In this work, Wavelet transform is utilized for D-PMU signal denoising.

2.2.2 Hampel Filter (HF)

Hample filter evaluates data in a sliding window of $2n'$ neighbor samples with n' in each side. The median absolute deviation (MAD) is computed as follows:

$$MAD = Median(|\bar{X} - Median(X)|) \tag{2.1}$$

where $\bar{X}$ represents the window in which the central element is the sample of interest x and $Median(.)$ calculates the local median in the input array. MAD provides an estimation of standard deviation on the window $\sigma = 1.4826MAD$.

If the difference between the sample and the median of the window, is bigger than three times of the estimated standard deviation σ, It is considered an anomaly as a convention for Hampel filter. For obtaining the margin of BD, the distance on the sample from the median is computed as $\delta x = x - Median(X)$. The logarithm of dividing this value to three times of estimated standard deviation gives a metric about the margin of the possibility of being an outlier. If this metric is positive, the sample is considered an outlier and vice versa. In the case that metric is positive, The bigger the metric, the more possibility to be an outlier from the Hampel filter perspective. If the value is close to zero, the Hampel filter is mostly indecisive about the sample. Hampel is prone to detect a false bad data anomaly in steady-state since the standard deviation in the window is fairly small and a tiny deviation from the median can be considered as a bad data anomaly. In that case, replacing with median will not be harmful. However, since we are interested to detect anomalies with different approaches and combine the results based on MB-MLE integration, we define the small constant ϵ_h to reduce the chance of detecting outlier in steady-state. Therefore margin bad data HF defines for each sample as

$$Margin_{HF} = \log(\frac{\delta x}{3\sigma + \epsilon_h}) \tag{2.2}$$

2.2.3 Quartile-based Anomaly Detection (QB)

Another statistical approach for anomaly detection is quartile-based anomaly detection. The advantage of this method is the fact that it does not have a normal distribution assumption of the data. To implement, data is analyzed in small windows, and the median and quarter of the window are computed. The sample is marked as an anomaly if they are 1.5 times bigger than the upper quartile or if they are 1.5 times below the lower quartile. For obtaining the margin metric of being an outlier, the distance of each sample from the median is computed (δx_i) and divided by the distance of the corresponding quartile from the median (δq). The logarithm of this division gives insight into how likely the sample is an outlier.

Similar to Hampel filter, QB is also prone to identify false positives in steady-state condition. By the same token, the constant ϵ_q is added to the denominator of the fraction. Therefore, margin quartile method can be computed for each sample as

$$Margin_{QB} = \log(\frac{\delta x_i}{\delta q + \epsilon_q}) \tag{2.3}$$

2.2.4 *Density-based Spatial Clustering of Applications with Noise (DBSCAN)*

DBSCAN has been deployed for determining outliers in finding the disturbances using the outputs of an unsupervised model [25]. DBSCAN uses two tunable parameters including threshold ϵ and Minimum Number of Points (MinPts). The data points within ϵ radius of cluster are merged into a single cluster, with the assumption of MinPts data points required to form a cluster. The data points without clusters are determined as outliers.

In this work, to find the margin-based value, the parameter threshold ϵ which makes each measurement categorized as an outlier is calculated based using grid search. Subsequently the parameter $Margin_{DBSCAN} = \log \frac{\epsilon}{\epsilon_0}$ is found to determine the margin of begin outlier with the DBSCAN method for each sample. The value of ϵ_0 can be estimated based on results from ground truth data.

2.2.5 *Margin-Based Maximum Likelihood Estimator (MB-MLE)*

Different calculated margins from base-detectors are combined together to make a robust decision about the likelihood of being bad data anomaly for each sample as follows:

$$MLE = \mathcal{N}(\sum_{i \in \mathfrak{B}} c_i Margin_i) \tag{2.4}$$

where set $\mathfrak{B} = \{HF, QB, DBSCAN\}$ shows the available based detectors and c_i is associate confidence factor which can be obtained based on historical data when the ground truth is know or using majority voting to estimate the ground truth, when the labels are not available. For the purpose of the current work, we assume $c_i = 1$ for all $i \in \mathfrak{B}$. Positive MLE indicates anomaly and vice versa and the biggest deviation from zero, indicates more confidence of MB-MLE bad data anomaly detector.

2.3 Online Tool

Although the introduced offline tools in the previous section, can determine the anomalies and reduce noise effect with smoothing techniques with good accuracy. The implementation of those techniques is time-consuming and they might not be suitable for the application when fast decision making is needed. Considering these facts, in this section, we develop an online D-PMU pre-processing tool using Koopman Mode Analysis (KMA) and adaptive Kalman Filter (KF) by adjusting the KF parameters based on the confidence of the fitted model with online monitoring. This technique uses both temporal and spatial correlations between synchrophasor measurement for signal processing.

2.3.1 Kalman Filter

KF is an optimal recursive bayesian estimator. If the assumptions of the KF (Linear Dynamic, Gaussian process and measurement noise) is held, it surpasses any other casual filter. It consists of two steps namely prediction and correction. In the prediction step, the model of the system is used to predict the expected value and uncertainty of the next step measurement.

$$\begin{aligned} \hat{x}_{n|n-1} &= A\hat{x}_{n-1|n-1} + Bu_n \\ P_{n|n-1} &= AP_{n-1|n-1}A' + Q_n \end{aligned} \tag{2.5}$$

Where A is the linear dynamic of the system, B is signature of the input, and u_n is control input at time n.$\hat{x}_{n-1|n-1}$ and $P_{n-1|n-1}$ shows the prior believe on the expected value and covariance of the estimated states. $\hat{x}_{n|n-1}$ and $P_{n|n-1}$ represents those at time n based on prediction and Q_n is process noise covariance.

In the correction step, the value of prediction is modified based on new observation, and the uncertainty of perdition reduces based on the characteristics of the observation.

$$
\begin{aligned}
S_n &= CP_{n|n-1}C' + R_n \\
K_n &= P_{n|n-1}C'{S_k}^{-1} \\
\hat{x}_{n|n} &= \hat{x}_{n|n-1} + K_n(y_n - C\hat{x}_{n|n-1}) \\
P_{n|n} &= P_{n|n-1}(I - K_nC)
\end{aligned}
\tag{2.6}
$$

Where C is signature of output. R_n is covariance of measurement noise and S_n characterizes the uncertainty of measurement. K_n is Kalman gain, y_n is measurement at time n and $\hat{x}_{n|n}$ and $P_{n|n}$ shows the posterior believe on the expected value and covariance of estimated states after observation.

In the proposed formulation, y is constructed by direct and indirect measurements from different D-PMU contaminated with noise and bad data anomaly and x indicates measurement without contamination and $\hat{x}$ is estimation of x, B is equal to zero and C equal to identity matrix with appropriate size therefore y has same size as x. In the next subsection, we explain how to obtain linear dynamic model A and adjust parameters R_n and Q_n attentively.

2.3.2 Koopman Mode Analysis

Koopman Theory asserts that any underlying nonlinear system can be described completely with infinite state space [19]. To find a finite approximation of the Koopman states, many data-driven mechanisms are suggested. In this work, we use Dy-

namic Mode Decomposition (DMD) to this end. DMD is a data-driven approach for fitting a linear dynamic to a set of measurements [26, 27]. Consider that the matrix $Z_1^{k'} = [z_1, ..., z_{k'}]$ represents a window of k' snapshots consisting n measurement at each snapshot where vector z_i shows the ith snapshot. The matrix $Z_1^{k'}$ is constructed by stacking different D-PMU's direct measurement data streams such as voltage/current magnitude/angel, and frequency and indirect measurement data stream including active/reactive power flow on each other. The notation Z_p^q shows a set of subsequent data where subscript p is the index of the starting snapshot and superscript q is the index of the last snapshot in the window. In this work, the assumption is that the sampling distance between every two successive snapshots is constant and is shown with δ. Suppose there exist a linear dynamic A, mapping data sample z_i to data sample z_{i+1} for $i = 1, \cdots, k' - 1$ as

$$z_{i+1} = Az_i \tag{2.7}$$

In general, measurements are generated from a nonlinear underlying dynamic. Using DMD, we attempt to find the best linear dynamic approximation that describes the relationship between the measurements over the processing window. In other words, we try to minimize the residual of the linear system defines as

$$r = Z_2^{k'} - AZ_1^{k'-1}{}_2 \tag{2.8}$$

Where $.2$ shows norm 2 of matrix. A naive approach to compute linear system is $A = Z_2^{k'} {Z_1^{k'-1}}^+$ where ${Z_1^{k'-1}}^+$ is Moore-Penrose pseudo-inverse of $Z_1^{k'-1}$. The matrix inverse can be obtained via the LQ method. Nonetheless, Singular Value Decomposition (SVD) allows more robust numerical stability [28]. However, it is not computationally efficient to calculate A directly because k' might be large. Therefore, the DMD algorithm finds eigen structure linear dynamic without computing it directly. By employing SVD, following equation is obtained:

$$Z_1^{k'-1} = U\Sigma V^* \tag{2.9}$$

Regardless of the possible hug dimensions of A, in most cases, the underlying dynamic of the window can be delineated with a few m dominant modes which are of the notable portion of event energy. Using these modes, the equivalent linear dynamic $\tilde{A}$ is computed as follows

$$\tilde{A} = U_m^* Z_2^{k'-1} V\Sigma^{-1} \tag{2.10}$$

Where U_m shows first m columns of U. $\tilde{A}$ has same eigenvalues as A. By eigenvalue decomposition of $\tilde{A}$ we have $\tilde{A}W = \Lambda W$ where diagonal matrix Λ contains eigenvalues of A and $\tilde{A}$ matrices. Columns of W are eigen vectors of $\tilde{A}$. Eigen vectors of A are computed as $\Psi = V_2^{k'} V\Sigma^{-1}W$. The jth diagonal entity of Λ indicates eigenvalue of jth mode shown by λ_j and corresponding eigen vector is jth column of

matrix Ψ denoted by ψ_j.

To Compute energy amplitude of each mode, SVD of the data window is obtained as $Z_1^{k'} = U_0 \Sigma_0 V_0^*$. The matrix $\Lambda_t = \exp([\lambda_{1,t}, \cdots, \lambda_{m,t}]^T)$ indicates evolution of each mode in the window where $\lambda_{i,t} = \lambda_j \bar{T}$ and $\bar{T} = [0, h, 2h, \cdots, (k'-1)h]^T$. Subsequently, we calculate matrix $\Xi_1 = \Upsilon^T \Upsilon \odot (\Lambda_t \Lambda_t^T)^*$ where $\odot$ denote Hadamard product operator, superscript $*$ indicates Hermitian transpose operator and $\Upsilon = U_0^T \Psi$. Further, $\Xi_2 = Diag(\Lambda_t V_0 \Sigma_0 \Upsilon)^*$ where the operator $Diag(.)$ place the diagonal entries of square input matrix into a vertical vector. Finally, m-length vector $\Xi = \Xi_1^{-1} \Xi_2$ is computed where absolute value of the jth element of Ξ is equal to σ_j which shows energy amplitude jth mode of linear system.

For finding the linear system A, representing the dynamic behaviour of the system, KMA is used by inputting a comprehensive set of historical data after preprocessing with offline anomaly and denoising techniques that were introduced in section 2. From each D-PMU direct measurements such as voltage/current magnitude/angle and frequency and indirect measurements such as active and reactive power flow in different phases are used to find the linear model. The reason for adding indirect measurement is that it is proved that by adding more observant, the better linear system can be fitted to describe the underlying dynamic of the system [26]. The fitted linear model, estimated by KMA, is further used for prediction in online phase using KF, for denoising and bad data detection.

2.3.3 Adaptive Adjustment of Kalman Filter Parameter

With the assumption of having Gaussian Noise, the mean of estimated value and measurement value overtime must have the same values. This can be used as an indicator that the fitted model A, does not describe the dynamic to the system precisely. The deviation of mean in all estimated measurements from noisy measurement over a is calculated as $N[n]$ in each sample n. The average of that over that in short interval on recent samples with length 5 is called $\bar{N}$. This metric is used to check the credence of the fitted linear model. Using this parameter, the metric ρ is defined as

$$\rho = \frac{2\sigma_{max}}{\exp\{-\sigma_{steep}\bar{N}\} + 1} - \sigma_{max} + 1 \tag{2.11}$$

In the above function, ρ goes to 1 as when N is zero and it goes to $\sigma_{max} + 1$ as N goes to ∞. σ_{steep} determines how fast this transition occurs. The process noise of the KF is then updated attentively as $Q = \rho Q_{base}$. It needs to be clarified, that A is computed offline and used in online pre-processing using KF. However, if ρ remains bigger than a preset threshold for a long period of time (1 day), that means that power grid topology or dynamic has been changed significantly and a new linear model A required to be fitted to obtain good results with online tool.

2.3.4 Online model update scheme

If the value of parameter N stays vastly out of range for a considerable amount of time, it can be interpreted that the linear model A cannot describe the behavior of the system anymore and it is essential to retrain the model. This will trigger offline analysis automatically to compute a new linear dynamic A. For that, we compute the parameter $\tilde{N} = \sum_{k=n-n_0}^{n} N[k]$ where $[n - n_0, n]$ indicates the time window that we accumulate the average deviation of mean in all estimated measurement from noisy measurement over. n_0 is chosen to have a time window that covers half and hour and the model is retrained if the $\tilde{N}$ if bigger that the threshold γ.

2.4 Testbed and Validation

2.4.1 Testbed and Usecase Development

For the evaluation purposes of the developed algorithms in this work, IEEE 33-node test feeder is chosen to be modeled and simulated in Opal-RT hardware-in-the-loop (HIL) simulator.

Fig. 2.4 shows the single-line diagram of the developed IEEE-33 node system in OPAL-RT with extra modifications compared to the standard model. The HYPERSIM interfacing simulator is used for software aspect of the model development

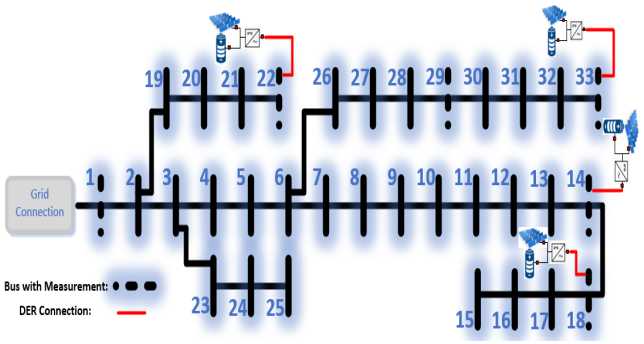


Figure 2.4: The modified IEEE 33-node system with D-PMU and DERs

with capabilities of DER integration and D-PMU simulation. The model is capable of being executed in real-time with 50 μs timestep and with hardware D-PMUs added to the loop of the simulation. The simulated Battery Energy Storage Systems (BESS) and PV units are modeled to be operating under the grid-forming configuration with maximum generation capability of 500kW each, making the distribution feeder to be able to operate under a 100% renewable scenarios in an islanded operation. The modifications on the standard IEEE system include adding four BESS and PV combined energy resources at four different loaded nodes of the feeder. The network has three radial feeders with possible mesh interconnection among themselves. Simulation D-PMUs are modeled and placed at generation nodes: $\{1, 14, 18, 22, 29, 33\}$.

Furthemore, load profiles are modeled to be connected at all the 33 buses with sinusoidal deviation of maximum 5% of the nominal value compared to the standard IEEE system. The monitored instantaneous time domain signals captured from the

Table 2.1: Stationary Operating Scenarios

Operating Scenario 1	All the renewables dispatching to the maximum capacity
Operating Scenario 2	All the four BESS discharging with 50% SOC
Operating Scenario 3	Zero renewable (loads being supplied by the grid)
Operating Scenario 4	BESS number 3 located at bus 22 is 50% charging

OPAL-RT solver are sent to simulated D-PMUs for phasor estimations.

A series of load flow stationary scenarios as well as a comprehensive list of practical dynamic events have been modeled and simulated and all the time-stamped measurements are fed to the algorithms for efficacy investigations. The stationary use case development is to have the system operate under the edge scenarios of either all the power being supplied from the grid or the system maintains the load by its own in an islanded operation. Table-2.1 shows a list of load flow operating scenarios, under which case several dynamic events have been simulated.

A wide range of dynamic use case scenarios have been modeled, including

breaker operations, load and capacitor bank switching, generations step up and down, different types of short circuited bolted and shallow faults as well as islanding events. The underlying idea behind the simulated anomalies is to cover a variety of events with distinct frequency, voltage, and current response. Corresponding D-PMU outputted measurements have been collected and pre-processed, further on timestamped measurements have been used for offline and online validations.

Additional emphasis is put towards three-phase to ground bolted fault, DER, and False Data Injection (FDI) Cyber-attack event. Table.4.1 summarizes the real-time simulated anomalies with more focus:

Table 2.2: Opal-RT simulated Dynamic Events

Event Type	Location	Total Number of Events
Bolted Fault	all the 33 buses	297
DER Switching	buses {14, 18, 22, 33}	36
FDI Cyber-attack	buses {1, 14, 29}	27

2.4.2 Simulation and Validation Analysis

For introducing noise and bad data to synthesis data from the OPAL-RT, the additive white Gaussian noise with the variance of $0.01 \times \mu_i$ added to each measurement i, where μ_i is the mean of that measurement. Non temporal-spatial anomaly bad data is acted uniformly to the measurements with rate 0.001 and change the value of the measurement to value $\mu_i + \kappa_i$ where random variable κ_i is chosen from interval $[-\eta_i, -0.5\eta_i] \cup [0.5\eta_i, \eta_i]$ with uniform distribution where $\eta_i = max(10\sigma_i, 0.2\mu_i)$ and σ_i and μ_i are standard deviation and mean of measurement i respectively.

Fig. 2.5 shows the performance of suggested offline wavelet smoothing and MB-MLE BD anomaly detection and compensation technique on the contaminated data with noise and BD anomaly. As can be seen, all of the bad data are detected and eliminated from compensated data (at seconds 10.58, 41.73, 51.42), and also, the noise level has been reduced. The proposed offline tool increases the signal to noise ratio (SNR) 15.241 dB for D-PMU 1 voltage magnitude measurement phase a (Fig. 2.5). Table 2.3 shows the performance of each base anomaly detector and their MB-MLE assemble statistically in terms of the confusion matrix, recall, and precision. Recall is defined as $Recall = \frac{TP}{TP+FN}$ and precision is defined as $Precision = \frac{TP}{TP+FP}$. As can be seen, the MB-MLE has improved both Recall and Precision compare to base detectors.

Table 2.3: Statistical comparison of different methods, Confusion matrix (True positive (TP), False positive (FP), False negative (FN), True Negative (TN)), Recall, and Precision

Method	TP	FP	FN	TN	Recall	Precision
HF	861	205	13	788713	0.9851	0.8077
QB	861	120	13	788798	0.9851	0.8777
DBSCAN	859	272	15	788646	0.9828	0.7595
MB-MLE	866	31	8	788887	0.9908	0.9654

Moreover, a comparative study is conducted for the performance evaluation on different types of the events. *Accuracy* and *Precision* are used as evaluation metrics and events are same as the ones in Table 4.1. For any particular event, two scenarios are considered as:

- **Scenario 1:** D-PMU data without noise and bad data used as base case
- **Scenario 2:** Manipulated D-PMU data with added noise and bad data and mitigation using the developed architecture

For all the event cases, noise and bad data processing performance in "Scenario 2" is compared with "Scenario 1". Table. 2.4 summarizes the event-based performance evaluation of the developed DPM-DBDM: **D**istribution **P**hasor **M**easurement

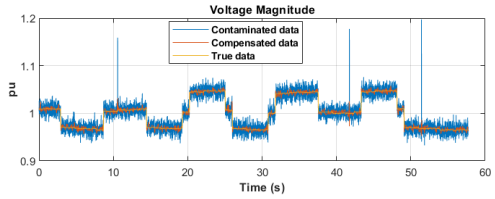


Figure 2.5: The result of offline pre-processing for denoising and bad data compensation using MB-MLE and wavelet analysis on D-PMU 1, voltage magnitude measurement phase a

Denoising, **B**ad Data **D**etection, and **M**itigation.

Table 2.4: Event-based Performance Evaluation of DPM-DBDM

Event Type	Accuracy	Precision
Bolted Fault	0.9872	0.9148
DER Switching	0.9733	0.8972
FDI Cyber-attack	0.9285	0.9005

For online analysis, the signal is processed with offline tools and a linear dynamic is fitted to describe the linear dynamic. The linear dynamic is subsequently used to reduce the noise of the system and compensate for bad data (at seconds 0.41, 8.45, 9.31, 17.29, 14.09, 46.1, and 57.41). Fig. 2.6 shows that the introduced system can reduce the effect of noise and bad data significantly and SNR increases 9.78 dB for

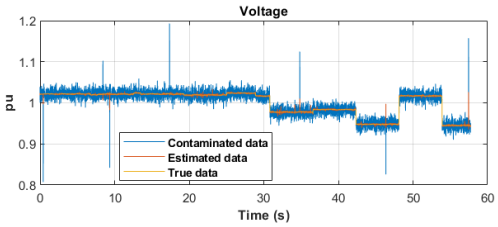
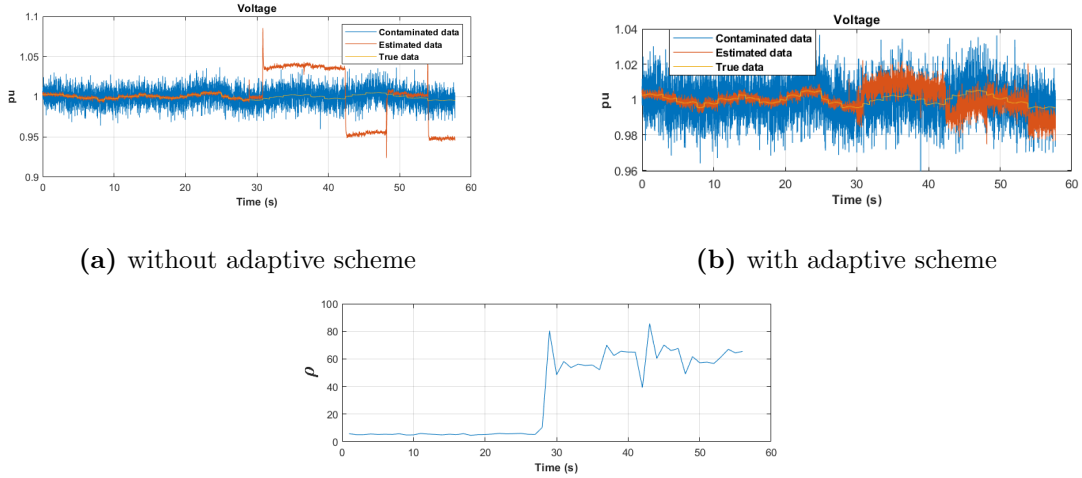


Figure 2.6: The result of online denoising and bad data compensation using adaptive Kalman filter on D-PMU 2, voltage magnitude measurements, phase b

voltage measurement D-PMU 2 phase b (Fig. 2.6). Although with online KF, BD is not replaced through interpolation as it happens in the offline tool, it yet reduces the magnitude of BD anomaly significantly since it utilizes the KMA-based fitted model for prediction and filter uncorrelated spatial and temporal jumps in the measurement.

When the dynamic behavior has not been seen in fitting the linear model A, the process noise parameters of the system are adjusted adaptively to avoid relying on non-precise parameters for prediction and rely more on the noisy measurement.

In Fig. 2.7 the performance of KF is shown when it is exposed to the dynamic events that have not seen before after 30 seconds. In 2.7a no compensation has been made for that and as can be seen, relying on an imprecise model leads to big bias. However, in 2.7b the deviation in the mean of estimated value and measurements are monitored and adjust the parameters of the KF. So after 30 seconds, we have a noisier estimation, however, we reduce the bias from the real value. That is because we trust



(a) without adaptive scheme

(b) with adaptive scheme

(c) parameter ρ

Figure 2.7: Online denoising with D-PMU with Kalman Filter, voltage angle of D-PMU 3 phase a. after 30 second unforeseen dynamic event in the system

less on the inexplicit model for prediction and rely more on the noisy measurement. 2.7c shows the ρ parameter where $Q = 0.1\rho I$, $R = I$ and I indicates the identity matrix with appropriate dimension.

Although offline tool provides more precise denoising and BDD than online KMA-based KF filter, it comes with the cost of having more expensive computational time. For example, in our simulation using MATLAB software with Core i7 computer, for processing 114 data stream for 60 seconds with the sample rate of 120 data points per second, it takes 145.642 seconds for the offline tool to process the data. However, the same process is done in 2.592 seconds using the online tool. Therefore, based

on the application, the appropriate tool can be selected for pre-processing of phasor measurements.

2.5 Summary

In this work, two novel approaches are proposed for denoising and bad data detection in distribution Phasor Measurement Units (D-PMU) data for offline (longer time window data processing) and online (shorter time window data processing) applications. Multiple base detectors of different types including Hampel filter, Quartile detector and DBSCAN are developed and integrated using a margin-based maximum likelihood estimator (MB-MLE). High-frequency noises are processed and removed using a wavelet denoising method. Koopman Mode Analysis is used to fit a model to the underlying dynamics based on an offline analysis and used for online denoising and bad data detection using Kalman Filter (KF) with adaptively adjusted parameters. Evaluation of performance using synthetic and realistic data generated by the modified IEEE 33-bus test system with multiple D-PMUs, simulated in real-time using OPAL-RT confirms the practicality of the proposed scheme. For the next step, the validation can be further elevated using a large distribution network with a decentralized approach to divide the network into local zones to test the scalability of the proposed scheme. Also, online parameter adjustment of adaptive KF can be done

using deep autoencoder and dynamic features to enhance the fast response of the adjustment loop before deviation causes modification in the parameters.

REFERENCES

[1] B. C. de Oliveira, I. D. Melo, and M. A. Souza, "Bad data detection, identification and correction in distribution system state estimation based on pmus," *Electrical Engineering*, pp. 1–17, 2021.

[2] L. Liu, D.-h. Zhai, and X. Jiang, "Current situation and development of the methods on bad-data detection and identification of power system," *Power System Protection and Control*, vol. 38, no. 5, pp. 143–147, 2010.

[3] X.-p. Ni and B.-h. Zhang, "A state estimation method for bad data detection and identification based on equivalent current measurement transformation [j]," *Power System Technology*, vol. 26, no. 8, pp. 12–15, 2002.

[4] Y. Gu, T. Liu, D. Wang, X. Guan, and Z. Xu, "Bad data detection method for smart grids based on distributed state estimation," in *2013 IEEE International Conference on Communications (ICC)*, pp. 4483–4487, IEEE, 2013.

[5] S. Som, R. Dutta, A. Gholami, A. K. Srivastava, S. Chakrabarti, and S. R. Sahoo, "Dpmu-based multiple event detection in a microgrid considering measurement anomalies," *Applied Energy*, vol. 308, p. 118269, 2022.

[6] S. Wang, L. Li, and P. Dehghanian, "Power grid online surveillance through

PMU-embedded convolutional neural networks," in *2019 IEEE Industry Applications Society Annual Meeting*, pp. 1–8, 2019.

[7] J. Chen and A. Abur, "Placement of pmus to enable bad data detection in state estimation," *IEEE Transactions on Power Systems*, vol. 21, no. 4, pp. 1608–1615, 2006.

[8] V. S. Bharath. Kurukuru, A. Haque, and M. A. Khan, "Fault classification for photovoltaic modules using thermography and image processing," in *2019 IEEE Industry Applications Society Annual Meeting*, pp. 1–6, 2019.

[9] A. Gholami, A. K. Srivastava, and S. Pandey, "Data-driven failure diagnosis in transmission protection system with multiple events and data anomalies," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 4, pp. 767–778, 2019.

[10] S. J. Hossain and S. Kamalasadan, "Combined deterministic-stochastic online subspace identification for power system mode estimation and oscillation classification," in *2019 IEEE Industry Applications Society Annual Meeting*, pp. 1–9, 2019.

[11] A. Gholami and A. K. Srivastava, "Comparative analysis of ml techniques for data-driven anomaly detection, classification and localization in distribution system," in *2021 North American Power Symposium (NAPS)*, 2021.

[12] M. Cramer, P. Goergens, and A. Schnettler, "Bad data detection and handling in distribution grid state estimation using artificial neural networks," in *2015 IEEE Eindhoven PowerTech*, pp. 1–6, 2015.

[13] Y. Peng, W. Qiao, L. Qu, and J. Wang, "Sensor fault detection and isolation for a wireless sensor network-based remote wind turbine condition monitoring system," in *2017 IEEE Industry Applications Society Annual Meeting*, pp. 1–7, 2017.

[14] A. Vosughi, A. Gholami, and A. K. Srivastava, "Denoising and bad data detection in distribution phasor measurements using filtering, clustering and koopman mode analysis," *IEEE IAS Annual Meeting*, Oct 2021.

[15] S. Avdakovic, A. Nuhanovic, M. Kusljugic, and M. Music, "Wavelet transform applications in power system dynamics," *Electric Power Systems Research*, vol. 83, no. 1, pp. 237–245, 2012.

[16] Y. An and D. Liu, "Multivariate gaussian-based false data detection against cyber-attacks," *IEEE Access*, vol. 7, pp. 119804–119812, 2019.

[17] F. A. Ghaleb, M. B. Kamat, M. Salleh, M. F. Rohani, and S. Abd Razak, "Two-stage motion artefact reduction algorithm for electrocardiogram using weighted adaptive noise cancelling and recursive hampel filter," *PloS one*, vol. 13, no. 11, p. e0207176, 2018.

[18] H. Liu, F. Hu, J. Su, X. Wei, and R. Qin, "Comparisons on kalman-filter-based dynamic state estimation algorithms of power systems," *IEEE Access*, vol. 8, pp. 51035–51043, 2020.

[19] I. Mezić, "Spectral properties of dynamical systems, model reduction and decompositions," *Nonlinear Dynamics*, vol. 41, no. 1-3, pp. 309–325, 2005.

[20] Y. Susuki, I. Mezic, F. Raak, and T. Hikihara, "Applied koopman operator theory for power systems technology," *Nonlinear Theory and Its Applications, IEICE*, vol. 7, no. 4, pp. 430–459, 2016.

[21] M. Netto, *Robust Identification, Estimation, and Control of Electric Power Systems Using the Koopman Operator-Theoretic Framework*. PhD thesis, Virginia Tech, 2019.

[22] R. Dubey, S. R. Samantaray, B. K. Panigrahi, and V. G. Venkoparao, "Koopman analysis based wide-area back-up protection and faulted line identification for series-compensated power network," *IEEE Systems Journal*, vol. 12, no. 3, pp. 2634–2644, 2016.

[23] S. P. Nandanoori, S. Kundu, S. Pal, K. Agarwal, and S. Choudhury, "Model-agnostic algorithm for real-time attack identification in power grid using koopman modes," *arXiv preprint arXiv:2007.11717*, 2020.

[24] A. Gholami, M. Mousavi, A. K. Srivastava, and A. Mehrizi-Sani, "Cyber-physical vulnerability and security analysis of power grid with hvdc line," in *2019 North American Power Symposium (NAPS)*, pp. 1–6, 2019.

[25] S. Pandey, S. Chanda, A. Srivastava, and R. Hovsapian, "Resiliency-driven proactive distribution system reconfiguration with synchrophasor data," *IEEE Transactions on Power Systems*, 2020.

[26] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *Journal of fluid mechanics*, vol. 656, pp. 5–28, 2010.

[27] M. R. Jovanović, P. J. Schmid, and J. W. Nichols, "Sparsity-promoting dynamic mode decomposition," *Physics of Fluids*, vol. 26, no. 2, p. 024103, 2014.

[28] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz, "On dynamic mode decomposition: Theory and applications," *arXiv preprint arXiv:1312.0041*, 2013.

CHAPTER 3. ENHANCE AWARENESS BY ANOMALY DETECTION, CLASSIFICATION, AND STATE ESTIMATION

3.1 Introduction

This chapter of the research aims to describe the methods used to distinguish anomalies from regular data using statistical, clustering and outlier factor based methods. Detailed investigation of DBSCAN, K-Means, Local Outlier Factor algorithm, Feature Bagging algorithm, and Robust Random Cut Forests are performed with help of real distribution scale system. It proposes the ensemble approach to leverage the merits of individual method for achieving the highest detection accuracy in comparison to solo method.Besides, research also discusses introduction of clustering based classification to make clear distinction between identified event sets considering the fundamental rules of the power system parameters. We analyzed data from multiple Phasor Measurement Unit devices (PMUs) for this exercise that are extracted from Bronzeville Community Microgrid consisting. This work draws higher level contributions addressing the existing issues from different techniques, and also provide the key advantages over one another.

A recent trend in distribution systems observed that high proliferation of the large scale deployments of distributed energy resources at each voltage level and consumer sectors across the globe. This resulted into the new era of distribution system that left sign of unprecedented scenarios and events incomparable to historical data. As a consequence, highly sophisticated and modern sensor technology penetrated through industry for capturing the unmatched event signatures. This also led to deployment of distribution phasor measurement unit (D-PMU) to liberate the timestamped phasor data for analyzing the dispersed event signatures and their relations, which analogous to PMUs in transmission to prevent catastrophic failures[**?**]. High resolution data is inherited from D-PMU and made available for system operators to extract useful information through data-driven techniques[**?**]. Various techniques have been reported in the literature to detect, localize and classify the significant events in the distribution systems.

3.1.1 Problem Statement

The problem pertinent to accurate and reliable event detection techniques without model or network information is still relevant to the industry and utilities, as existing methods mostly rely on the training data sets and its accuracies with respect to real-time emulation of the same set of events in the lab. However, there is no guaran-

tee of detection efficiencies since systems lose their behavior models in the process of modeling. Moreover, Supervised and semi-supervised demands real data for achieving the better performance and also need event labels further to prove their outstanding performances in case of localization and classification. This is always challenging to obtain from utilities. This chapter proposed the unsupervised methods to fill the research gap by considering two different types of methods i.e. outlier factor and clustering techniques. Uniqueness of this work is to propose the ensemble approach to leverage the individual approach merits and meet high accuracy in detection that drives reliability of classification.

3.2 Related Works

The advancement and deployment of Phasor Measurement Units(PMUs) across the power system improves the situational awareness of the system. On one hand, they can provide useful information regarding events, on the other hand the data packages could be prone to interruption. So, researchers have proposed various approaches especially statistical and machine learning based techniques for Anomaly detection and classification. The authors of [**?**] developed a Synchrophasor Anomaly Detection and Classification (SyADC) tool based on a hierarchical structure composed of three unsupervised methods- isolation forest, Kmeans & LoOP. The Data Streams

from multiple PMUs were fed to the isolation forest detector which conveys the generated scores to the other detectors to classify the incoming data as outlier, event, PDC (Phasor Data Concentrator) error or normal data. Likewise, an unsupervised clustering technique based on Principle Component Analysis (PCA) is introduced in [**?**] for power system event detection, classification and localization. Furthermore, the work in [**?**] incorporates unlabeled, partly labeled and completely labeled data from μPMUs and proposes a Hidden Structure Semi-Supervised Machine Learning (HS^3M) Method to detect and classify events in power system. Besides, a supervised learning approach has been proposed in [**?**] based on explicit labelling through utility records and expert knowledge. In [**?**], a matrix partitioning based online algorithm has been built to identify not only the type but also the location of events in distribution grid via change in admittance matrix that corresponds to topology change. The authors of [**?**] present a strategy of applying shape based Fisher-Rao Registration Method (FRMM) to align multiple time signals that improves the feature selection process by clustering algorithm to better detect and classify events in power system. While, the aforementioned research works mostly scrutinize the performance of a particular approach on anomaly detection or classification, a consolidated method unifying the performance of multiple statistical, clustering and LOF based techniques and leveraging their advantages through an ensemble approach is yet to be explored in the existing literature of power system application. Since the mea-

surements from metering devices, sensors and PMUs go into different power system application such as state estimation function which further impact significantly on various decision making processes, it is imperative to accurately detect and classify anomalies for an effective operation and control of power system. The authors of [**?**] develops a distribution system state estimation method utilizing branch currents measured through Advanced Metering Infrastructure (AMI). Research in [**?**] proposes a neural network based learning approach for state estimation. In addition to that, some researchers have discussed the strategies and necessities of pseudo measurements for efficient estimation of power system states [**?** **?**]. But the above-listed works do not comment on the robustness of the state estimation function in case of anomalies and outliers. Consequently, the current work first focuses on the ensemble approach for improved efficiency of anomaly detection and classification followed by enriching the performance of state estimation function resilient to anomalies.

3.2.1 Contributions

Contribution of the research includes

1. Development of Ensemble anomaly detection technique to capture the diverse events through combination of clustering and scoring based methods with improved robustness

2. Proposal of an enhanced State Estimation methodology based on Weighted Least Square (WLS) method with anomaly detection

3. Development of a modified event classification method is developed to categorize real events with simulated utility data.

4. Streamlining all the modules to produce a single package.

3.3 Overall architecture of the Proposed solution

The architecture in Fig.3.1 depicts the actual way of implementation of all the methods in sequential manner for each data processing window. This outlines the real information exchange and also transmission of data from one block to another. First, data will be passed to preprocessing stage where the filtering of bad data is performed from raw-data and then send to dimensional reduction depending upon requirement from each base detector.Filtered data is passed to four major base detectors (i.e. DBSCAN, K-means, Net-LOF and RRCF as described in previous D1 and D2-D3 reports) for segregating the anomalous list and normal operational data. Anomalous data is fed to base anomaly classification stage for sorting out the real events and outliers.Thereafter, the true event list is given as input to the broader level classification module to categorize the events according to their classes. The aforementioned procedure will be repeated over definite time period of data cycles.

There is always scope for making the entire process in parallel processing windows so that interested data window will be checked computationally less intensive fashion.

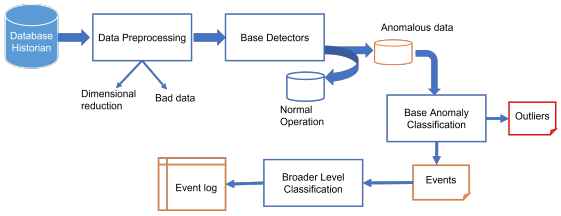


Figure 3.1: Data flow diagram for the implementation

3.4 Methodology of anomaly detection and State Estimation approaches

3.4.1 Clustering Methods

3.4.1.1 DBSCAN

Density-based Spatial clustering of applications with noise (DBSCAN) is a density-based clustering algorithm that is used to group the close points and mark the low density points far apart as anomalies [**?**]. It has been deployed in D-PMU anomaly detection scheme as to determine outliers in finding the disturbances from the signal [**?**]. DBSCAN has two hyper-parameters - epsilon optimal distance (ϵ) and minimum number of points ($minPts$). If at least $minPts$ points are within distance ϵ of a point p, then p is assigned as the core point. For each point q in this space is with lower density and cannot be directly reachable, then this data sample q will be

marked as an outlier.

3.4.1.2 K-Means

K-Means Clustering anomaly detection is a signal processing-driven vector quantization approach. In this method, the effort is needed towards partitioning the n number of total observations into k number of clusters, with the condition of observations being assigned to the closest cluster based on the calculated mean distance. As far as the scope of this project, "Centroids" are defined as the observation, containing the most impact of the anomaly, and the other observations are being compared to the "Centroids" using Euclidean distances. With x_i being every single observation and S_j the corresponding cluster, K-means objective is to be formulated as the following equation:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i \tag{3.1}$$

Elbow Method: A primary requirement for the aforementioned clustering approach, is the pre-determined number of clusters, which in the case of this project, might not be feasible to define. For this purpose, Elbow Method is used by plotting of the variations of distances versus the number of clusters. This method, aims towards finding the knee point of the curve, where the increase in the number of clusters, does not improve in reducing cluster distance. The plotted distances, is based on the following equation:

$$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 \tag{3.2}$$

3.4.2 Scoring Methods

3.4.2.1 Local Outlier Factor with Feature Bagging (Net LOF)

Local Outlier Factor is an unsupervised anomaly detection algorithm that is based on the concept of "Local Density" of a point and it's n neighbors [**?**]. We compare the density of a point with respect to the density of it's neighbors. Regions containing data points that are clustered together each have a higher local density compared to regions that are sparsely populated. To detect anomalies, this algorithm assumes that data points that have a high local density are more correct / normal than those data points that have a lower local density.

We calculate Local Outlier Factor as follows:

1. Let the n-dist(p) for a data point 'p' be the distance of the data point to it's n^{th} nearest neighbor. Let us denote the set of n neighbors for p as $S_n(p)$.

2. We can now use the definition of n-dist(.) to calculate the "reachability distance", denoted as r-dist(.) of a point 'q' from point 'p', as

$$\text{r-dist}_n(p, q) = max(\text{dist}(p, q), \text{n-dist}(q)) \tag{3.3}$$

where dist(p, q) is any distance function, preferably euclidean distance.

3. Using r-dist(.), we can now calculate the local reachable density of a data point p with respect to it's nearest n neighbors as:

$$lrd_n(p) = \left[\frac{1}{|S_n(p)|} \sum_{q \in S_n(p)} \text{r-dist}_n(p, q) \right]^{-1} \tag{3.4}$$

that is, the inverse of the mean of the reachability distance of the nearest n points from the data point p.

4. Given Equation (3.4), we can now define the Local Outlier Factor for a data point p as the average of the ratio of the local reachable distance of all the nearest neighbors of p to the local reachable distance of the data point p. This is calculated as

$$lof_n(p) = \frac{1}{|S_n(p)|} \left[\frac{\sum_{q \in S_n(p)} lrd_n(q)}{lrd_n(p)} \right] \tag{3.5}$$

If we use the LOF algorithm without making any changes to the dataset, we call the approach **single LOF**

Feature Bagging is a feature selection algorithm used to reduce correlations present between features of the dataset [**?**]. Multiple data points can have similar changes reflected across various features. This can cause errors in our unsupervised algorithms. We randomly select a feature set that contain lesser number of features than the original dataset. We use the following algorithm to select random features:

The new dataset with the random features can now be passed to an outlier detection algorithm(LOF algorithm), and obtain an **outlier vector**. We run our

Algorithm 1: Implemented Algorithm for Feature Bagging

Require: $N \geq 1$

Ensure: $D \geq 2$

$D \leftarrow len(\text{Dataset\_Features})$

$N \leftarrow random(\frac{D}{2}, D)$

$i \leftarrow 0$

$feature\_sub \leftarrow []$ **while** $i < N$ **do**

$f = random(0, D)$ **if** $Dataset\_Features[f] \notin feature\_sub$ **then**

$feature\_sub.append(\text{Dataset\_Features}[f])$;

$i := i + 1;$

else

continue

return $feature\_sub$

feature bagging algorithm multiple times to obtain different random features, and a unique outlier vector for each set of random features. The final result is an average of all the outlier vectors. As this approach runs the LOF algorithm multiple times over a unique selection of features, we call it **ensemble LOF**.

Net Local Oultier Factor (LOF) is a combination of single LOF and ensemble LOF. Single LOF performs well in recognising unique outliers, such as large/immediate changes in data, but it does not recognise patterned outliers well. Ensemble LOF can detect patterned outliers, but the outlier scores for anomalous datapoints often occur extremely close to the threshold and can be mislabled as normal scores. We can observe that single LOF and ensemble LOF compliment one another in calculating the outlier scores. We combine the scores by multiplying the outlier vectors from single and ensemble LOF together and finding it's square root.

$$net\_lof = \sqrt{single\_lof * ensemble\_lof} \tag{3.6}$$

A normal data point will have scores below the threshold for single lof and ensemble lof approaches. An anomalous data point shall have scores above the threshold for single lof and ensemble lof approaches. If any of the two approaches result in an incorrect output, we rely on the other approach to reduce the error.

3.4.2.2 Robust Random Cut Forest

Robust Random Cut Forest[**?**] is an ensemble algorithm that uses a collection of independent Robust Random Cut Trees to structure data into tree graphs. Anomalies are detected by comparing the changes the addition of a data point as a leaf/node makes to the structure of the graph. This algorithm is designed used to detect outliers in streaming data. It is designed to detect anomalies in streaming data [**?**]. It works well with high-dimensional data, and reduces the influence of irrelevant dimensions in the dataset. The outlier scores are returned in various ranges. Therefore, to assist in our detection the scores are then scaled using a min-max scaler. This is to provide a form of uniformity. The scaled values for the array out outlier scores O are calculated as:

$$O_{scaled} = \frac{O - O_{min}}{O_{max} - O_{min}} \tag{3.7}$$

To assess whether outliers are indicated by the RRCF scores, we analyze the distribution of the scores of the data points. If no anomalies are present in the data, then the scores will have an even distribution across the data points. With an even distribution of data points, we can assume that the difference between the mean of the data points and the median of the data points will be extremely small. The next check we perform is to compare the resulting RRCF scores against the line of best fit for all the scores calculated. If anomalies are present in the dataset, their scores will be abnormally higher than the scores of regular datapoints. If there are no

anomalies present in the dataset, then the scores shall be spread out evenly, eventually dropping down to a low value near the end of the timeframe. We can use this scoring distribution to plot a line of best fit for all the calculated datapoints. A line of best fit, when plotted against regular data will have a non-positive slope. If anomalies are present, they will cause a large enough deviation for the line to have a positive slope. We calculate the slope and intercept of the line of best fit by using least squares [**?**]. We can thus check if any anomalies are present in the RRCF scores by checking the difference between the mean and median and checking the slope of the line of best fit. If the checks satisfy that anomalies are present, we can catagorize specific data points as anomalous by checking how far the RRCF score is from the value of the line of best fit at that index.

3.4.3 Weighted Least Square (WLS) Method

The goal of the State Estimation of distribution system in the current research work is defined as high accuracy estimation of the system voltage magnitudes, targeting the environments with high noise levels and presence of measurement outliers. The outliers considered in this work are implemented in terms of missing measurements or wrong phasor estimations. For the purpose of estimation of the voltage magnitudes, Weighted Least Square (WLS) technique is adopted.

Assuming a linear regression model, representing the system equations with Y_i and X_i defined as system variables for $i = 1, \ldots, n$:

$$Y_i = \alpha_0 + \alpha_1 X_i + \varepsilon_i \tag{3.8}$$

with $\varepsilon_i \sim N\left(0, \sigma^2/w_i\right)$, the parameters α_0 and α_1 are estimated using minimization of the E_w function, with constant $w_1, \ldots, w_n$, defined as below:

$$E_w\left(\alpha_0, \alpha_1\right) = \sum_{i=1}^{n} w_i \left(y_i - \alpha_0 - \alpha_1 x_i\right)^2 \tag{3.9}$$

In this methodology, the considered weights are assumed to be inversely proportional to the variances, consequently, higher weights are assigned to the point with low variants and lower weights are assigned to higher variant points.

Furthermore, the final estimated values are given as:

$$\begin{aligned} \widehat{\alpha}_0 &= \bar{y}_w - \widehat{\alpha}_1 \bar{x}_w \\ \widehat{\alpha}_1 &= \frac{\sum w_i \left(x_i - \bar{x}_w\right)\left(y_i - \bar{y}_w\right)}{\sum w_i \left(x_i - \bar{x}_w\right)^2} \end{aligned} \tag{3.10}$$

where $\bar{x}_w$ and $\bar{y}_w$ are defined as the mean values with considered assigned weights.

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} \quad \bar{y}_w = \frac{\sum w_i y_i}{\sum w_i} \tag{3.11}$$

The aforementioned methodology for WLS in the current work is extracted in a general matrix formulation with $\boldsymbol{W}$, defined as a diagonal matrix with corresponding weights, $w_1, \ldots, w_n$, as diagonal elements. The exploitation of the Distribution System State Estimation (DSSE) technique is built upon the original WLS methodology

and the comparative results with outliers and with filtered outliers are presented in the results section of this chapter.

In order to increase the accuracy and robustness of the DSSE methodology, an outlier compensation mechanism is embedded as part of the DSSE block for the anomalous data point x_t, following the below steps:

- x_t labeled as anomalous data point
- x_i and x_j are to be labeled as the two previous adjacent normal data points
- $\tilde{x}_t$=mean(x_i, x_i)
- replace x_t with $\tilde{x}_t$

3.5 Ensemble Technique and Event Classification

3.5.1 Ensemble Approach (EA)

Generation of the anomaly detection results by individual base detectors, is always prone to certain assumptions and limitations of the specific technique. In order for the accuracy of the anomaly detection to be increased, Ensemble Approach (EA) is proposed by integration of the detection results from separate detectors.
Detection of anomalies can be interpreted as identification of high noise levels, miss-

ing or bad data, as well as actual physical power events. Specifics of each detection technique, presented in this chapter, is tailored for the processing requirements, being drawn by the used detection strategy as an independent anomaly detection scheme. Initially, a primary scan of the D-PMU extracted data is being conducted by each of the detectors separately, followed by score-conversion of the results. In the next step, the anomaly scores are being normalized, aiming towards decreasing the dependency of detection quality to hyperparameter tuning.

The proposed EA starts by acquisition of the vector of the anomaly scores with regard to the simulation observations, requiring to be outputed by respective base detectors. Primary challenge of this step is the nature of the detection process, used by each of the anomaly scores, resulting in various ways of score generation process. A synchronization step is needed towards normalization of the generated vectors by standalone detectors, keeping the final score values in a coherent range for integration. Several approaches are studied in a nuanced manner and applicability of the implementation on real-field distribution dataset.

In the final step of detection, the scores are being integrated using MOA (Maximum of Average) technique and the outputted results are to be used as detection criteria. As part of the base detector development and performance evaluation process, it has been observed by the research team that anomalous observations are remaining hidden in outlier function-based techniques, while clustering methods (i.e. DB-SCAN,

and K-Means/Elbow) are efficiently performing. In the other scenarios, anomalous data points are occurring in a window of data manner, hence generated clusters are not capable of differentiating between a bad or missing data and actual feeder events. Preliminary studies to highlight the strengths and shortcomings of the detectors have been conducted and MOA is chosen as the most efficient way of outlier score integration, with respect to the developed base-detectors in the current research. As far as the MOA ensemble integration of the scores are concerned, stand-alone base detectors are being divided into two groups of 1) Clustering-based: including DB-SCAN and K-Means, and 2) Outlier Score-based: including Net LOF and RRCF.

After finding the average score for each group of detectors, MOA accounts for the resulted degree of anomalous, being interpreted as a normalized number between 0 and 1, corresponding to measurements at each timestamp.

3.5.2 Broader level event classification

After the bad data are filtered through the base anomaly classification, the detected event data fragments are sent to broader-level event classification, and each event is labeled according to the corresponding physical rule. The proposed broader level classification is shown in Figure 3.2.

The developed algorithm executes feature extraction by DBSCAN, where it

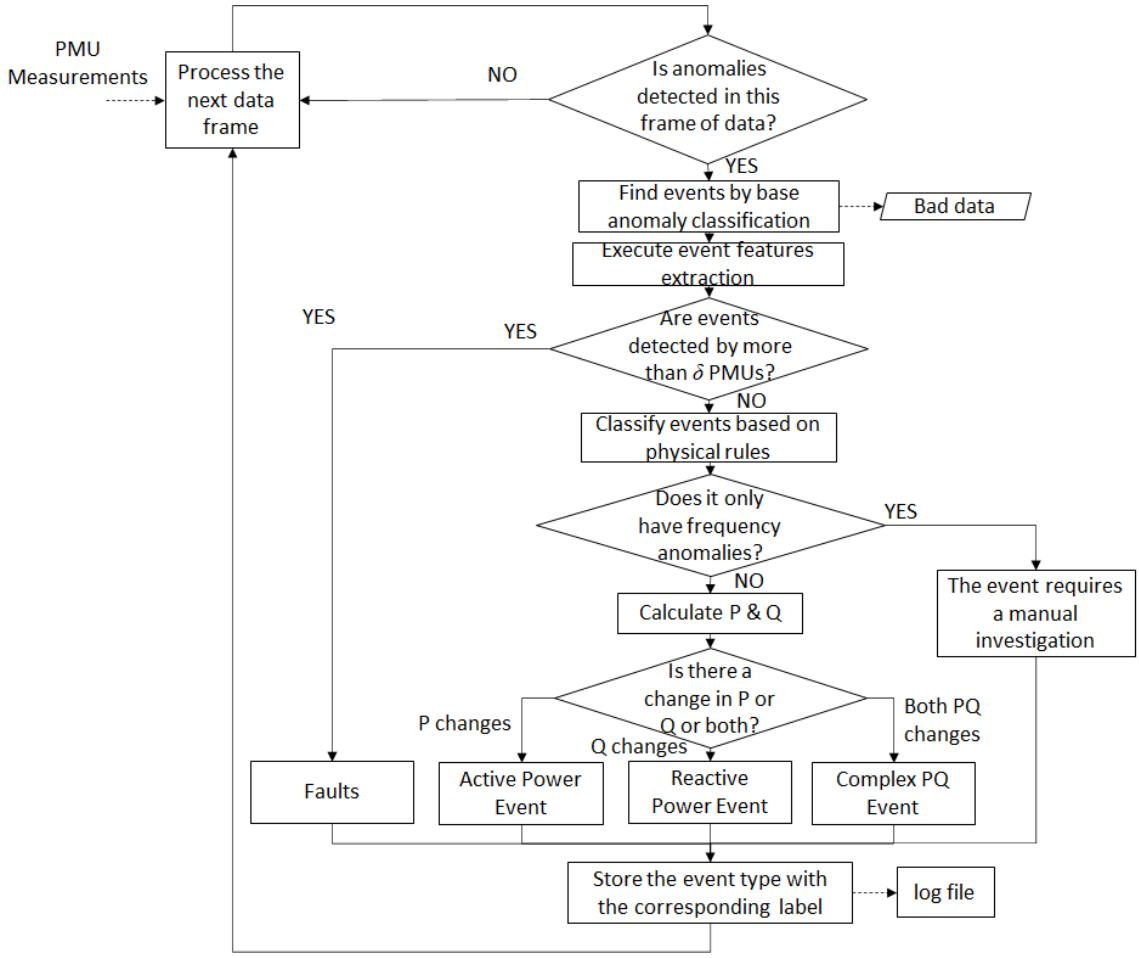


Figure 3.2: Broader-level classification architecture

detects the changes in V, I, P, or Q in those event data fragments. Then the extracted signatures are used for physical rule based event classification, and those events can be classified into following categories:

- Active power event: if a event signal has a significant change in active power change, then it is assigned to P event category. E.g., increasing the power

generation can lead P changes.

- Reactive power event: if a event signal has a significant change in reactive power change, then it is assigned to Q event label. E.g., opening capacitor bank will have a big change in Q.

- Complex power event: if a event signal has a significant change in both active and reactive power change, then it consider as a complex event and will be assigned with a PQ event label.

- Faults: fault classification is completed before the physical-rule based anomaly classification. Faults are classified on the basis of collecting local and spatial information. If an event can be detected by more than a specified number of PMUS, it can be considered a fault.

In the given test D-PMU dataset, three-phase voltage magnitude and angles, three-phase current magnitude and angles, frequency, and rate of change of frequency are included. Based on those attributes, we calculated the active power injection and reactive power injection. The processed data are then fed to DBSCAN for feature extraction. Feature extraction is one prerequisite step for event classification that obtain the main signatures from a piece of signal. The process of feature extraction in this project is achieved by a fine-tuned DBSCAN. Recall that the hyperparameter of DBSCAN are optimal epsilon distance ϵ and the minimum number of samples

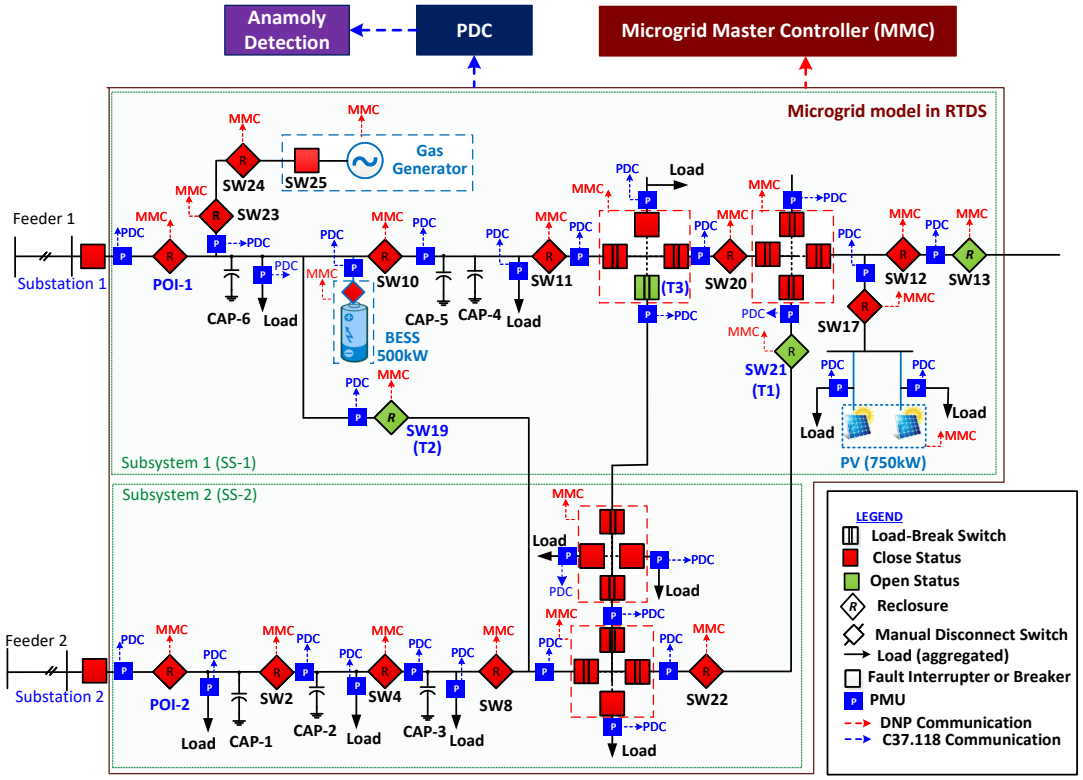


Figure 3.3: One-line diagram of the Bronzeville Community Microgrid

$MinPtr$.

3.6 Test systems and performance metric for evaluation

3.6.0.1 Bronzeville Community Microgrid

Commonwealth Edison (ComEd), a utility serving more than 4 million customers in northern Illinois, is deploying Bronzeville Community Microgrid (BCM) on

south side of Chicago. The BCM serves critical infrastructures, such as Chicago Police Department headquarters along with approximately 1,000 residences, businesses, and institutions. The BCM integrates solar photovoltaics (PV), battery energy storage system (BESS), a controllable generator, and a microgrid master controller to connect/disconnect from ComEd's grid and keep the local power flowing in the event of interruption from the main grid. In this research, the D-PMU data from real-time simulation of BCM at ComEd's Grid Integration Technology (GrIT) lab facility are utilized to test and validate the proposed method.

Figure 3.3 shows the schematic diagram of the BCM. It consists of two feeders fed from the two substations via point of interconnection (POI) switches, POI-1 and POI-2. The top feeder, SS-1, consists of loads, distribution automation (DA) devices, a 4.8 MW gas generator, 500 kW/2 MWh battery storage system, and 750 kW solar photovoltaic (PV) system. The lower feeder, SS-2, consists of loads and DA devices. These two feeders can be interconnected via tie-switches (T1, T2 or T3) to transfer load of one feeder to the other. Since all the distributed energy resources (DERs) are located in the top feeder, SS-1 can island from rest of the system to form partial island. In full island mode, the two sub-systems are interconnected and both POI switches are opened to isolate the entire microgrid from utility supply.

As a part of initiative, ComEd has utilized synchrophasor technology to improve its business operation. In BCM, 46 D-PMUs are strategically placed to stream time-

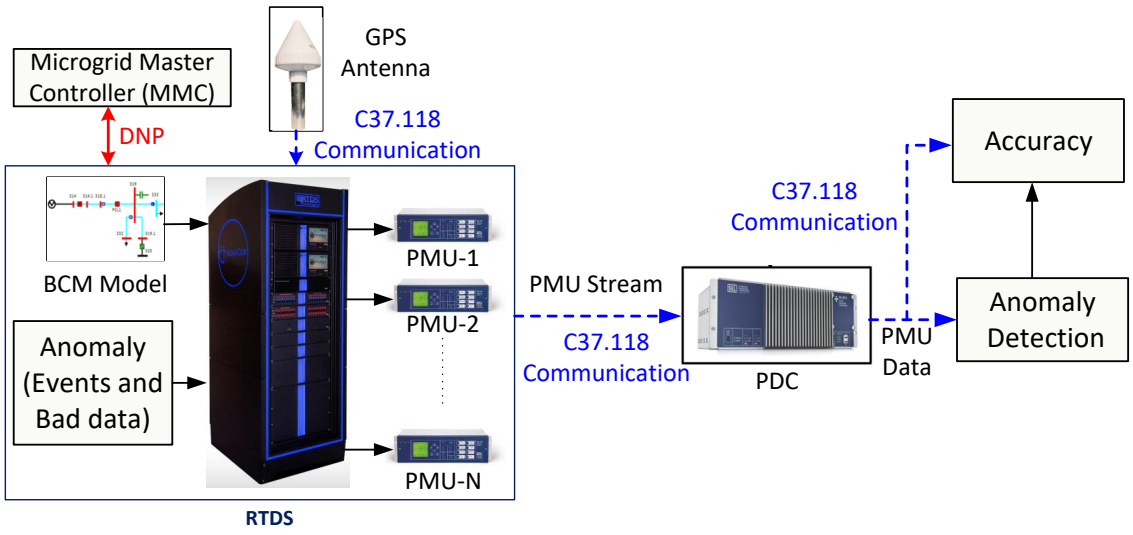


Figure 3.4: RTDS test bed architecture

stamped measurements from various nodes in the microgrid to the utility analytics platform and storage. Broadly, these data have two types of main functions and applications: real-time and offline applications. The real time D-PMU data deal with time-sensitive functions, such as system event detection, protection and control, and system monitoring, which typically require event-occurrence-to-application-outcome latency of several milliseconds to a few seconds. Some of the offline applications include post event analysis, load modeling, and load forecasting.

3.6.0.2 RTDS Test Bed Architecture

Figure 3.4 shows the test-bed architecture at ComEd's lab facility. The power circuit of the BCM and the D-PMUs are modeled and and implemented in RTDS.

The GTNET card simulate the D-PMUs in RTDS model that stream real-time time-stamped data to a phasor data concentrator (PDC) using C37.118.2 communication protocol. The aggregated data in the PDC is then forwarded to anomaly detection and event classification algorithm discussed in this chapter.

3.6.0.3 Test Cases

Several events are applied on the BCM RTDS model and the D-PMU data are collected to test and validate the proposed algorithm. Followings are the details of each test cases.

Test Case I: In this test, 23 events are applied at a regular interval of 45 seconds. The events include connection and disconnection of the DERs, DER power changes, load changes, switching of capacitor banks, and faults (LG, LL and LLLG) at several nodes.

Test Case II: This test includes 23 events such as connection and disconnection of the gas generator, load changes, switching of capacitor banks, faults (LL, LLG, LLL, LLG and LLLG type), and switching of the POI and tie switches. The events include changes in the circuit configuration, such as partial islanding, load transfer, full islanding, blackout, and grid reconnection of the microgrid. The test events occur at a regular interval of 45 seconds.

Test Case III: The events in Test Case III include faults (LG, LL, LLG, LLL, LLLG type), capacitor switching, load changes, connection/disconnection of DERs,

DER power changes, and operation of POI and tie switches for load transfer, partial islanding, partial blackout, and grid reconnection. There are 23 events at a regular interval of 45 seconds.

3.7 Results and Summary

3.7.1 Anomaly Detection

Based on the developed testbed and the corresponding usecases I to III, performance evaluation of the developed detection and classification techniques have been conducted. Two evaluation metrices reflecting "Accuracy" and "Precision" are evaluated using the four terms— a) True Positives (TP) are positive instances correctly classified, b) False Positives (FP) are negative instances classified as positive, c) True Negatives (TN) are negative instances correctly classified as non-positive, d) False Negatives (FN) are positive instances classified as negative. A "recall" is the ratio of TP to the sum of TP and FN and "Accuracy" is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.12}$$

Table.3.1 presents the confusion matrix based on the anomaly detection performance of all the stand-alone base detectors as well as the developed "Ensemble Approach".

Table 3.1: Anomaly Detection Performance

		Accuracy	Precision
Testcase 1	DB-SCAN	0.7826	0.7826
	K-Means	0.7635	0.7635
	Net LOF	0.7462	0.5518
	RRCF	0.7192	0.8111
	Ensemble	0.9153	0.8942
Testcase 2	DB-SCAN	0.8261	0.8261
	K-Means	0.7927	0.7927
	Net LOF	0.7442	0.4710
	RRCF	0.6950	0.3980
	Ensemble	0.9484	0.9397
Testcase 3	DB-SCAN	0.7826	0.7826
	K-Means	0.7462	0.7462
	Net LOF	0.7465	0.5775
	RRCF	0.7002	0.4687
	Ensemble	0.9248	0.8925

3.7.2 Distribution System State Estimation Results with Measurement Outliers

In this section, the technical methodology and the developments that are presented in the previous sections of this chapter are deployed as a single DSSE tool that is robust to outliers and noise. An outlier compensation methodology is also considered and the artificial noise and outliers are introduced to the D-PMU measurements that are extracted from the Bronzeville Community Microgrid. Among the two main feeders in the test system, substation one is considered as the main feeder and is used for DSSE evaluation purposes.

Multiple outliers are introduced in a dynamic period of the system operation for a single example bus and in the next step, with the presence of outliers, the performance of WLS is evaluated in Fig.3.5. Furthermore, in this figure, the DSSE results are provided for the case after the detection, filtering, and compensation of the outliers, labeled as Outlier Filtered WLS (OF-WLS).

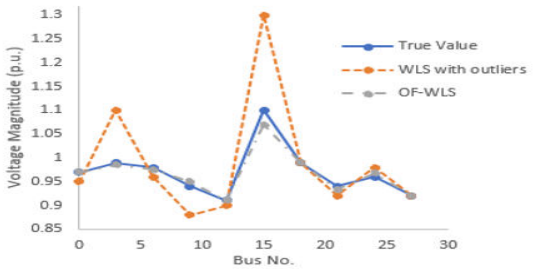


Figure 3.5: Performance comparison of the conventional WLS and OF-WLS

3.7.3 Broader-level Classification Results

This procedure is to verify the usability of the proposed event classification algorithm. The designed event classifier is validated through five different event datasets provided by ComEd/ NuGrid. The DBSCAN-based feature extraction algorithm can determine how many PMUs can detect the current processing event. According to the estimated δ number of PMUs based on a set of ground truths, the classifier distinguished whether the event was a fault. In the test power system, δ was set to be five by the hyperparameter estimation, i.e., if an anomaly is detected by more than 5 D-PMU in the given test system, then this event will be identified as a fault. The average classification results for the simulated test cases are 82.2%. The developed broader-level classification algorithm was validated using the given 5 test cases created by the RTDS test bed introduced in Section 3.6.0.2. The results using accuracy metrics with a total of 104 events are illustrated in Table 3.2.

For complex PQ events, including DER connection/reconnecting, DER power changes, topology changes, and grid reconnection, the average classification accuracy of the developed physics-driven broader-level event classifier provided 10% accuracy. For P and Q events, the classifier stands the accuracy at 69.23% and 76.92%, respectively. In addition, the classifier performed well in identifying the fault events, in which the classification accuracy for the given 41 fault events was up to 100%. The

Table 3.2: Classification results for each specific power distribution event.

Estimation Model	# of Event	Accuracy (%)
DER connection/reconnection	12	91.67
DER power changes	11	81.82
Switching capacitor banks	13	69.23
Load changes	13	76.92
Faults (LL, LLL, LG, LLG, LLLG)	41	100
Full/ Partial islanding	6	83.33
Partial blackout	2	100
Grid reconnection	6	83.33

results are evidence that exclusive events like active and reactive still produce the variations in counterparts, given that the distribution system is neither pure reactive nor resistive. Moreover, a few events are not appropriately classified or unclassified entirely because of errors in determining the dominant PMUs and invisible variations.

3.8 summary

In this research, an attempt has been made to examine the practicality of statistical and clustering based anomaly detectors against the real event datasets from D-PMUs connected with simulated Bronzeville Community Microgrid (BCM) network. A detailed step-wise algorithm is discussed and validated for various test case scenarios to investigate true performance of considered methods. A performance comparison is also carried out by utilizing the common validation metric to showcase relative superiority of methods under different test cases. An extension of this work will emphasize the ensemble approach to improve detection accuracy by taking advantage of strength in different methods. Furthermore, anomaly classification, localization and root cause analysis is to be performed in future.

CHAPTER 4. CYBER-PHYSICAL AWARENESS AND EVENT CLASSIFICATION

4.1 Introduction

This chapter proposes a Real-time Anomaly Location and ClassificatiON (RALCON) for power distribution systems to determine the type of anomaly (i.e., short-circuit fault, cyber attack, DER switching) and its location simultaneously. RALCON collects and aggregates measurement data from multiple sources that operate at different sampling rates, such as protection relays and D-PMUs. The output of the data aggregation is then fed into a multi-task long-based short-term memory (LSTM) to classify the type of the anomaly and the location in two separate tasks. The proposed approach can be utilized in real-time operation in order to distinguish between normal operation and other anomalies. The RALCON is validated in a modified IEEE 33-bus test feeder benchmark, that integrates solar generation and energy storage. The results show that RALCON can accurately locate and classify anomalies for several operation conditions. Further experiments highlight the impact of aggregating multiple sources of data in the accuracy of the results.

4.1.1 Background and Problem Description

Given the enhanced automation and ongoing penetration of inverter-based resources to distribution system, further integration of advanced measurement devices such as Distribution Phasor Measurement Units (D-PMUs) and protection relays are required to increase the visibility of the distribution system status, properly capture the real-time anomalies of the system [1], and enable distribution automation. Situational awareness utilizing data from sensors capturing physical system data and cyber data provides an opportunity for real-time cyber-physical monitoring and subsequent decision support. Time-synchronized high-resolution measurements can be exploited for model validation, topology detection, state estimation, event analysis and other applications [2, 3]. The expanding deployment of communication networks and digital devices, however, has made power distribution systems vulnerable to cyber attacks [4, 5] such as the coordinated cyber attack against the Ukrainian power grid [6], and the cyber attack on the communication network at a large renewable energy operator in Utah [7]. Hence, it is of crucial importance to take into account the cyber attack as an individual type of anomaly in power distribution system alongside with the others such as a natural fault. The distribution system measurements (i.e., D-PMUs and protection relays) can be used for cyber-physical anomaly detection. However, the variety of monitoring devices operating at different sampling rates introduce new

challenges in the design of methodologies for the detection of anomalies.

4.1.2 Literature Review

Multiple studies have been conducted in distribution systems focusing on three different tasks of anomaly detection, classification, and location [8, 9]. The majority of the previously-established approaches concentrate on physics-based or data-driven techniques [10]; however, given the complex evolving nature of the distribution system induced by high penetration of Distributed Energy Resources (DERs), e.g., energy storage (ES) and solar generation (SG) units, and integration of unconventional controllable loads [11], the requirement for a practical tool, achieving high accuracy for detection, classification, and location of anomalies in a coherent manner, still remains open to research [12].

The precision and solutions offered by analysis of streaming D-PMU data facilitate operators to reveal short-time events that would otherwise be undetected [13]. The issue of anomaly analysis in the distribution system is addressed in [14] and an algorithm is proposed combining binary KNN, multiclass SVM, and decision trees. The drawback of this approach is that the algorithm fails to account for the spatial location of buses in the system. This might result in the detection of events at the wrong locations. A compensation theorem-based approach is presented in [15] for

event detection and location. This approach considers the spatial location of buses but ignores the temporal nature of the measurements obtained from the system. A novel optimization algorithm is proposed in [16] to make a bridge between supervised, semi-supervised, and unsupervised learning approaches. The proposed event detection approach in this chapter is validated on a real feeder with actual installed D-PMUs. The Principle Component Analysis (PCA) technique, decision tree, and combination of PCA and SVM with autoencoder are proposed in [17, 18, 19, 20] for event classification, respectively. An approach is presented in [21] based on fast localized spectral filtering using the spatial location of buses. These methods consider the spatial or temporal nature of the measurements but do not jointly exploit both to achieve accurate event detection, location, and classification.

Considering a practical distribution system operating scenario, due to a low number of D-PMU in-field installations, the total number of high-resolution incoming measurements do not provide full observability of the system [22]. Hence, further multi-source measurement integration is required before proceeding to detailed anomaly investigations. Multiple data aggregation strategies have been established in literature, centering around applications of "Bayesian Filtering" with focus on "Kalman Filtering" [23][24] [25]. For the purpose of the multi-source data aggregation in this work, a non-linear assumption of the underlying system model is made, aiming towards enhancement of the data aggregation accuracy with comparison to

the similar works with linear model assumption [26], and a data-driven approach is used for aggregation of incoming information from D-PMUs as well as Supervisory Control and Data Acquisition (SCADA).

The technical literature mostly tackles fault and cyber attack detection problems individually only using D-PMU data , and the few efforts that integrate fault and attack classification do not accomplish the location task [27, 28, 29]. Furthermore, little attention has been given to combining asynchronous sources of data such as protection relays and D-PMUs, which are typically transmitted at different sampling rates and require additional pre-processing efforts in order to utilize for anomaly detection, location and classification. In addition, the majority of cyber-physical anomaly detection methods in the literature focus on the problem in power generation and transmission systems [30, 31, 32], and less attention has been given to developing a comprehensive model integrating fault and attack, as well as DER event classification and location in distribution systems. Finally, the proposed cyber-physical anomaly detection and location approaches in distribution systems are post-failure and offline approaches [10].

4.1.3 Contributions

This research proposes the Real-time Anomaly Location and ClassificatiON (RALCON) system to classify the type and location of different anomalies in distribution system. More specifically, the proposed RALCON system continuously receives physical data that is being collected in the control centner, including the protection relays and D-PMUs, to discern the presence of an anomaly in distribution systems, and determine its type and location. The RALCON system is capable of identifying and locating short-circuit faults, cyber attack on protection relays and DER switching in distribution systems. At its core, the RALCON system includes two significant modules, namely, data aggregation and multi-task learning-based long short-term memory (MTL-LSTM). The data aggregation module is utilized to integrate the asynchronous measurement from various measurement resources with different sampling rates (i.e., protection relays and D-PMUs) and generate uniform high-resolution time-series data. The aggregated measurements with the enhanced resolution are provided to the second module. The MTL-LSTM enables utilizing a single machine learning structure for two individual tasks, the type and the location of the anomaly in distribution system. Employing an LSTM recurrent neural network in the core of the proposed RALCON also enables the advantage of real-time anomaly classification and location, leading to a fast identification in distribution

system. Utilizing LSTM over other artificial neural networks provides the advantage of modeling a sequence of data (i.e. time series of measurements) while each data of a time window can take the dynamic behaviour of the power distribution and low range of anomaly disturbance into account. Additionally, specifying a layer for each time step in LSTM reduces the number of the trainable weights helping to accelerate the training procedure.

The rest of the chapter is organized as follows: Section 4.2 presents an overview of the proposed RALCON system and defines the types of anomalies classified by the model. The multi-source measurement aggregation module in the RALCON system is presented in Section 4.3. The MTL-LSTM module in the RALCON system is presented in Section 4.4. The proposed model is implemented on the IEEE 33-bus test feeder in Section 4.5, which is utilized to locate and classify multiple types of anomalies. In addition, Section 4.5 presents results of comparing the performance of the proposed model for different scenarios of input data. Finally, Section Summary concludes the chapter.

4.2 Overview of the Proposed Model

The architecture of the proposed RALCON is illustrated in Fig. 4.1. The proposed RALCON is implemented in the utility control center to monitor the measure-

ments data and identify the operation status in real-time. The RALCON receives the current data from the field measurements, including protection relays and D-PMUs installed throughout the entire distribution network. In addition, it receives limited samples of historical data collected by the outage management system (OMS) from relays and D-PMUs. Along with the transmission system standards, the communication between D-PMUs is assumed to be via IEEE c37.118 protocol. On the other hand, for a sensible enhancement in the system observability, the installed protection relays are assumed to be connected to the control room using DNP3 communication protocol. A variety of reporting rates have been considered, and 10 samples per second are chosen to comply with the existing distribution feeder communication structure. The aforementioned assumptions on the incoming data aim towards rational development of a data aggregation scheme, focusing on integration of the incoming data from high-resolution D-PMU (120 samples/sec) and low-resolution relay (10 samples/sec) measurements. Using the present and limited samples of historical data from D-PMUs and protection relays, the data aggregation module up-samples the protection relay data to 120 samples/sec. The aggregated data is then fed into the MTL-LSTM to determine the operation mode and the location of the anomaly (if there is any). Operation mode is defined as the operation state of the distribution system including, normal operation, short circuit fault, cyber attack, DER switching, and intermediate mode. The inclusion of the normal operation mode enables the real-

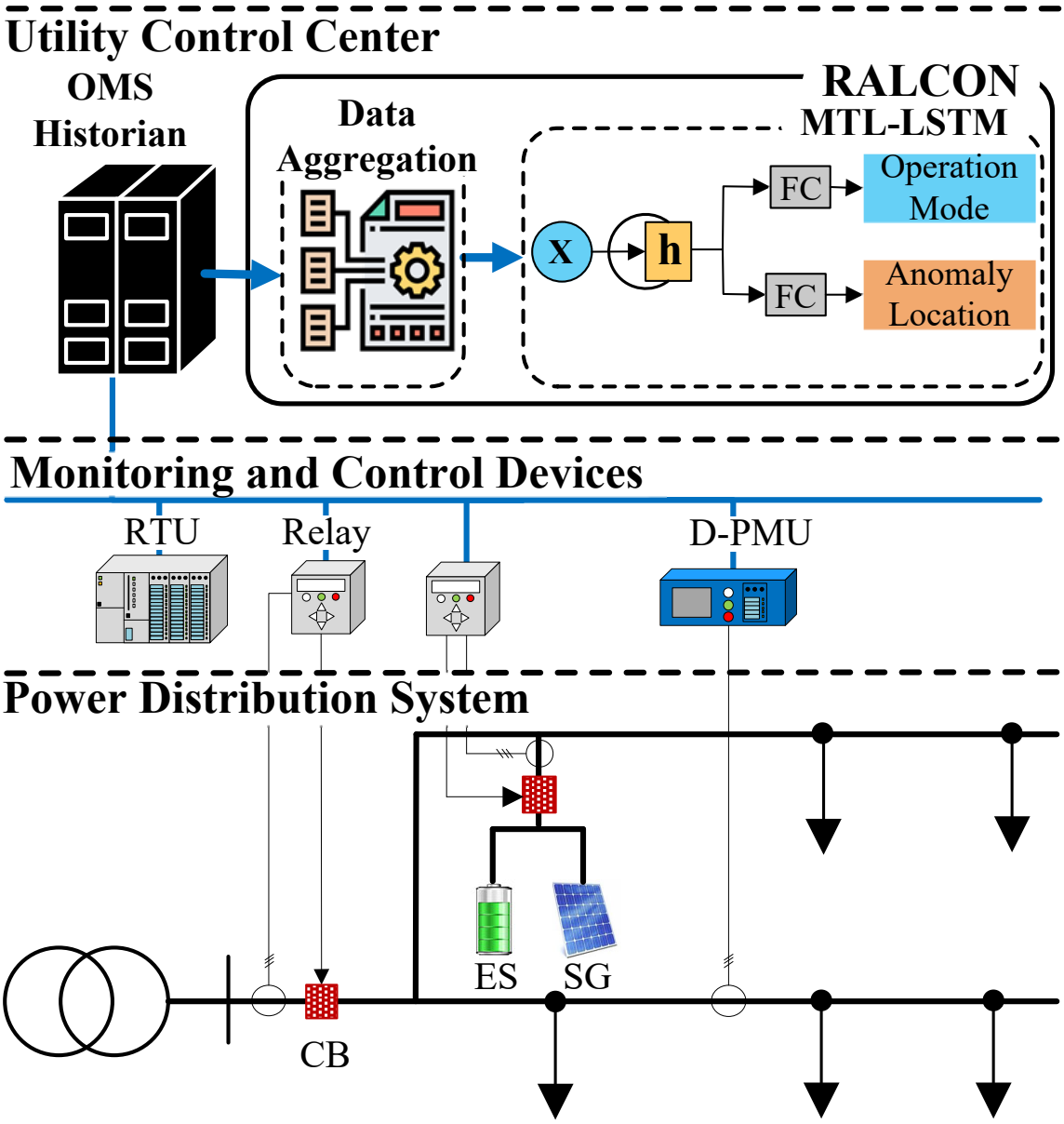


Figure 4.1: Architecture of the proposed real-time anomaly location and classification (RALCON) system

time application of the proposed model. The location of the anomaly is also classified simultaneously to illustrate the predicted location of the anomaly to the operator. The training process of the proposed RALCON is performed offline. Nevertheless, after training the model, it can be used in the real-time operation and the output of the classification tasks can be visualized on the monitoring station to enhance situational awareness and accelerate the anomaly identification process. The data aggregation and MTL-LSTM modules are discussed in sections III, IV, respectively.

4.2.1 Input data

Let us consider a distribution system with a set of lines $\mathcal{L} = \{1, ..., L\}$, a set of measurement sensors $\mathcal{S} = \{1, ..., S\}$, and a set of DERs, $\mathcal{D} = \{1, ..., D\}$, where L is the number of lines, S is the number of sensors including protection relays and D-PMUs, and D is the number of DERs in the distribution system. The set of sensors $\mathcal{S}$ consists of a subset of protection relays $\mathcal{R} = \{1, ..., R\}$, and a subset of D-PMUs $\mathcal{U} = \{1, ..., U\}$, where R is the number of protection relays and U is the number of D-PMUs.

The collected data consists of the measurement data from relays and D-PMUs, which are continuously being received and stored in the OMS database. The RALCON receives the data from the OMS historical system for a specific time window.

The data employed by RALCON includes incoming measurements from both high-resolution and low-resolution measurement devices, elaborated as follows:

4.2.1.1 D-PMU High-resolution Measurements

Three-phase voltage and current magnitudes and angles, frequency and rate of change of the frequency (ROCOF) are transmitted to the control center with the rate of 120 samples/sec.

4.2.1.2 Protection Relay Low-resolution Measurements

Three-phase voltage and current magnitudes and frequencies are transmitted to the control center with the rate of 10 samples/sec.

Let us define $\mathbf{Q}(t) = [\mathbf{x}_1(t), ..., \mathbf{x}_s(t)...., \mathbf{x}_S(t)] \in \mathbb{R}^{14\times S}$ as the measurement matrix for a distribution system, where t denotes the time index; and $\mathbf{x}_s(t)$ denotes the sensor measurement vector defined as the following:

$$\begin{aligned}\mathbf{x}_s(t) = [&I_{s,a}(t), I_{s,b}(t), I_{s,c}(t), \alpha^I_{s,a}(t), \alpha^I_{s,b}(t), \alpha^I_{s,c}(t),\\ &V_{s,a}(t), V_{s,b}(t), V_{s,c}(t), \alpha^V_{s,a}(t), \alpha^V_{s,b}(t), \alpha^V_{s,c}(t),\\ &f_s(t), rf_s(t)]^T \end{aligned} \tag{4.1}$$

where $I_{s,a}(t)$, $I_{s,b}(t)$, $I_{s,c}(t)$, $V_{s,a}(t)$, $V_{s,b}(t)$, $V_{s,c}(t)$ denote the current and voltage magnitudes; $\alpha^I_{s,b}(t)$, $\alpha^I_{s,c}(t)$, $\alpha^V_{s,a}(t)$, $\alpha^V_{s,b}(t)$, $\alpha^V_{s,c}(t)$ denote the current and voltage angles for sensor $s \in \mathcal{S}$ at time t in three phases a, b, c, respectively; $f_s(t), rf_s(t)$ denote the frequency and ROCOF for sensor s at time t, respectively. All the arrays

of vector in $x_s(t)(\forall s \in \mathcal{R})$ considered to be 0 if the relay does not provide measurement data at time sample t. Correspondingly, any features that are not being measured by relays, i.e. angle values and ROCOF, considered to be 0.

4.2.2 Anomaly Classes in RALCON

The RALCON system identifies and locates multiple types of anomalies and distinguishes between normal operation and an event in the distribution system. Two different tasks are taken into account in RALCON; 1) Detection of the distribution system operation mode 2) Identification of the location of an event. Five different operation modes are taken into account in this research as the followings:

4.2.2.1 Normal Operation

The RALCON system includes the normal operation as one of the operation modes in order to enable the real-time application of anomaly detection and location. Since the distribution system is usually being operated without any anomaly, a label is needed for the normal operation. Additionally, the location of this operation mode is set to 0, highlighting a specific class for this operation mode in the anomaly location task.

4.2.2.2 Short-Circuit Fault

The RALCON system takes the short-circuit fault as the second operation mode. Also, the location of short-circuit faults for each line is taken into account as the labels of the location of the anomaly. Once a fault happens in line $l \in \mathcal{L}$, the protection relays and D-PMUs sense a short-circuit fault current, and the closest relay sends a trip command to the associated breaker or recloser after a specific time delay. The number of fault location classes equals the number of lines in distribution system.

4.2.2.3 Cyber Attack Types

The proposed RALCON system takes cyber attacks into account as an anomaly type where the adversary can bypass the network security and launch an attack directly on protection relays or in the communication layer [10]. The false-data injection attack (FDIA) is assumed to happen on the protection relays trip command. This kind of attack would directly cause an interruption in the distribution system. Every protection relay in the distribution system, especially the ones in critical locations on the system, can be the target of a cyber attack. Therefore, the number of cyber attack location classes equals the number of protection relays in the network. FDIA on D-PMUs are not considered here, as they would not trigger any trip command.

4.2.2.4 DER Switching

The proposed RALCON system considers DER switching as an event type where a DER gets connected/disconnected to the distribution system. Knowing about the exact time of the DER switching can provide noteworthy information for the operator. Every bus connected to a DER in distribution system is considered as a location for a DER switching event. Therefore, the number of DER switching location classes equals the number of buses with DER.

4.2.2.5 Intermediate Operation Mode

In addition to the above-mentioned operation modes, another mode is defined as an *intermediate* mode in which a part of the MTL-LSTM samples belong to an operation mode and the rest belong to another one. For instance, when a short-circuit fault happens in the middle of a time window of MTL-LSTM, a part of the samples belong to a normal operation and the others belong to a short-circuit fault. In order to improve the performance of the proposed method and take the transition time window into account, the intermediate mode is utilized.

The total number of classes for operation mode task in RALCON includes five different labels as explained, i.e., $K^o = 5$. Additionally, the total number of classes for the anomaly location in RALCON equals to: $K^l = \max\{L, R, D\} + 1$, where we consider a label for no location in addition to the maximum number of location classes

for different operation modes.

4.3 Multi-source Measurement Aggregation

The proposed RALCON framework begins with acquiring the available data in the distribution system, i.e., relay measurements from SCADA and D-PMUs. The aforementioned data are primarily being captured in streaming and non-streaming manners with various reporting rates.

4.3.1 Inputs and Outputs of Measurement Aggregation Module

The formulation and the steps that are presented in this section, tackle the data aggregation problem with an state estimation approach by creating the observation model during the dynamic events. Prediction of the system state using the available observations, accompanied by the assumption of the system state at time $t+1$ being only dependent to the system state at time t, makes the estimation of the unknown variables in relay measurements feasible. As a result, resolution improvement of the measurements with incorporating phase angle estimations is achieved.

Fig. 4.2 presents an overview of the developed data aggregation scheme with the corresponding required inputs and final generated outputs. The overall sensor measurement aggregation technique, implemented in the current research work for

the purpose of distribution system anomaly detection, classification and location, is considered as the preliminary step to preprocess the data for further anomaly investigations.

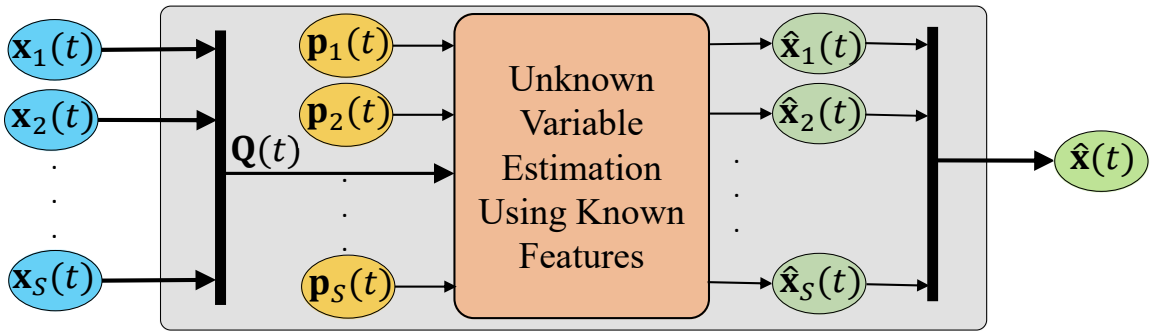


Figure 4.2: Overview of the Data Aggregation scheme at time snapshot of t

4.3.2 Application of the Proposed "Data Aggregation Scheme" to Distribution System Measurements

Streaming data from D-PMUs are in high-resolution with measurements of voltage and current phasor values. However, the incoming relay data are originally obtained using potential and current transformers, containing low-resolution information on voltage and current magnitudes with missing phase angles. The goal of this section is to develop an estimation strategy for the missing phase angle values, captured by relays. Moreover, aggregation of the Root Mean Square (RMS) values of the measured voltage and current magnitudes by relays in a correlative manner with

similar D-PMU measurements to synchronize the data-capturing resolution of all the resources (i.e. 120 samples/sec). The assumption of the total number of incoming high resolution measurements (captured by D-PMUs) is greater than the total number of the estimation variables needs to be maintained to assure high aggregation accuracy.

The data aggregation strategy in this work is adopted through a combination of multiple steps of estimation and re-estimation of unknown variables, by introducing the concept of "Proximity Vector", denoted by $\mathbf{p}_s \in \mathbb{R}^{S\times 1}$, for sensor s. The proximity vector is constructed by the adjacency comparison between sensors based on the electrical distances, aiming towards identification of the closest neighboring D-PMU with respect to each relay device. Given a radial feeder configuration, calculation of the electrical distances is interpreted as the summation of the line impedances between a relay and all the installed D-PMUs. Consequently, the D-PMU with the least upstream or downstream impedance is chosen as the adjacent D-PMU. Accordingly, the proximity vector for sensor s is defined as: $\mathbf{p}_s = [0, \ldots, 1, \ldots, 0]^T_{1*S}$ with all the elements except the element j equal to zero, implicating the adjacency of the sensor s to D-PMU index j.

In the developed strategy, incoming measurements at any given timestamp are considered as a vector of captured features by the corresponding sensor. For the

purpose of the mathematical formulation of the aggregation approach, vector $\mathbf{x}_s(t)$ is considered as the feature vector for sensor s at t.

In conjunction with other parameter estimation approaches, two phases of prediction and updating of the prediction is considered to be accomplished respectively. Estimation variables are considered to be the relay unknown features, constructing a vector of voltage and current phase angles as well as the missing RMS values for voltage and current magnitude measurements and frequency. Assuming that the measurement aggregation is performed for $t-1$, the model of the implemented data aggregation scheme in this work consists of the following steps for time step t:

1. **Step One:** Initialization

   (a) Collection of all measurements from D-PMUs and relays in $\mathbf{x}_s(t-1)$ for $s = 1, \ldots, S$;

   (b) Construction of $\mathbf{Q}(t-1)$ using $\mathbf{x}_s(t-1)$.

2. **Step Two:** Estimation

   Calculation of the estimations using the obtained measurements at time t:

$$\widehat{\mathbf{x}}_s(t) = \mathbf{Q}(t-1)\mathbf{p}_s(t-1). \tag{4.2}$$

3. **Step Three:** Update

   This step is to conduct a comparison between the RMS values of the last available actual relay measurements for the relay s, with the estimated values from

D-PMUs as followings:

$$\epsilon_s = \left| \frac{\widehat{\mathbf{x}}_s(t) - \widehat{\mathbf{x}}_s(t-1)}{\widehat{\mathbf{x}}_s(t-1)} \right|, \tag{4.3}$$

where ϵ_s denotes the estimation error for sensor s. The percentage error between the values is calculated and compared with a preset threshold, denoted by ε_{th}. The threshold parameter, needs to be properly tuned. This will serve as the criteria for increasing the relay measurement resolution either by D-PMU based estimations or repetition of previous relay estimation.

- $\varepsilon_{th} \geq \epsilon_s$: No Update is required, i.e. $\widehat{\mathbf{x}}_s(t) = \mathbf{Q}(t-1)\mathbf{p}_s(t-1)$
- $\varepsilon_{th} < \epsilon_s$: Update using the last estimation, i.e. $\widehat{\mathbf{x}}_s(t) = \widehat{\mathbf{x}}_s(t-1)$.

After updating the unknown measurements in all sensors, the estimated measurement vector at time t is constructed as: $\widehat{\mathbf{x}}(t) = [\widehat{\mathbf{x}}_1^T(t), ..., \widehat{\mathbf{x}}_s^T(t), ..., \widehat{\mathbf{x}}_S^T(t)]$ and is fed into the MTL-LSTM. The overall algorithmic implementation of the data aggregation scheme, as part of the proposed RALCON is presented in Algorithm.2.

4.4 The Structure of the Multi-task Learning-based LSTM

The proposed RALCON utilizes the concept of RNN and Multi-Task Learning (MTL) simultaneously in the core of the MTL-LSTM module. The MTL-LSTM con-

Algorithm 2: Data Aggregation Scheme

Input the obtained measurements by D-PMUs in $\mathcal{U}$ and Relays in $\mathcal{R}$, at $t-1$;

Build the measurement matrix $\mathbf{Q}(t-1)$ using the uploaded measurements;

for ($s = 1$ to S; $s++$):

  if $s \in \mathcal{U}$:

    $\widehat{\mathbf{x}}_s(t) = \mathbf{x}_s(t)$

  if $s \in \mathcal{R}$:

    Estimate the unknown measurements using the "Proximity Vector ($\mathbf{p}_s$)":

    $\widehat{\mathbf{x}}_s(t) = \mathbf{Q}(t-1)\mathbf{p}_s(t-1)$

    $\epsilon_s = \left|\left[\widehat{x}_s(t) - \widehat{x}_s(t-1)\right] / \widehat{x}_s(t-1)\right|$

    if $\epsilon_s > \varepsilon_{th}$:

      $\widehat{\mathbf{x}}_s(t) = \widehat{\mathbf{x}}_s(t-1)$

$\widehat{\mathbf{x}}(t) = [\widehat{\mathbf{x}}_1^T(t), ..., \widehat{\mathbf{x}}_s^T(t), ..., \widehat{\mathbf{x}}_S^T(t)]$

Return $\widehat{\mathbf{x}}(t)$

sists of a LSTM layer representing the recurrent neural network structure. Furthermore, two parallel Fully Connected (FC) layers and two *softmax* layers are followed by the LSTM to implement the operation mode and anomaly location tasks as shown in Fig. 4.3.

In particular, MTL is a machine learning approach in which several learning tasks are solved simultaneously by optimizing multiple loss functions at once. The standard LSTM is modified in this research to meet the requirements for anomaly classification and localization in distribution system. Rather than training an independent model for each task, a single model is implemented in MTL to optimize the total loss function. MTL has gained extensive attention across different research areas such as natural language processing, compute vision, recommendation systems among the others [33]. MTL can be implemented with sharing partially similar neural network. In the proposed MTL-LSTM, two tasks (e.g., operation mode and anomaly location) share the LSTM layer as the main hidden layer enabling the time-series input features. The output of the LSTM is then fed into two parallel FC layers as the individual layers for each task. The layers of the proposed machine learning model as well as the functions are elaborated in the following subsections.

4.4.1 Long-Short Term Memory Layer

A recurrent neural network is a neural network that is specialized for processing a sequence of input data $\widehat{\mathbf{x}}(1), ..., \widehat{\mathbf{x}}(t), ..., \widehat{\mathbf{x}}(\tau)$ with the time step index t ranging from 1 to τ where τ indicates the time steps of the RNN. In contrast to traditional feed-forward neural networks, RNNs have self-connected recurrent connections modeling the temporal evolution. The output response $\mathbf{h}(t)$ of a recurrent hidden layer can be formulated as follows:

$$\mathbf{h}(t) = \sigma_h \left(\mathbf{W}_{\widehat{x}h} \widehat{\mathbf{x}}(t) + \mathbf{W}_{hh} \mathbf{h}(t-1) + \mathbf{b}_h \right), \tag{4.4}$$

where $\widehat{\mathbf{x}}(t)$ is the output of the data aggregation module as well as the input of the classifier at time t; $\mathbf{W}_{\widehat{x}h}$ and $\mathbf{W}_{hh}$ are mapping matrices from the current inputs $\widehat{\mathbf{x}}(t)$ to the hidden layer h and the hidden layer to itself; $\mathbf{b}_h$ denotes the bias vector; σ_h denotes the activation function in the hidden layer h. The structure of the RNN presented above has difficulty in learning long range dependencies due to vanishing gradient effect hampering the learning of long data sequences. To overcome this limitation, recurrent neural networks using LSTM have been designed to mitigate the vanishing gradient problem and to learn the long-range contextual information of a

temporal sequence. The LSTM transition equations are as the followings [34]:

$$\mathbf{i}(t) = \sigma\left(\mathbf{W}_i\widehat{\mathbf{x}}(t) + \mathbf{U}_i\mathbf{h}(t-1) + \mathbf{V}_i\mathbf{c}(t-1)\right), \tag{4.5}$$

$$\mathbf{f}(t) = \sigma\left(\mathbf{W}_f\widehat{\mathbf{x}}(t) + \mathbf{U}_f\mathbf{h}(t-1) + \mathbf{V}_f\mathbf{c}(t-1)\right), \tag{4.6}$$

$$\mathbf{o}(t-1) = \sigma\left(\mathbf{W}_o\widehat{\mathbf{x}}(t) + \mathbf{U}_o\mathbf{h}(t-1) + \mathbf{V}_o\mathbf{c}(t)\right), \tag{4.7}$$

$$\tilde{\mathbf{c}}(t) = tanh\left(\mathbf{W}_c\widehat{\mathbf{x}}(t) + \mathbf{U}_c\mathbf{h}(t-1)\right), \tag{4.8}$$

$$\mathbf{c}(t) = \mathbf{f}^i(t) \odot \mathbf{c}(t-1) + \mathbf{i}(t) \odot \tilde{\mathbf{c}}(t), \tag{4.9}$$

$$\mathbf{h}(t) = \mathbf{o}(t) \odot tanh(\mathbf{c}(t)), \tag{4.10}$$

where $\mathbf{i}(t)$ denotes input gate; $\mathbf{f}(t)$ denotes forget gate; $\mathbf{o}(t)$ denotes output gate; $\mathbf{c}(t)$ denotes memory cell; $\mathbf{h}(t)$ denotes hidden state; σ denotes the logistic sigmoind function and $\odot$ denotes elementwise multiplication. The entries of the gating vectors $\mathbf{i}(t)$, $\mathbf{f}(t)$ and $\mathbf{o}(t)$ are in [0, 1]. Intuitively, the forget gate controls the amount of memory that is erased from each memory cell, the input gate controls how much each unit is updated, and the output gate controls the exposure of the internal memory state.

4.4.2 Multi-task Learning Layer

The final output vector of the LSTM layer is fed into two parallel FC layers to train the corresponding weights for each task. The output of the FC layer for the

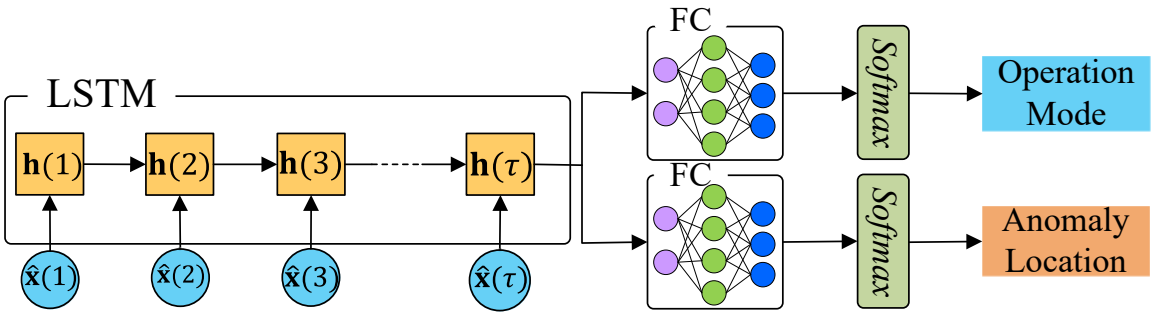


Figure 4.3: Overview of the MTL-LSTM utilized for anomaly classification and location

operation mode task is given by:

$$\mathbf{z}_1^o = \sigma\left(\mathbf{h}(\tau)\mathbf{\Theta}_0^o + \mathbf{b}_0^o\right), \tag{4.11}$$

$$\mathbf{z}_{m+1}^o = \sigma\left(\mathbf{z}_m^o\mathbf{\Theta}_m^o + \mathbf{b}_m^o\right), \tag{4.12}$$

$$\mathbf{z}_M^o = \mathbf{z}_{M-1}^o\mathbf{\Theta}_{M-1}^o + \mathbf{b}_{M-1}^o, \tag{4.13}$$

where $\mathbf{h}(\tau) \in \mathbb{R}^C$ is the output vector of last LSTM layer with C hidden features; $\mathbf{\Theta}_0^o \in \mathbb{R}^{C\times C_1}$ indicates the trainable weight matrix mapping C input features to the output channels C_1; $\mathbf{b}_0^o \in \mathbb{R}^{C_1}$ denotes the trainable bias for the first FC layer; $\mathbf{\Theta}_m^o \in \mathbb{R}^{C_m\times C_{m+1}}, \mathbf{b}_m^o \in \mathbb{R}^{C_{m+1}}$ denote the trainable weight and bias in layer m; $\sigma(\cdot)$ denotes an activation function; $\mathbf{z}_m^o \in \mathbb{R}^{C_m}$ denotes the output of FC in layer m with C_m number of channels; $\mathbf{z}_M^o \in \mathbb{R}^{C_M}$ denotes the output of the FC with M layers and C_M hidden features for the operation mode task. The output of the FC layer for the

anomaly location task is given by:

$$\mathbf{z}_1^l = \sigma\left(\mathbf{h}(\tau)\mathbf{\Theta}_0^l + \mathbf{b}_0^l\right), \tag{4.14}$$

$$\mathbf{z}_{m+1}^l = \sigma\left(\mathbf{z}_m^l\mathbf{\Theta}_m^l + \mathbf{b}_m^l\right), \tag{4.15}$$

$$\mathbf{z}_M^l = \mathbf{z}_{M-1}^l\mathbf{\Theta}_{M-1}^l + \mathbf{b}_{M-1}^l, \tag{4.16}$$

where $\mathbf{\Theta}_0^l \in \mathbb{R}^{C\times C_1}$ indicates the trainable weight matrix mapping C input features to the output channels C_1; $\mathbf{b}_0^l \in \mathbb{R}^{C_1}$ denotes the trainable weight and bias for the first FC layer; $\mathbf{\Theta}_m^l \in \mathbb{R}^{C_m\times C_{m+1}}, \mathbf{b}_m^l \in \mathbb{R}^{C_{m+1}}$ denote the trainable weight and bias in layer m; $\sigma(\cdot)$ denotes an activation function; $\mathbf{z}_m^l \in \mathbb{R}^{C_m}$ denotes the output of FC in layer t with C_m number of channels; $\mathbf{z}_M^l \in \mathbb{R}^{C_M}$ denotes the output of the FC with M layers and C_M hidden features for the anomaly location task.

The *softmax* activation function is used to generate the probability for the predicted label vector for each task in multi-task network which is given by:

$$\begin{cases} \hat{\mathbf{y}}^o = softmax(\mathbf{z}_M^o) \\ \hat{\mathbf{y}}^l = softmax(\mathbf{z}_M^l) \end{cases}, \tag{4.17}$$

where $\hat{\mathbf{y}}^o, \hat{\mathbf{y}}^l$ denote the prediction probability vector for operation mode and location, respectively.

4.4.3 Loss Function

One of the most common approaches to define the cost function in MTL is to formulate the linear combination of individual cost functions as the following:

$$\phi = \lambda^o L(\hat{\mathbf{y}}^o, \mathbf{y}^o) + \lambda^l L(\hat{\mathbf{y}}^l, \mathbf{y}^l), \tag{4.18}$$

where λ^o, λ^l denote the weights for the tasks; $L(\hat{\mathbf{y}}^o, \mathbf{y}^o), L(\hat{\mathbf{y}}^l, \mathbf{y}^l)$ denote the operation mode and anomaly location losses; $\mathbf{y}^o, \mathbf{y}^l$ denote the true labels, respectively.

Both of the above-mentioned tasks are multi-class classification tasks. In each multi-class classification task, the parameters of the neural network are being trained to minimize the cross-entropy loss of the predicted and true labels. The cross-entropy loss is given by:

$$L(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{m=1}^{M}\sum_{c=1}^{C} y_{m,c}\log(\hat{y}_{m,c}), \tag{4.19}$$

where $y_{m,c} \in \{0, 1\}$, $\hat{\mathbf{y}}_{n,c} \in [0, 1]$ are the target and predicted labels for m^{th} sample and c^{th} class respectively; M denotes the number of training samples and C denotes the class number.

4.5 Case Study

4.5.1 Developed Real-time Testbed

The proposed RALCON model is implemented on multiple testbeds and the results in this section are presented based on the simulation outcomes on IEEE 33-bus test feeder. This system has 37 line sections, supplying a total load of 5084 kW and 2547 kVAr. The system is served by the upstream transmission network, as well as by four Solar Generation (SG) units connected to buses $14, 18, 22, 33$ through controllable inverters with the rated active power of $250, 150, 150, 250$ kW, respectively. Four ES units with the same power ratings of 400 kW are coupled to the same buses. The single line diagram with the base configuration of the system, including the location of the circuit breakers, reclosers, protection relays, D-PMUs, and DERs, is shown in Fig. 4.4, and the detailed system data is provided in [35]. Three relays control the operation of circuit breakers and reclosers in the lines connecting buses $1-2$, $13-14$, and $28-29$. In addition, three D-PMUs monitor and stream the measurements in the lines connecting buses $1-2$, $17-18$, and $32-33$. The line connecting buses $1-2$ is equipped with both D-PMU and protection relay since it corresponds to the substation. As a consequence, for monitoring purposes we will not use the relay data in this bus but instead the high-resolution data from D-PMUs, such that the number of D-PMUs is $U = 3$ and the number of relays is $R = 2$ for the test system.

In total, 5 sensors are fed into the "Data Aggregation" module. The developed testbed is modeled in Hypersim, interfacing with OPAL-RT real-time Hardware-in-the-loop (HIL) simulator. The testing environment is capable of executing real-time simulations with a time-step of up to 50μs. Hardware D-PMUs have been added to the loop of the simulation with high resolution data capturing. Additionally, relay operation models are considered and transient events are sequentially scripted and run respectively. A varying load profile is modeled to be connected to all the buses and full control models are adopted for all the SG and ES. We consider normal scenario alongside with different evens including, three-phase-to-ground short-circuit bolted fault at the end of all lines ($L = 32$), DER switching on buses $\{14, 18, 22, 33\}$ ($D = 4$), as well as relay cyber-attacks on buses $\{1, 14, 29\}$ ($R = 3$). Hence, the total number of classes for the anomaly location equals to: $K^l = \max\{32, 3, 4\}+1 = 33$. All the scenarios are simulated considering different load profile. Table 4.1 summarizes number of samples generated in the simulation process.

4.5.2 Data aggregation performance

The aggregated measurements obtained from relays and D-PMUs with synchronized sampling rates for all the modeled events have been evaluated with respect to the D-PMU high resolution measurements. As far as the validation of the data aggre-

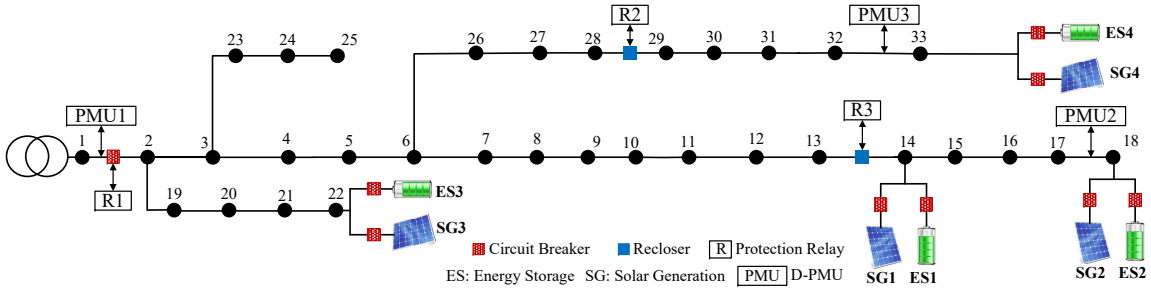


Figure 4.4: IEEE 33-bus test feeder

Table 4.1: Opal-RT simulated Dynamic Events

Event Type	Location	Number of Anomalies
Short-circuit Fault	end of all lines	297
FDIA Cyber-attack	buses {1, 14, 29}	27
DER Switching	buses {14, 18, 22, 33}	36

gation is concerned, two scenarios have been considered in the modified IEEE 33-bus system:

- **Scenario 1:** D-PMU measurements at buses $\{14, 29\}$ with no relay measurements.
- **Scenario 2:** Relay measurements at buses $\{14, 29\}$, being aggregated with other measurements using the proposed approach for synchronization of the

sampling rates.

For all the developed usecases, Scenario 1 is considered as the base case and a statistical comparison is conducted between the measurements in both scenarios. Mean and standard deviation of the generated error is calculated for voltage and current phasors and average values are shown in Table.4.2. Mean value of the error between measurements is tuned to be below 1, reflecting a close estimation of the actual values. In the case of standard deviation of the error measurements, aggregated results have been performing close to D-PMU outputs in case of faults, while DER events and cyber attacks have more deviations.

Table 4.2: Measurement Aggregation Comparative Performance

Performance	Faults	Cyber attacks	DER Events
Avg Mean	0.71	0.94	0.79
Avg Std	9.88	14.93	12.51

Given the nature of the simulated FDIA events, a topology reconfiguration is inevitable, resulting in more required tuning effort for $\varepsilon_{\text{threshold}}$ as well as recalculation of the proximity vector after each attack. Furthermore, DER events are primary considered as low-scale events in grid-connected mode, as the amount of the switched active and reactive power by the event are less than 14% of the total system load. This,

in turn, result in increase of the standard deviation of the aggregated measurements.

4.5.3 LSTM Dataset Generation

In order to evaluate the performance of the proposed MTL-LSTM, we need to re-organize and label the samples for a fixed time step window. The MTL-LSTM dataset generation is as follows: We select the samples of the simulation from time $t = 0$ to $t = \tau$ for each operation mode and label the operation mode and the anomaly location to generate a sample for the MTL-LSTM dataset. For the second sample we shift t one time step ahead (i.e., $t = 1$ to $t = \tau + 1$) and repeat the process. We repeat the process for all the simulated scenarios. Table 4.3 illustrates the number of generated samples for each operation mode for $\tau = 60$.

Table 4.3: Number of samples for each operation mode for $t = 60$

Operation mode	Number of samples
Normal Operation	4740
Short circuit Fault	27000
Cyber attack	2160
DER Switching	7200
Total	41100

4.5.4 Hyper Parameter Selection

The hyper-parameters of MTL-LSTM in the proposed RALCON are selected through 5-fold cross-validation. The cross-validated layer architecture includes LSTM layers with 60 samples time window followed by a dropout regularization. Two parallel FC layers, each with two hidden layers with $ReLU$ activation function and dropout, and an output layer with *softmax* activation for operation mode and anomaly location are followed by the LSTM layers. The weight multiplier for each task is set as one throughout the experiment.

Initially, the dimensionality of the LSTM layer output, as well as the FC hidden layers, are set to 100 while the dropout is set to 0 for all the layers. The number of 64 samples is considered to create a batch. Different optimizers such as Adam, RMSprop, SGD were tested for various learning rates to find the best activation function and optimizer. The best results for the initial structure obtained for the RMSprop optimizer, which has 94.01% validation accuracy for the operation mode and 96.01% validation accuracy for location classification (accuracy is defined as the number of true predicted samples divided by the total number of samples).

Using these results, the other hyper-parameters are selected based on a trial search, where a different number of layers, neurons and dropout probability for scenarios are tested to find the best MTL-LSTM structure. According to the results, the

best number of hidden features for LSTM is 150, the number of neurons for operation mode FC layer is 200, 100 and the number of neurons for the event location is 150, 100 and the dropout rate is 0.3.

4.5.5 RALCON Testing Result

The proposed MTL-LSTM is trained for 50 epochs using the obtained hyper-parameter from the previous section, resulting in training and testing accuracy of 98.25%, 98.16% for operation mode task, and 96.98%, 96.68% for event location task, respectively. The batches of samples are fed into the proposed MTL-LSTM. Selecting the batch size is a vital decision since it has a substantial impact on the performance of the model. Therefore, the proposed model is trained for different batch sizes in order to investigate the impact of batch size on the performance of the proposed MTL-LSTM.

The accuracy of the prediction for both training and test dataset for different batch sizes is reported in Table 4.4. As can be seen, the RALCON performs fairly robust against different batch sizes. However, the model with a batch size of 64 outperforms other scenarios with a prediction accuracy of 98.16% in operation mode and 96.68% in location classification.

The confusion matrix for the operation mode task for the testing data is de-

picted in Fig. 4.5, where the diagonal elements represent the percentage of accurate prediction, while the off-diagonal elements are those that are mislabeled by the classifier. In Fig. 4.5, the majority of samples are predicted correctly in all of the events aligning with the high accuracy of the proposed model in anomaly classification. The misclassification has mostly occurred between intermediate and DER events, showing that it is slightly more possible to misinterpret the intermediate and the DER event. There are cases that RALCON classifies an intermediate mode as other operation modes, where the signatures of the intermediate mode bear similarities to different modes.

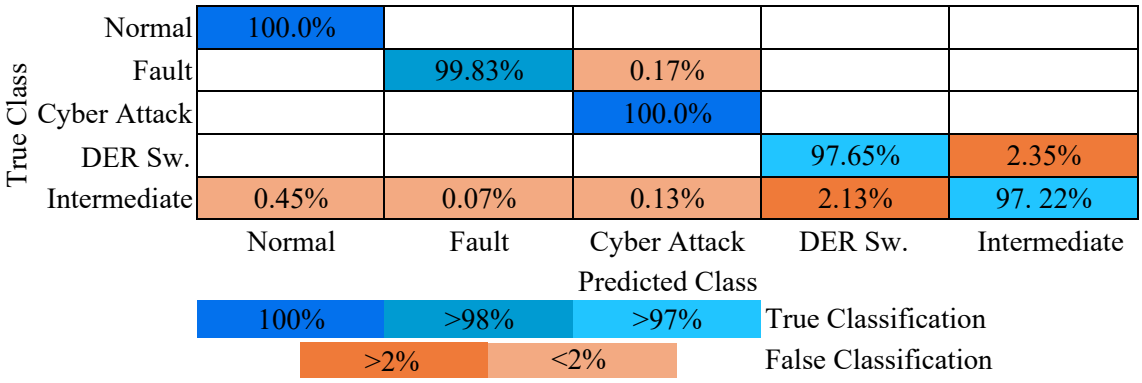


Figure 4.5: Confusion matrix for the proposed RALCON system

Table 4.4: The accuracy of the RALCON for different batch-sizes

Dataset	Task	Batch Size			
		16	32	64	96
Training Accuracy(%)	Operation mode	97.49	97.46	**98.25**	97.44
	Anomaly location	96.63	93.56	**96.98**	94.42
Test Accuracy(%)	Operation mode	97.56	97.18	**98.16**	97.44
	Anomaly location	95.31	93.50	**96.68**	95.97

4.5.6 The Impact of Time Steps on RALCON Performance

The number of consecutive samples required for the RALCON, namely time steps, is another important hyper-parameter to select. In addition to the performance

of the MTL-LSTM, the time steps gain more importance since it specifies the amount of measurement that need to be stored on the storage system.

The performance of the proposed RALCON with different operation modes is investigated here through the following five cases, which indicate different testing datasets:

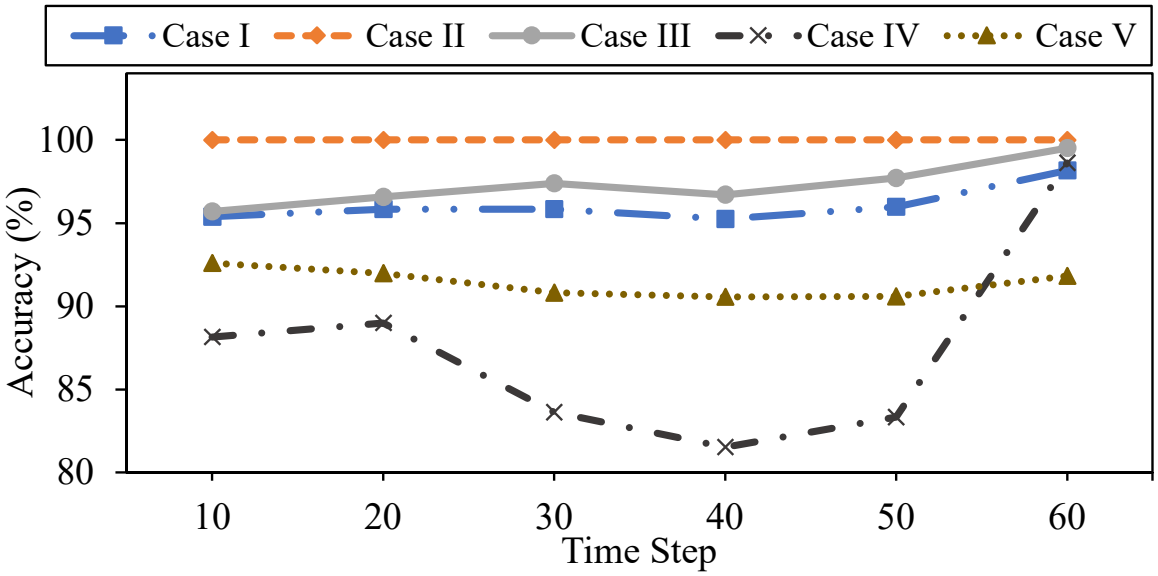
- Case I: The accuracy of the testing sample is evaluated for all the operation modes.
- Case II: Only the normal operation samples are selected and the accuracy of the samples is calculated.
- Case III: Only the short-circuit fault samples are selected and the accuracy is calculated.
- Case IV: Only the cyber attack samples are selected and the accuracy is calculated.
- Case V: The DER switching samples are selected and the accuracy is calculated.

The performance of RALCON operation mode and location classification for Cases I-V of different test datasets and time steps between 10 and 60 are summarized in Fig. 4.6. In Fig. 4.6, the accuracy of all cases except with 60 time steps gain the highest value for both operation mode and anomaly location tasks comparing the

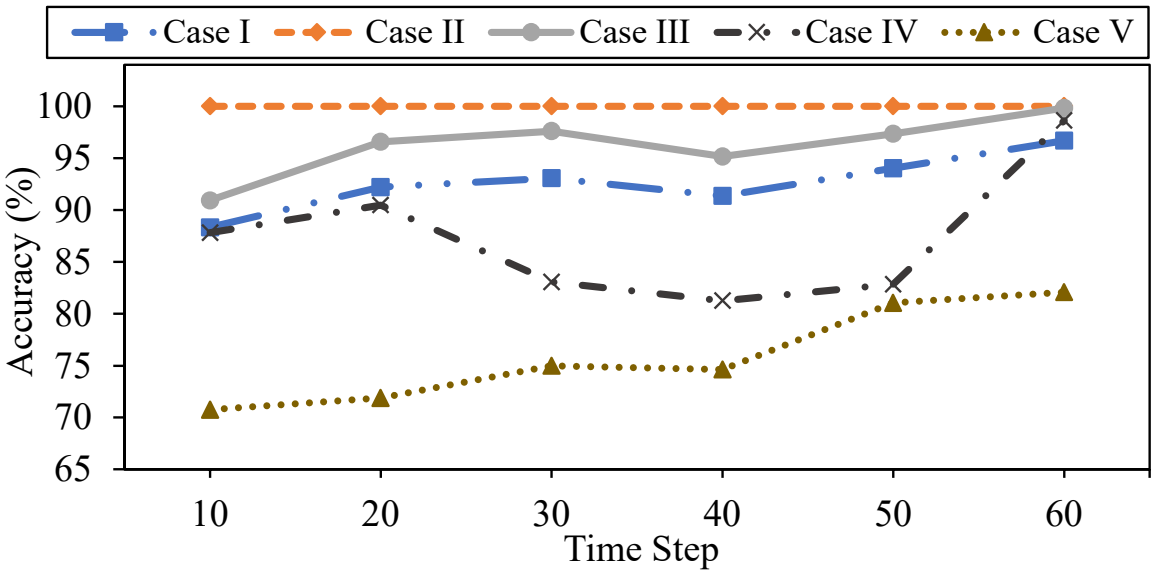
other time steps illustrating that selecting 60 time steps for the MTL-LSTM leads to the best performance of the proposed approach. While Cases I-III experience almost consistent performance for different time steps, Case IV and Case V are highly dependant on the time step size. In Fig. 4.6, the accuracy of Case IV where only the cyber attack samples are evaluated drops dramatically where the size of time step increases from 10 to 50 and then recovers sharply with 60 time steps. In Fig. 4.6a, the operation mode accuracy of Case V decreases slightly where the size of the time step increases from 10 to 50 and then increases with 60 time steps. In Fig. 4.6b, the location accuracy of Case V obtains the lowest value for all time steps comparing the other cases. Nevertheless, selecting a higher value for the time step provides better performance. The result shows that selecting 60 time steps for the proposed model provides the best performance for operation mode and location classification tasks illustrating that this number of time steps include the required samples for the proposed MTL-LSTM for classification.

4.5.7 Impact of Input Data on RALCON

The impact of input data on the performance of the proposed RALCON is investigated here through three different scenarios. Only the measurements from D-PMUs are used to train the MTL-LSTM and the protection relay data is removed in



(a) Operation mode classification accuracy of the test dataset



(b) Anomaly location classification accuracy of the test dataset

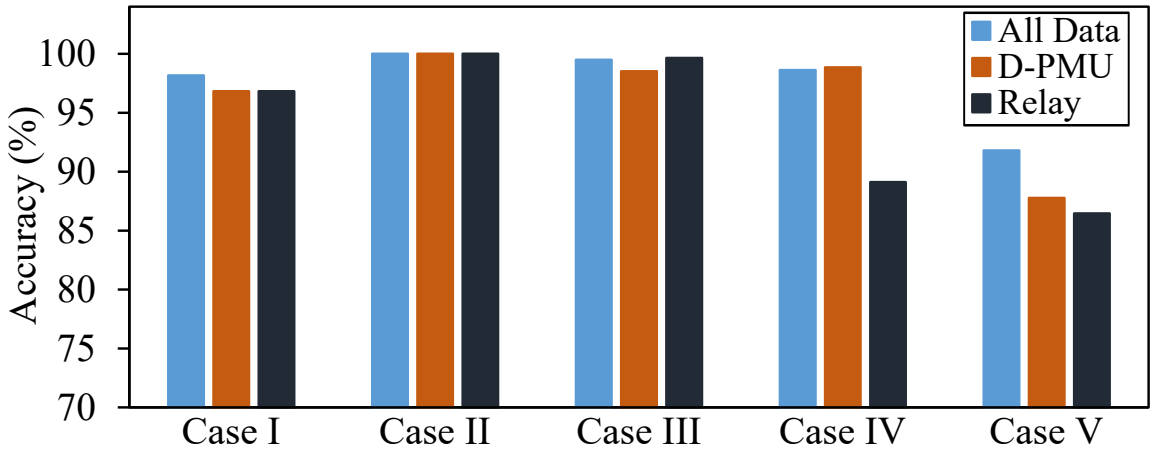
Figure 4.6: The accuracy of the proposed RALCON for different time steps

a scenario. In another scenario, only the protection relay data is used to train the MTL-LSTM. The obtained results are compared with the previous section in which both D-PMU and protection relay data are used to train the proposed network.

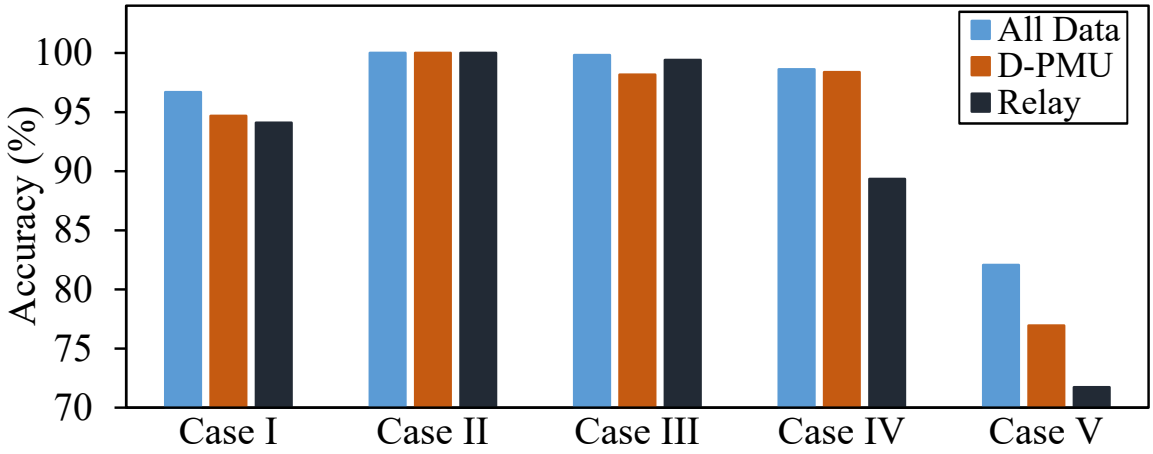
The performance of RALCON for different input data for both anomaly classification and location are summarized in Fig. 4.7. In Fig. 4.7, the RALCON with all input data outperforms using only D-PMU and relay data in all cases in both operation mode and anomaly location classification, illustrating the significance of both data types. The RALCON experiences dramatic accuracy reduction for cases IV, V where only the relay data is used, illustrating the importance of the D-PUM data to classify and locate cyber attacks and DER switching. In Case III, obtaining high accuracy with only protection relays data compared to the D-PMU data depicts the significance of using protection relay data for fault classification and location

4.6 Summary

This research is a collaborative work with University of Utah under th DOE SolarSTARTS project and proposes the RALCON system to classify and distinguish between normal operation and several anomalies, including short-circuit fault, cyber attack and DER switching and locate the anomaly in a real-time application. More specifically, the proposed RALCON first integrates the measurement data from D-



(a) Operation mode classification accuracy of the test dataset



(b) Anomaly location classification accuracy of the test dataset

Figure 4.7: The accuracy of the proposed RALCON for different input data

PMU and protection relay devices with different sampling rates using an aggregation strategy with existing data and then feeds the integrated data to a carefully-tuned multi-task learning-based LSTM to classify and locate the potential anomaly in the

distribution system. The proposed model can be integrated and utilized in real-time monitoring to enhance situational awareness. The simulation results on the IEEE 33-bus test feeder demonstrate that the proposed model classifies the operation mode of the distribution system and their location with high accuracy of more than 98% and 96%, respectively. The implementation of RALCON on the IEEE 33-bus test system illustrates that the normal operation and the cyber-attack samples are correctly classified. Correspondingly, the high percentage of the other operation modes are correctly predicted in the test dataset. In addition, experiments conducted for a different number of samples illustrate the importance of the time window on the performance of the LSTM, especially on the cyber-attack and DER switching classification and location. Further, experiments conducted for different sets of input data illustrate the importance of both D-PMU and protection relay data in the distribution system to enhance the anomaly classification and location performance. Moreover, the protection relay data has a remarkable impact on fault classification and location, while the D-PMU data plays a critical role in cyber-attack and DER switching classification and location.

REFERENCES

[1] M. Farajollahi, A. Shahsavari, E. M. Stewart, and H. Mohsenian-Rad, "Locating the source of events in power distribution systems using micro-pmu data," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6343–6354, 2018.

[2] R. Arghandeh, M. Gahr, A. von Meier, G. Cavraro, M. Ruh, and G. Andersson, "Topology detection in microgrids with micro-synchrophasors," in *2015 IEEE Power & Energy Society General Meeting*, pp. 1–5, IEEE, 2015.

[3] A. Gholami, A. K. Vosughi, and A. K. Srivastava, "Denoising and detection of bad data in distribution phasor measurements using filtering, clustering and koopman mode analysis," *IEEE Transactions on Industry Applications*, pp. 1–1, 2022.

[4] M. Parvania, G. Koutsandria, V. Muthukumary, S. Peisert, C. McParland, and A. Scaglione, "Hybrid control network intrusion detection systems for automated power distribution systems," in *2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, pp. 774–779, 2014.

[5] M. Dawson, M. Omar, and J. Abramson, "Understanding the methods behind cyber terrorism," in *Encyclopedia of Information Science and Technology, Third Edition*, pp. 1539–1549, IGI Global, 2015.

[6] D. U. Case, "Analysis of the cyber attack on the ukrainian power grid," *Electricity Information Sharing and Analysis Center (E-ISAC)*, 2016.

[7] "SECURITY: First-of-a-kind U.S. grid cyberattack hit wind, solar," 2019.

[8] A. Gholami, A. K. Srivastava, and S. Pandey, "Data-driven failure diagnosis in transmission protection system with multiple events and data anomalies," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 4, pp. 767–778, 2019.

[9] M. Zhou, Y. Wang, A. K. Srivastava, Y. Wu, and P. Banerjee, "Ensemble-based algorithm for synchrophasor data anomaly detection," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2979–2988, 2019.

[10] M. Ganjkhani, M. Gilanifar, J. Giraldo, and M. Parvania, "Integrated cyber and physical anomaly location and classification in power distribution systems," *IEEE Transactions on Industrial Informatics*, 2021.

[11] S. Pandey, A. K. Srivastava, and B. G. Amidan, "A real time event detection, classification and localization using synchrophasor data," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4421–4431, 2020.

[12] J. C. Mayo-Maldonado, J. E. Valdez-Resendiz, D. Guillen, M. Bariya, A. von Meier, E. A. Salas-Esquivel, and A. Ostfeld, "Data-driven framework to model

identification, event detection, and topology change location using d-pmus," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6921–6933, 2020.

[13] S. Som, R. Dutta, A. Gholami, A. K. Srivastava, S. Chakrabarti, and S. R. Sahoo, "Dpmu-based multiple event detection in a microgrid considering measurement anomalies," *Applied Energy*, vol. 308, p. 118269, 2022.

[14] A. Shahsavari, M. Farajollahi, E. M. Stewart, E. Cortez, and H. Mohsenian-Rad, "Situational awareness in distribution grid using micro-pmu data: A machine learning approach," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6167–6177, 2019.

[15] M. Farajollahi, A. Shahsavari, E. M. Stewart, and H. Mohsenian-Rad, "Locating the source of events in power distribution systems using micro-pmu data," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6343–6354, 2018.

[16] Y. Zhou, R. Arghandeh, and C. J. Spanos, "Partial knowledge data-driven event detection for power distribution networks," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5152–5162, 2018.

[17] M. Rafferty, X. Liu, D. M. Laverty, and S. McLoone, "Real-time multiple event detection and classification using moving window pca," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2537–2548, 2016.

[18] D. Nguyen, R. Barella, S. A. Wallace, X. Zhao, and X. Liang, "Smart grid line event classification using supervised learning over pmu data streams," in *2015 Sixth International Green and Sustainable Computing Conference (IGSC)*, pp. 1–8, IEEE, 2015.

[19] A. Gholami and A. K. Srivastava, "Comparative analysis of ml techniques for data-driven anomaly detection, classification and localization in distribution system," in *2020 52nd North American Power Symposium (NAPS)*, pp. 1–6, 2021.

[20] I. Niazazari and H. Livani, "Disruptive event classification using pmu data in distribution networks," in *2017 IEEE Power & Energy Society General Meeting*, pp. 1–5, IEEE, 2017.

[21] K. Chen, J. Hu, Y. Zhang, Z. Yu, and J. He, "Fault location in power distribution systems via deep graph convolutional networks," *IEEE Journal on Selected Areas in Communications*, 2019.

[22] A. Vosughi, A. Gholami, and A. K. Srivastava, "Denoising and bad data detection in distribution phasor measurements using filtering, clustering and koopman mode analysis," IAS Annual Meeting 2021, Oct 2021.

[23] M. Kezunovic, C. Zheng, and C. Pang, "Merging pmu, operational, and non-operational data for interpreting alarms, locating faults and preventing cas-

cades," in *2010 43rd Hawaii International Conference on System Sciences*, pp. 1–9, IEEE, 2010.

[24] A. Sahu, Z. Mao, P. Wlazlo, H. Huang, K. Davis, A. Goulart, and S. Zonouz, "Multi-source data fusion for cyberattack detection in power systems," *arXiv preprint arXiv:2101.06897*, 2021.

[25] S. Siamak, M. Dehghani, and M. Mohammadi, "Dynamic gps spoofing attack detection, localization, and measurement correction exploiting pmu and scada," *IEEE Systems Journal*, pp. 1–10, 2020.

[26] S. Pan, T. Morris, and U. Adhikari, "Classification of disturbances and cyber-attacks in power systems using heterogeneous time-synchronized data," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 650–662, 2015.

[27] I. Siniosoglou, P. Radoglou-Grammatikis, G. Efstathopoulos, P. Fouliras, and P. Sarigiannidis, "A unified deep learning anomaly detection and classification approach for smart grid environments," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1137–1151, 2021.

[28] E. Khaledian, S. Pandey, P. Kundu, and A. K. Srivastava, "Real-time synchrophasor data anomaly detection and classification using ¡italic¿isolation forest, kmeans, and loop¡/italic¿," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2378–2388, 2021.

[29] M. Gilanifar, J. Cordova, H. Wang, M. Stifter, E. E. Ozguven, T. I. Strasser, and R. Arghandeh, "Multi-task logistic low-ranked dirty model for fault detection in power distribution system," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 786–796, 2020.

[30] Y. Guo, Z. Pang, J. Du, F. Jiang, and Q. Hu, "An improved alexnet for power edge transmission line anomaly detection," *IEEE Access*, vol. 8, pp. 97830–97838, 2020.

[31] A. Gholami, M. Mousavi, A. K. Srivastava, and A. Mehrizi-Sani, "Cyber-physical vulnerability and security analysis of power grid with hvdc line," in *2019 North American Power Symposium (NAPS)*, pp. 1–6, 2019.

[32] M. Zhou, Y. Wang, A. K. Srivastava, Y. Wu, and P. Banerjee, "Ensemble-based algorithm for synchrophasor data anomaly detection," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2979–2988, 2018.

[33] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[34] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," *arXiv preprint arXiv:1605.05101*, 2016.

[35] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems

for loss reduction and load balancing," *IEEE Power Engineering Review*, vol. 9, no. 4, pp. 101–102, 1989.

CHAPTER 5. OUTAGE EVENT IDENTIFICATION AND ROOT CAUSE ANALYSIS

5.1 Introduction

Power outages have been an inevitable part of power delivery systems. In case of large scale outages with conflicting information in the distribution network with distributed energy resources (DER), recognition of the root causes becomes increasingly complex. Root cause detection is very important for routing the crew and for faster restoration. Number of studies have been conducted on analysis of the outages, planning predictive actions, and improving the restoration process for minimizing the load outages using various statistical and data processing. A comprehensive methodology with the ability to find root causes in quick and efficient manner is still missing. This work proposes an Outage Root Cause Analysis (ORCA) tool, with various internal modules. The proposed ORCA utilizes the available data from diverse set of sensors (Distribution Phasor Measurement Units, metering and relay) within the distribution system (DS) to: a) aggregate using data fusion algorithm: an Ensemble Extended Kalman Filter (EEKF) approach, b) classify events at broader level (fault, breaker

operation or cyber-attack with/without load outage) using hierarchical agglomerative clustering in online and offline manners, and c) narrow down the outage root cause to the possible nuanced identified causes (vegetation, animal, weather, wildfire, protection, equipment failure, planned) using the available historical data and the Frequent Pattern-Growth data mining approach. Simulation results demonstrate the superiority of the proposed approach compared to the other existing key approaches using two different test systems and multiple outage case scenarios.

Reliable and resilient operation of electric Distribution System (DS) requires minimizing number of outages and outage duration. However, preventing all the power delivery interruptions on a daily DS operation seems to be impossible, due to several unpredictable physical and environmental factors. Unexpected events are threats to the distribution system operation, causing disconnection of critical loads as hospitals, and fire departments impacts resiliency. Maintaining the reliability and the resiliency of the system is challenging, and it is critical to avoid prolonged outage periods. The root cause of the outage can be variety of possibilities specially with increasing edge devices and distributed energy resources (DERs), each of which are accountable for a significant set of power disruptions that consumers encounter. As long as the cause of the outage is not detected, the restoration process cannot be implemented in an effective manner. Therefore, Root Cause Analysis (RCA) of the load outages are critical for electric utilities to decrease the duration of the power

interruptions.

Over the last several years with the increased accessibility to the recorded power interruption information, recorded by utility companies, the amounts of research works addressing the issue of outage management and RCA has exponentially grown[1, 2, 3]. The conventional approach to extract a level of understanding from the outage data used to be more around statistical strategies and derivation of the system model, however, these approaches are mostly associated with huge computational costs, as it requires performing a range of mathematical analysis, therefore making it time consuming and not appropriate for live applications. Furthermore, exploiting highly sophisticated methods such as Cox and Logistic Regression are potentially able to bring higher nuanced insight of the outages and the corresponding possible root causes (RC's), nevertheless, these methods are associated with certain computational challenges in addition to remarkable quantities of running times as well [4, 5, 6, 7, 8, 9].

The proposed Outage Root Cause Analysis (ORCA) in this work, consists of different modules, aiming towards an efficient usage of the available data in DS, in streaming and non-streaming manners, and identification of the outage events with corresponding RC. The primary objective of the non-streaming module of the proposed ORCA is to contribute towards detection of the large outages with conflicting information, which majorly cannot be identified using conventional approaches. The

structure of the proposed methodology is organized on a hierarchical manner starting with accumulation of the data, followed by data processing and clustering, and finally rule-based mining of the outage historical data, resulting in identification of the detailed RCs. The contributions of the work are summarized in the following key points:

- Developed algorithms to aggregate data from diverse set of distribution sensors (e.g. Distribution Phasor Measurement Units (D-PMU), metering and relay) and data fusion tailored for analyzing large scale outages using the Ensemble Extended Kalman Filter (EEKF) assisted by Hankel Alternative View of Koopman (KAVOK) approach [10]

- Developed outage event root cause broader level classification (fault, breaker operation or cyber-attack with/without load outage) using a bottom-top hierarchical clustering in offline and online manners

- Developed algorithms for extracting frequent patterns in outage data using unsupervised frequent pattern-growth algorithm for detailed identification of the outage root cause (vegetation, animal, weather, wildfire, protection, equipment failure, planned and multiple sub-root causes) assisted by root cause hypothesis generation and feature selection

5.2 Related Work

The proposed ORCA addresses the issue of outage management with regard to detection and root cause analysis. The scope defines under any DS circumstances, leading to a load outage, both in direct and indirect manners. This work categorizes the events of interests by the criteria of leading to a load outage, and further investigates the nuanced possible causes by processing the available historical data, which is invoked by the results of the outage detection.

The related established works in the literates are primarily divided into load flow identification, [11, 12], or data mining approaches[13, 14, 15, 16, 17]. Authors of [18] narrow down the load outage causes to only fault and develops a rapid detection technique to recognize the outaged line. Ahmed et al. develops a probabilistic approach, assuming various line outages happening at the same time [19]. The invention of high resolution measurements, triggers the work on [20, 21, 22, 23, 24, 25, 26] to find the optimal locations of sensors, which can provide sufficient help on finding the outages. A conventional outage detection study has been completed in [27], using DS load flow measurements. A detailed RC-based outage analysis in accomplished by the authors of [28], aiming towards prediction and prevention of animal-related causes. In a similar work, [29], concentrates the analysis on vegetation-related root causes, that is continued by [30] in a data-driven manner. One of the early researches on us-

ing data mining approaches for outage detection, has completed a nuanced study on vegetation-related outages ,[31], with focus on tree-induced power interruptions. By recent advancements in Machine learning (ML) technology, more attention has been attracted to possible applications in DS outage management. Given the high temporal and spatial dependencies among the measurements in DS, authors of [32], develop a cause and effect scheme to identify the outages; however, this work only considers outages triggered by fault causes and do not include other events. [33] addresses the issue of outages due to severe weather conditions and develops a ML-based predictive model to prevent the power disruptions. In a more comprehensive study, an extensive analysis of different ML techniques, including supervised, semi-supervised, and unsupervised, approaches is done in [34], in which the authors bring domain-based insights on potential benefits and drawbacks for each of the analyzed techniques with respect to the application.

While the outage management studies have been conveyed for several years, still a comprehensive methodology that takes use of the modern data mining and ML advancements together as a single package is lacking. Fig.5.1 shows the framework of the proposed ORCA, including the three stages of Data Fusion, Outage Detection (Online and Offline), and identification of the detailed RC in a single integrated tool that can be used in utility applications and has the capability of being implemented over the data streams as well.

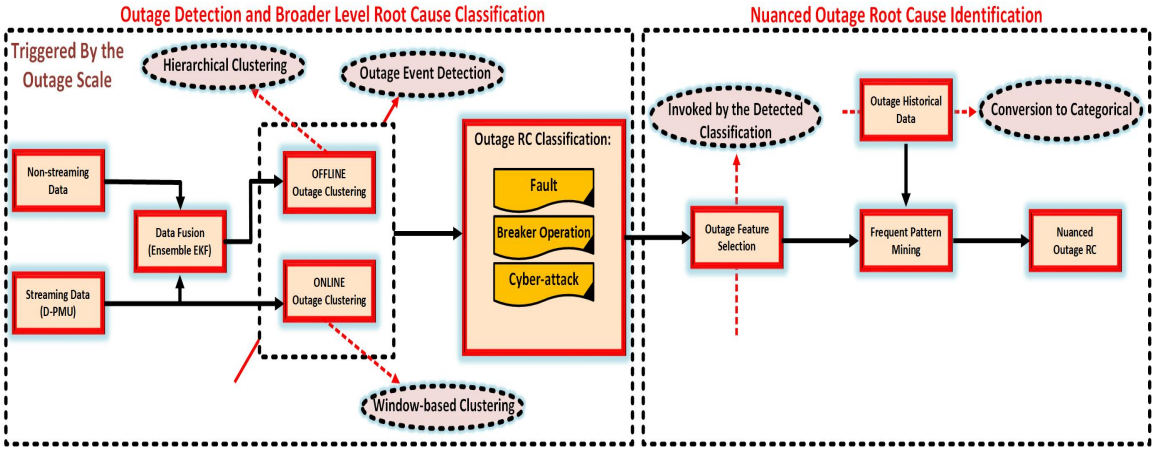


Figure 5.1: Overview of the Proposed ORCA: Outage Root Cause Analysis for Distribution System

5.2.1 Input and Output Data

Considering a distribution system with a set of measurement sensors $\mathcal{S} = \{1, ..., S\}$, where S is the number of total sensors including protection relays and D-PMUs. The aggregation of the measurements for the offline ORCA in the current research work is based on the fusion of D-PMU data and protection relay data, while the developed EEKF has the capability of data aggregation from multiple measurement resources. In the offline mode the OMS historian data are to be triggered by the system operator through a post-outage analysis. The set of sensors $\mathcal{S}$ in this work consists of a subset of protection relays $\mathcal{R} = \{1, ..., R\}$, and a subset of D-PMUs $\mathcal{U} = \{1, ..., U\}$, where R is the number of protection relays and U is the number of

D-PMUs.

The collected data consists of the measurement data from relays and D-PMUs, which are continuously being received and stored in the OMS database. The online ORCA continuously receives and processes the streaming D-PMU data, while the offline module receives the data from the OMS historical system for a specific time window. The data employed by offline ORCA includes measurements from both high-resolution and low-resolution measurement devices, elaborated as follows:

5.2.1.1 D-PMU Measurements

Three-phase voltage and current magnitudes and angles, frequency and rate of change of the frequency (ROCOF) are transmitted to the control center with the rate of 120 samples/sec.

5.2.1.2 Protection Relay Data

Three-phase voltage and current magnitudes and frequencies are transmitted to the control center with the rate of 10 samples/sec.

Let us define $\mathbf{Q}(t) = [\mathbf{x}_1(t), ..., \mathbf{x}_s(t)...., \mathbf{x}_S(t)] \in \mathbb{R}^{14\times S}$ as the measurement matrix for a distribution system, where t denotes the time index; and $\mathbf{x}_s(t)$ denotes

the sensor measurement vector defined as the following:

$$\begin{aligned}\mathbf{x}_s(t) = [&I_{s,a}(t), I_{s,b}(t), I_{s,c}(t), \alpha^I_{s,a}(t), \alpha^I_{s,b}(t), \alpha^I_{s,c}(t),\\ &V_{s,a}(t), V_{s,b}(t), V_{s,c}(t), \alpha^V_{s,a}(t), \alpha^V_{s,b}(t), \alpha^V_{s,c}(t),\\ &f_s(t), rf_s(t)]^T \end{aligned} \tag{5.1}$$

where $I_{s,a}(t)$, $I_{s,b}(t)$, $I_{s,c}(t)$, $V_{s,a}(t)$, $V_{s,b}(t)$, $V_{s,c}(t)$ denote the current and voltage magnitudes; $\alpha^I_{s,b}(t)$, $\alpha^I_{s,c}(t)$, $\alpha^V_{s,a}(t)$, $\alpha^V_{s,b}(t)$, $\alpha^V_{s,c}(t)$ denote the current and voltage angles for sensor $s \in \mathcal{S}$ at time t in three phases a, b, c, respectively; $f_s(t), rf_s(t)$ denote the frequency and ROCOF for sensor s at time t, respectively. All the arrays of vector in $x_s(t)(\forall s \in \mathcal{R})$ considered to be 0 if the relay does not provide measurement data at time sample t. Correspondingly, any features that are not being measured by relays, i.e. angle values and ROCOF, considered to be 0.

5.3 Measurement Processing Module

In a real-world daily operation of a DS, an essential consideration of the system status lies under an appropriate interpretation of the receiving measurements. Without a proper synchronization of data, acquired data-based knowledge might end up with deficiencies, resulting in a flawed decision-making process. Hence, a significant primary step is to process the incoming data from different measurement resources with different sampling rates and develop an approach to effectively integrate, clean,

and synchronize all the incoming information, based on which the proposed ORCA applications can be studied and applied with higher levels of competency. In the current research work, the developed data fusion approach to address this issue, called Ensemble Extended Kalman Filter (EEKF), is schematically shown in Fig. 5.2. When dealing with imperfect raw data, data fusion can be deployed in order

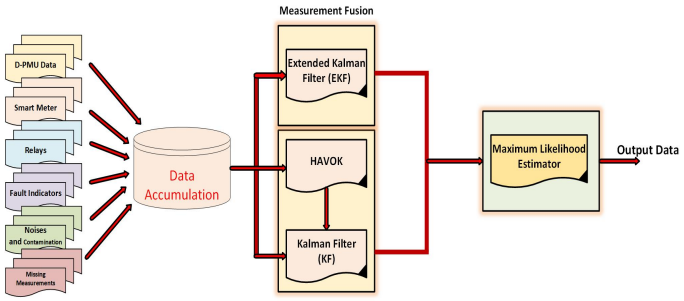


Figure 5.2: The Developed EEKF Data Fusion Approach

to help increase the reliability and credibility of the information.Contrary to classical probabilistic data fusion techniques, deep learning, which automatically learns from past experiences without explicitly programming, remarkably renovates fusion techniques by providing the capability of predicting and computing. In spite of this, no comprehensive study is available to show the application of the latest advances in machine learning for data fusion with focus on Distribution System measurement specifics. Therefore, as a part of the goals of the current research work, we have done a comprehensive investigation on Machine Learning developments in the area of data fusion and have tried applying the methods for distribution system data. In this sec-

tion, we will first go through the architecture of the data fusions, tailored for DS data and then will introduce the methodology that we have used for this purpose. Such Probabilistic fusion methodologies as Bayesian fusion, evidential belief reasoning fusion, and rough set-based fusion are examples of conventional data fusion approaches. With the advancements in ML, an extensive formulation and costly solving process of the data fusion problem in DS is not needed, and instead, ML allows the computer system to find out the underlying structure of the data with supplied information to target the discovering of the connection between the input available data points. In the case of the DS data with multiple sensors being deployed and data coming from different resources, the fusion techniques that is used is "Data-In, Data-Out" (DAI-DAO), which means the input data are considered to be the sensor raw data, expecting the output of the fusion to be still with the same features and underlying relations, while with higher resolutions and less noise levels.

The data fusion technique, proposed as part of the ORCA, is built upon the conventional Kalman Filter (KF) approach with specific modifications to make it customized for the DS system measurements. Furthermore, integration of the KF estimations, with forced system linearization, with the Extended Kalman Filter (EKF) approach and finally the fusion results will be outputed. The proposed method is called Ensemble Extended Kalman Filter (EEKF) and uses a multi-measurement data fusion strategy to boost the precision as well as the integrity of the data by detection

and removal of the noise levels, and synchronizing the data resolution from different sensors. This methodology for the fusion of data, incorporates the information from numerous sensing units with their corresponding streaming or non-streaming information to attain an enhanced precision, which can be further used for such other operating purposes as outage detection and root cause analysis. Considering the final purpose of the fusion, which is to be applied in the control room, the current work puts more focus on centralized type of data aggregation, while maintaining the criteria for distributed applications as well. With the centralized approach, all the raw information obtained from different sensors at various locations of the system are to be sent out to a single integration unit to be merged; however, in the case of the decentralized fusion, the filtering system procedure is regionally split between neighbors that operate the fusion strategy in addition to a master center to aggregate the regional estimates to generate an enhanced result for the entire system.

5.3.0.1 Original and Extended Kalman Filter

The two steps of prediction and correction are considered, for which a linear approximation of the system model is calculated to predict the measurements as well as the uncertainty.

With the DS being highly dynamic with continuous changes in the system status and amount of load and generation as well as switches status, data manipulation with noise and missing values is inevitable. Furthermore, sensor malfunctions lead-

ing to conflicting data exists in some cases as well. In a fuzzy environment multi-measurement fusion using EKF is considering a range of diverse sensing units as input elements, EKF suits a better fit to address the high non-linearity of the system, compared to the original KF.

$$\begin{aligned} \hat{x}_{n|n-1} &= L\hat{x}_{n-1|n-1} + I'u_n \\ P_{n|n-1} &= LP_{n-1|n-1}L' + C_n \end{aligned} \tag{5.2}$$

Where L is to represent the linear approximation of the system underlying model, I' is considered to be a signature of input, and u_n is the control input at time $n.\hat{x}_{n-1|n-1}$ and $P_{n-1|n-1}$ represents the prior prediction of the estimated value and covariance of estimated states, and C_n is the covariance of the processed noise at time n. [35]

Previous step is followed by an estimation correction step in which the value of the prediction is modified based on the incoming observations.

In the proposed architecture, the inputs to the Kalman Filter (KF) and Extended Kalman Filter (EKF), are direct and indirect measurements from different DS measurement resources, including D-PMUs, with possibilities of being contaminated with noise and bad data, as well as streaming and non-streaming measurements from smart meters, fault indicators, etc. The detailed explanation of the correction step and how to obtain A is elaborated in our previous publication [14], and we pre-

vent repeating those information for better exploitation of information within the current research work.

In order to tailor the method for systems with high levels of non-linearity, EKF with an iterative estimation is adopted resulting in removal of the linearity constraint:

$$\begin{aligned} x_k &= f(x_{k-1}, u_k) + w_k \\ z_k &= h(x_k) + v_k \end{aligned} \tag{5.3}$$

With w_k and v_k defined as noise and u is the control vector. Therefore:

$$\begin{aligned} \hat{y_k} &= z_k - h(\hat{x}_{k|k-1}) \\ S_k &= H_k P_{k|k-1} H_k^T + R_k \\ K_k &= P_{k|k-1} H_k^T S_K^{-1} \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k \hat{y_k} \\ P_{k|k} &= (I - K_k H_k) P_{k|k-1} \end{aligned} \tag{5.4}$$

Where $\hat{y}_k$ and S_k are residual values of measurement and covariance and $K, \hat{x}$, and P are the gain, state and covariance estimated values, respectively.

5.3.1 Synchronization and Ensemble Approach

The generated data from EKF and KF needs to be synchronized for $\mathbf{x}_s(t)$ using Maximum Likelihood Estimator technique at any time snapshot of t, to synchronize

the sampling rate and time stamp for both D-PMU and Relay measurements.

5.3.1.1 Maximum Likelihood Estimator (MLE)

The final step to generate the data fusion results in Ensemble Extended Kalman Filter (EEKF) is to integrate the results of the two data fusion base detectors (i.e. KF and EKF).

Given a set of X with random variables, $X = (X_1, X_2, ..., X_P)$, interpreted as the non-measured sensor features, with a probability distribution of $f(X|\theta)$, the likelihood function and the MLE will get defined as follows:

$$\begin{aligned} L(\theta|X) &= f(X|\theta) \\ \hat{\theta}(x) &= argmax L(\theta|X) \end{aligned} \tag{5.5}$$

Estimated values of $X = (X_1, X_2, ..., X_P)$ are used to fill-up the zero elements of $\mathbf{Q}(t) = [\mathbf{x}_1(t), ..., \mathbf{x}_s(t)...., \mathbf{x}_S(t)] \in \mathbb{R}^{14\times S}$, hence synchronous final EEKF generated output data will be achieved. Fig. 5.3 demonstrates the integration and synchronization of the input D-PMU and relay measurements as part of the developed EEKF.

5.4 Broader Level Outage Root Cause Classification

In this section, the effort is made towards creating an scheme for detection of the load outage event in a hierarchical manner, by detection of a broader level event in DS

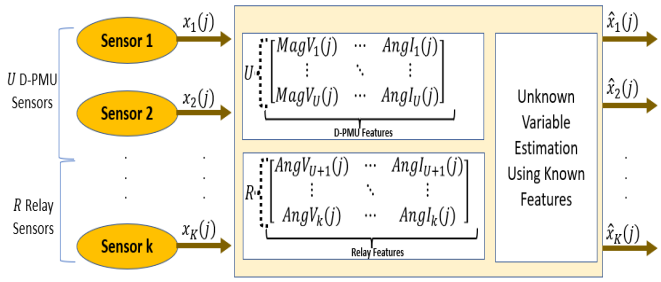


Figure 5.3: Integration of input measurements at time snapshot of t

and then sub-categorizing the classified disturbances with the criteria of either or not a load outage has occurred as a result of that event. The proposed outage detection scheme is able to classify the Faults, Breaker Operations, and False Data Injection (FDI) Cyber-attacks in the first step and then further analysis is done throughout the second step to identify if any of the aforementioned cyber-physical events do result in a load outage as well. The proposed framework includes Load Outage (LO) as a separate event within another event and this will result in creation of an Outage Detection dendrogram scheme as shown in Fig. 5.4.

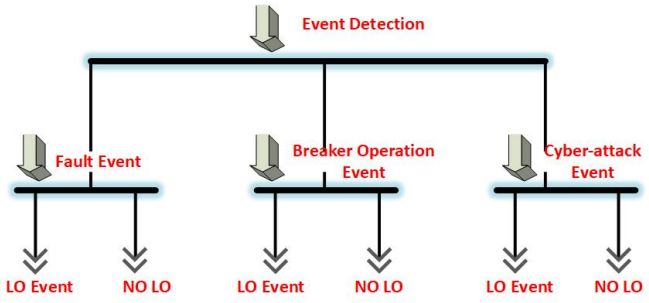


Figure 5.4: Dendrogram of the Outage Detection and Broader Level Root Cause Classification

5.4.1 Offline Phase

With the DS measurements being primarily unlabelled and dynamic nature of the system, resulting in insufficient available ground truth for the occurring disturbances, an unsupervised clustering approach is exploited in this work. The proposed outage detection approach is framing the clustering solution on a bottom-top fashion. Hence, a hierarchical clustering strategy is used based upon an agglomerative structure. An essential action in any type of clustering method is to calculate the distances among data points, which will impact the level of resemblance within clusters. Euclidean distances is chosen to be used as a distance metric in this work which best suits the multi dimentionality nature of the measurements in DS. Assuming K number of measuring points, agglomerative clustering starts with considering every measurement as a single cluster and then after calculating the similarity between the clusters, the two C_i and C_j clusters with the maximum similarity are to be combined to create a bigger cluster. This cluster merging process continues until the defined number of clusters are achieved. As far as the goals of this research work is concerned, an "Average Linkage" strategy is used to calculate the similarity between clusters. Given the fact that the load outage might be included as a sub-event category of any of the broader level events (i.e. Fault, Breaker Operation, and Cyber-attack), the "Average Linkage" technique considers all the distances among all pairs of the

clusters and calculates the average distance to merge. Assuming $\alpha_1, \alpha_2, ..., \alpha_n$, and $\beta_1, \beta_2, ...\beta_p$ are observations within clusters C_i and C_j, then the distance between the two clusters is calculated as follows:

$$D_ij = \frac{1}{np}\sum_{x=1}^{n}\sum_{y=1}^{p} D(\alpha_i, \beta_j)^2 \tag{5.6}$$

Where $D(\alpha, \beta)$ calculates the distances between the observation vectors.

5.4.1.1 Determining the Optimal Number of Clusters

The outputs of the EEKF consisting of the aggregated and cleaned measurements are primarily unknown, in terms of the number and type of the events as well as the labels. Therefore, a significant step in the proposed ORCA scheme is to find the optimal number of clusters, covering all the events of interest in this work as well as the LO sub-events. To that purpose, this work defines a window of clusters with various clustering numbers based on which a Silhouette score is assigned and the clustering results with the maximum score is chosen to be the final number of clusters for the entire dataset. The final score is expected to be a value between -1 and 1. The closer the value is to 1 has the implication of a better clustering efficiency. Efficiency of the two clusters is evaluated using the Silhouette score as follows:

$$Si_{score} = \frac{\sigma - \beta}{\max(\sigma, \beta)} \tag{5.7}$$

Where σ is the average distance between all the clusters and β is the average distance between the points within the same cluster.

As a rule of a clustering performance evaluation for the purpose of this work, the further apart the clusters are and the smaller the size of the final clusters become, a better clustering is conveyed.

5.4.2 Online Phase

This section expands the clustering method that was used for the offline phase and the goal is to make it customized for streaming input data, and tailor the methodologies for any possible change on any of the previously-formed clusters. The incoming data streams are being grouped in the window sets of 200 samples per window and the mean value for all the features are being calculated within the window.Assuming a d_{max} values as the maximum distance between variables in the current clusters, the distances for the incoming variable α_n is getting calculated from all the existing variables, $\alpha_1, \alpha_2, ..., \alpha_k$, in the clusters. With the $d_{n,max} < d_{max}$, no changes in the previous clusters is needed and the new variable is to be added to the least distance cluster. However, in case of the $d_{n,max} > d_{max}$, a new cluster will be created including α_n. In this case, a new calculation of the distances, between clusters and within clusters, needs to be done. Given the fact that this stage is processing the data streams with minimum amount of storage usage, the calculation of the distances are required to be done based on the stored data within the window size, as previous data are

not available. In this step, aim is towards possible re-clustering of the time series by finding the maximum distances to the new sets of clusters, for which, Hausdorff distance calculation strategy is used to calculate the variable similarities in a rapid manner using the following equations:

$$H(\alpha, \beta) = \max(h(\alpha, \beta), h(\beta, \alpha)) \tag{5.8}$$

Where h is defined as:

$$\begin{aligned} h(\alpha, \beta) &= \max(\min(i - j)) \\ h(\beta, \alpha) &= \max(\min(j - i)) \end{aligned} \tag{5.9}$$

In which case, i, and j are features of variables α, and β respectively.

5.5 Identification of the Detailed Outage Root Cause

In this section of ORCA, the results of the previous sections are to be used to find the major category of the outage and classifying it under three categories of Fault-induced, Breaker Operation-induced, and Cyber-attack-induced outages. For further analysis of the detailed outage root cause, access to tabular historical data, consisting different characteristics of the historical recorded outages is required. The data can contain a variety of features including the outaged phases, season, month, time of the day, temperature, weather conditions at the time of the outage, and etc. For this section, these data are being processed using an unsupervised Association

Rule Mining method. This analysis is programmed to be done in an offline manner and is appended to the offline phase of the hierarchical clustering that was presented in the previous section.

5.5.1 Outage Hypothesis Generation and Feature Selection

The broader level classification of the outage Root Cause (RC) can be primarily narrowed down to a series of further accurate potential root causes. This step is to create a scenario of possible detailed RC's for each of the classifications that has been accomplished in the previous sections, and can make the supply restoration and outage recovery process to be faster and more feasible. Table. 5.1 presents a list of hypothetical nuanced outage main causes and sub-causes that needs to be mapped to any of the previously-detected broader level classifications in the next step:

The methodologies in this research work has the capability of processing a variety of outage features, based on the availability. However, the most effective characteristics of the outages and the proposed features are as follows:

- Outage Phase
- Month
- Time of the Day

- Activated Devices
- Weather Condition
- Temperature
- OH or UG
- Wildlife Proximity
- Level of Plant Coverage

Table 5.1: Outage Root Cause Hypothesis

Main Root Cause	Sub Root Cause
Vegetation-related	• Above ground touch • Root of the tree • Water • Etc.
Weather-related	• Excessive heat • Flood • Hurricane • Etc.
Animal-related	• Wildlife • Bird
Wildfire	• Transformer outage • Transmission-induced
Protective Devices	• Relay • Transformer fuse • Line recloser • HPP/MHP breaker • Transformer CSP • Cyber-attack • Line fuse • Etc.
Equipment Failure	• Physical • Communication • Cyber-attack
Planned Outages	• Local • Transmission-based

Moreover, all the above features need to be cleaned and converted into categorical meaningful values, based upon which, the data mining method can be applied.

5.5.2 Rule Mining Using Frequent Pattern-Growth (FP-Growth) Unsupervised Learning

preparation of the outage historical data, with regard to the aforementioned features mentioned in the previous section, paves the way towards the ORCA unsupervised data mining with the goal of identification of the detailed outage root cause. This section of the work gets use of an Association Rule Mining technique called Frequent Pattern-Growth (FP-Growth). This is a tree-based data mining strategy which, in ORCA, is tailored for the DS outage data with domain-based adjustments. The proposed solution approach, redefines the problem of the RCA as an itemset mining problem. While the goal remains on finding the most accurate model predicting the outage root cause, the key step is narrowed down to find the frequent itemsets in terms of associated rules between the measurements and mapping of the extracted rules into known root causes. As far as the rule mining aspect of the proposed ORCA is concerned, several base detectors, that have been vastly developed and used in literature, are tested and evaluated for potential beneficial results in the scope of DS Outage RCA.Considering such domain criterias as the level of availability of the data, computational time, and memory usage, Table. 5.2 brings a comparison between the

two mostly-used, methodologies in the area of rule mining. The itemset mining technique that is used in this work consist of two primary actions of first, creation of the tree framework based on all the prepared items which is called FP-tree, and second, to apply the FP-growth formula to mine the leaves of the tree and draw out all the itemsets that are being correlated, defined by a certain threshold. The constructed tree is created as a result of scanning all the database to make a decision on the items that has the most frequency of occurrence, which is called Support,S_{α_i}, assuming α being an item in the dataset. Furthermore, a secound round of scanning is done over the all the items, $_1, \alpha_2, ..., \alpha_n$, and the tree is being constructed. The higher the value of S_{α_i}, node α_i will have a higher chance of being combined with other α's in the tree, which in turn, has the implication of α_i,having a more chance of being in an itemset with other items. In the context of root cause analysis the interpretation is if α_j is encountered, as a single feature of an outage, most likely α_i is also the case, which will help to add more corresponding features to a same variable, and in this case outage root cause.

Furthermore, due to the hierarchical underlying nature of the proposed ORCA, outage events are continuously being scanned in the EEKF output data. Bottom layer of the hierarchy, consisting of the measurements at multiple locations, are considered separate events, inherently embedding the capability of multiple outage detection at the same time in ORCA.

Table 5.2: FP-Growth vs Apriori

	FP-Growth	**Apriori**
Memory Usage	Small	Large (stores all the itemsets)
Candidate Generation	Does not Exist (tree structure)	Exist
Speed	Faster for large datasets	Speed exponentially decreases with data size
Patterns Finding	Conditional trees are being mined	Threshold on the minimum support
Number of Scans	Two	Several (depending on the data size)

Algorithm 3: Outage Data Mining Algorithm

Input: Outage Cleaned Data with Categorical Features $\alpha_{1,2,\ldots,n} \in \mathcal{O}_{1,2,\ldots,k}$;

find S_α for $\alpha_{1,2,\ldots,n}$;

for $(t = 1$ to $k; t++)$:

  for $(j = 1$ to k and $i = 1$ to n $; j++, i++)$:

    find Max $S_{\alpha_{1,2,\ldots,n}} \in \mathcal{O}_{1,2,\ldots k}$;

    List the Ascending order of α's $\in \mathcal{O}$'s;

    if $\alpha_i \in$ any other $\mathcal{O}$:;

      $S_{\alpha_i}++$;

  if $\alpha_i < S_{min}$;

    remove α_i ;

  else ;

    $\overline{U_t} \leftarrow \alpha_i$;

Return $\overline{U}$

5.6 Testbed and Validation

5.6.1 Test Systems

For the validation of different phases of the proposed ORCA, two separate testcases have been considered and various scenarios are developed for each of the testcases. Both cases are simulated in the OPAL-RT hardware-in-the-loop simulator and actual harware D-PMU is added to the loop of the simulation, for the purpose of the "Online Phase" validation.

Hypersim, OPAL-RT interface, sample 24-node feeder test system, is considered as the first testcase, and is modified by adding DERs and connection of D-PMUs. The one-line diagram of the system is as shown in Fig. 5.5.

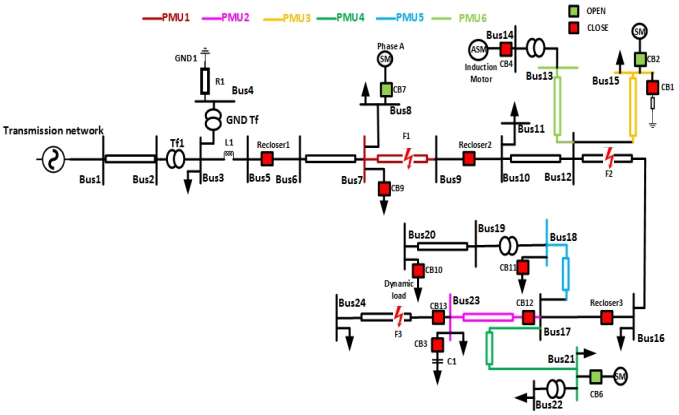


Figure 5.5: 1-line diagram of the testcase1

As the second test system, the IEEE 33-node modified feeder is developed as

shown in Fig. 5.6.

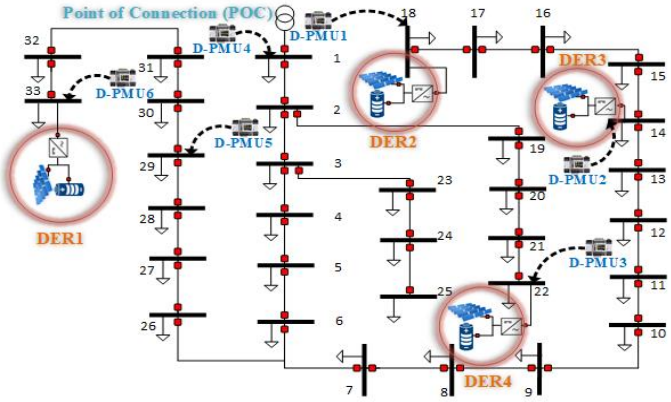


Figure 5.6: 1-line diagram of the testcase2

Four operating scenarios are considered for the test systems as follows:

- Scenario 1: Fault and Breaker Operation Events
- Scenario 2: Fault and Cyber-attack Events
- Scenario 3: Cyber-attack and Breaker Operation Events
- Scenario 4: Fault, Breaker Operation, and Cyber-attack Events

5.6.2 Evaluation Results

For the validation purposes four terms are defined as— a) True Positives (TP) are positive instances correctly classified, b) False Positives (FP) are negative instances classified as positive, c) True Negatives (TN) are negative instances correctly

classified as non-positive, d) False Negatives (FN) are positive instances classified as negative. Therefore, the final performance of the detection is defined as follows:

$$RI = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.10}$$

Furthemore, Precision is defined as the ratio of TP to the sum of TP and FP. A recall is defined as the ratio of TP to the sum of TP and FN. Average precision metric is used to compare the performance of the proposed EEKF, with respect to the conventional KF and EKF, as shown in Fig.5.8 . Moreover, various number of clusters have been tested and the Silhouette score is calculated. The number of clusters that achieves the highest score, will be selected for the next step. Fig.5.7 shows the scores for a range of $K = [2, 3, 4, 5, 6]$, and $K = 3$ is selected. This is for Scenario 1 of the first test system and a similar procedure is done for all the other scenarios as well.

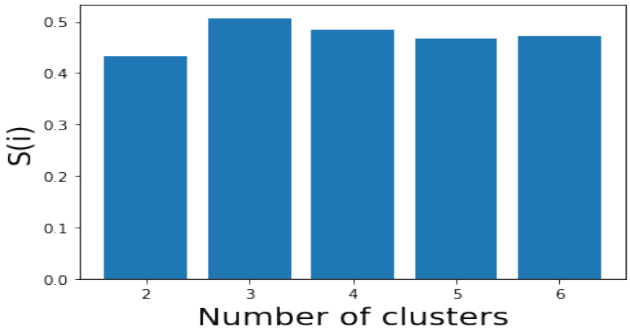


Figure 5.7: Choosing the Optimal Number of Clusters

To evaluate the detection performance, Rand Index (RI) is used. A total number of n outage variables present in the set O, $O = \{o_1, o_2, ..., o_n\}$, including subsets of

$A = \{A_1, A_2, ..., A_p\}$, and $B = \{B_1, B_2, ..., B_s\}$, with each A_i and B_j being a vector of o's. As a comparative metric to evaluate the performance of the proposed detection to the existing methods, harmonic mean of precision and recall is used as F_1score. The results are summarized in the following Table 5.3.

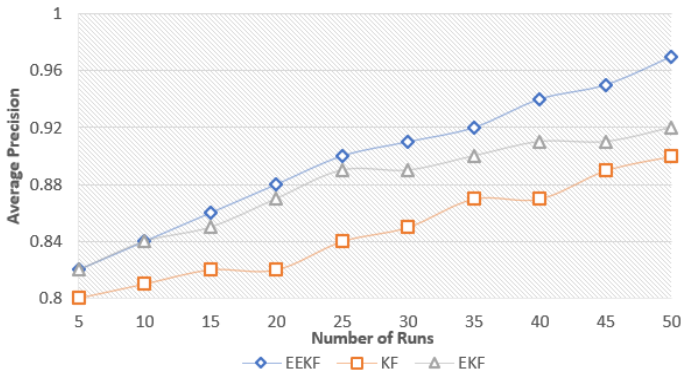


Figure 5.8: Performance Comparison of the Proposed EEKF with EKF and KF

Table 5.3: Detection Evaluation Results

Scenario	Test System	# Clusters	Rand Index	$F_1 Score$
1	1	5	0.9	98.6
2	1	5	0.9	98.1
3	1	5	0.9	98.9
4	1	7	0.8	90.3
1	2	5	0.9	97.9
2	2	5	0.9	96.6
3	2	5	0.9	96.2
4	2	7	0.7	87.1

Scenario 1 of the test system 2 is considered for a comparison purposes and other methods are used to see how the outage detection performance of ORCA, outweights the other methods. Table 5.4 shows the comparative results, with respect to F_1Score:

Table 5.4: Comparison of ORCA vs Other Methods for Outage Detection

Method	$F_1 Score$
Decision Tree	83.5
Support Vector Machine	85.7
Principal Component Analysis	70.1
ORCA	**97.9**

The next step of the evaluation of ORCA, is to examine the performance of the data mining section for a identification of the outage Detailed Root Cause (DRC). In this section we have used artificial data and have converted all the features to categorical values with labels. The final outage table consists of the values for each of the converted features, based on which, the mining method is implemented. Table 5.5 shows the results of the data mining on the artificial data, as well as the associated feature labels. In this table, the metric Confidence is defined as the ratio of the total number of a certain DRC associating with specified features over the total number of outages by DRC. In this table features are mentioned with their associated labels, for instance DA means Device Activated, and other features are chosen in a way to be self explanatory.

Table 5.5: ORCA: Mining of the Nuanced Outage Root Cause

Features	DRC	Confidence
{WEATHER,DA,DEW}	Vegetation	0.9208
{WEATHER,HUMIDITY}	Animal	0.8438
{MONTH,WEATHER,TEMP}	Animal	0.8961
{WIND}	Vegetation	0.7098
{MONTH,TEMP,HUMIDITY}	Wildfire	0.9745
{WIND,WEATHER}	Vegetation	0.9127
{DEW,HUMIDITY}	Equipment	0.9355
{MONTH,TIME,TEMP,DA}	Animal	0.9426
{DEW}	Vegetation	0.7422
{DA,HUMIDITY}	Equipment	0.9003
{MONTH,TEMP}	Wildfire	0.9249
{WEATHER,TEMP,HUMIDITY,WIND}	Equipment	0.9736

5.7 Summary

In this work, set of algorithms have been developed as a sub-modules for ORCA: Outage Root Cause Analysis in DER-rich Electric Distribution System. Variety of outages in Electric Distribution Systems (EDS) are analyzed including large-scale load interruptions for the root causes. ORCA is developed using data-driven and physics-based approaches, containing various modules on data fusion, outage detection, broader level root cause classification, and identification of the nuanced outage root cause. Ensemble Extended Kalman Filter (EEKF) is used to aggregate and fuse data from diverse set of distribution sensors. Hierarchical clustering is used for broader level classification of root causes and an unsupervised frequent pattern-growth algorithm is developed assisted by hypothesis generation and feature selection for detailed outage root cause identification. The detection stage is tailored for offline

and online applications and has the capability of detecting outages using streaming D-PMU data. The developed algorithm is validated using test systems in OPAL-RT real-time simulator, as well as synthetic but realistic data for evaluation of the different developed modules. Results demonstrate relative advantage of the developed approach compared to state-of-the-art approaches. Future work will include validation with utility data and implementation.

REFERENCES

[1] M. S. Bashkari, A. Sami, M. Rastegar, and M. J. Bordbari, "Distribution power system outage diagnosis based on root cause analysis(invited paper)," *Scientia Iranica*, vol. 26, pp. 3672–3680, 2019.

[2] U. J. Minnaar, F. Nicolls, and C. T. Gaunt, "Automating transmission-line fault root cause analysis," *IEEE Transactions on Power Delivery*, vol. 31, no. 4, pp. 1692–1700, 2016.

[3] A. N. Samudrala, M. H. Amini, S. Kar, and R. S. Blum, "Distributed outage detection in power distribution networks," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5124–5137, 2020.

[4] Y. Cai, M.-Y. Chow, W. Lu, and L. Li, "Statistical feature selection from massive data in distribution fault diagnosis," *IEEE Transactions on Power Systems*, vol. 25, no. 2, pp. 642–648, 2010.

[5] A. Gholami, A. K. Srivastava, and S. Pandey, "Data-driven failure diagnosis in transmission protection system with multiple events and data anomalies," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 4, pp. 767–778, 2019.

[6] Q. Cui, S. M. Y. Hashmy, Y. Weng, and M. Dyer, "Reinforcement learning

based recloser control for distribution cables with degraded insulation level," *IEEE Transactions on Power Delivery*, vol. 36, no. 2, pp. 1118–1127, 2021.

[7] A. Ashok, M. Govindarasu, and V. Ajjarapu, "Online detection of stealthy false data injection attacks in power system state estimation," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 1636–1646, 2016.

[8] N. V. Mago, S. Santoso, and M. F. McGranaghan, "Assessment of feeder voltage regulation using statistical process control methods," *IEEE Transactions on Power Delivery*, vol. 23, no. 1, pp. 380–388, 2008.

[9] S. Som, R. Dutta, A. Gholami, A. K. Srivastava, S. Chakrabarti, and S. R. Sahoo, "Dpmu-based multiple event detection in a microgrid considering measurement anomalies," *Applied Energy*, vol. 308, p. 118269, 2022.

[10] S. Qian and C.-A. Chou, "A koopman-operator-theoretical approach for anomaly recognition and detection of multi-variate eeg system," *Biomedical Signal Processing and Control*, vol. 69, p. 102911, 2021.

[11] B. Alnajjab, A. N. Samudrala, C. Chen, R. S. Blum, S. Kar, and E. M. Stewart, "Outage detection for distribution networks using limited number of power flow measurements," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 2, pp. 315–324, 2020.

[12] R. Moxley and D. Dolezilek, "Case studies: Synchrophasors for wide-area monitoring, protection, and control," in *2011 2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies*, pp. 1–7, IEEE, 2011.

[13] M. Doostan and B. H. Chowdhury, "Power distribution system equipment failure identification using machine learning algorithms," in *2017 IEEE Power & Energy Society General Meeting*, pp. 1–5, IEEE, 2017.

[14] A. Gholami, A. K. Vosughi, and A. K. Srivastava, "Denoising and detection of bad data in distribution phasor measurements using filtering, clustering and koopman mode analysis," *IEEE Transactions on Industry Applications*, pp. 1–1, 2022.

[15] F. Agostinelli, M. Hoffman, P. Sadowski, and P. Baldi, "Learning activation functions to improve deep neural networks," *arXiv preprint arXiv:1412.6830*, 2014.

[16] Y. Zhou, R. Arghandeh, and C. J. Spanos, "Partial knowledge data-driven event detection for power distribution networks," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5152–5162, 2018.

[17] A. Vosughi, A. Gholami, and A. K. Srivastava, "Denoising and bad data detec-

tion in distribution phasor measurements using filtering, clustering and koopman mode analysis," in *2021 IAS Annual Meeting*, 2021.

[18] X. Jiang, Y. C. Chen, V. V. Veeravalli, and A. D. Domínguez-García, "Quickest line outage detection and identification: Measurement placement and system partitioning," in *2017 North American Power Symposium (NAPS)*, pp. 1–6, 2017.

[19] X. Jiang, Y. C. Chen, V. V. Veeravalli, and A. D. Domínguez-García, "Quickest line outage detection and identification: Measurement placement and system partitioning," in *2017 North American Power Symposium (NAPS)*, pp. 1–6, 2017.

[20] A. N. Samudrala, M. H. Amini, S. Kar, and R. S. Blum, "Sensor placement for outage identifiability in power distribution networks," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 1996–2013, 2020.

[21] C.-W. Ten, J. Hong, and C.-C. Liu, "Anomaly detection for cybersecurity of the substations," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 865–873, 2011.

[22] A. Gholami, M. Mousavi, A. K. Srivastava, and A. Mehrizi-Sani, "Cyber-physical vulnerability and security analysis of power grid with hvdc line," in *2019 North American Power Symposium (NAPS)*, pp. 1–6, 2019.

[23] J. Valenzuela, J. Wang, and N. Bissinger, "Real-time intrusion detection in power system operations," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1052–1062, 2012.

[24] M. Wu and L. Xie, "Online detection of low-quality synchrophasor measurements: A data-driven approach," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2817–2827, 2016.

[25] N. M. Manousakis and G. N. Korres, "Optimal pmu placement for numerical observability considering fixed channel capacity—a semidefinite programming approach," *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 3328–3329, 2016.

[26] M. Jamei, A. Scaglione, C. Roberts, E. Stewart, S. Peisert, C. McParland, and A. McEachern, "Anomaly detection using optimally placed ¡inline-formula¿¡tex-math notation="latex"¿μPMU ¡/tex-math¿¡/inline-formula¿ sensors in distribution grids," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 3611–3623, 2018.

[27] Y. Zhao, R. Sevlian, R. Rajagopal, A. Goldsmith, and H. V. Poor, "Outage detection in power distribution networks with optimally-deployed power flow sensors," in *2013 IEEE Power & Energy Society General Meeting*, pp. 1–5, IEEE, 2013.

[28] M.-y. Chow and L. S. Taylor, "Analysis and prevention of animal-caused faults in power distribution systems," *IEEE Transactions on Power delivery*, vol. 10, no. 2, pp. 995–1001, 1995.

[29] D. T. Radmer, P. A. Kuntz, R. D. Christie, S. S. Venkata, and R. H. Fletcher, "Predicting vegetation-related failure rates for overhead distribution feeders," *IEEE Transactions on Power Delivery*, vol. 17, no. 4, pp. 1170–1175, 2002.

[30] M. Doostan, R. Sohrabi, and B. Chowdhury, "A data-driven approach for predicting vegetation-related outages in power distribution systems," *ArXiv*, vol. abs/1807.06180, 2018.

[31] L. Xu, M.-Y. Chow, and L. S. Taylor, "Data mining and analysis of tree-caused faults in power distribution systems," in *2006 IEEE PES Power Systems Conference and Exposition*, pp. 1221–1227, IEEE, 2006.

[32] Y. Cai and M. Chow, "Cause-effect modeling and spatial-temporal simulation of power distribution fault events," *IEEE Transactions on Power Systems*, vol. 26, pp. 794–801, 2011.

[33] R. Eskandarpour and A. Khodaei, "Machine learning based power grid outage prediction in response to extreme events," *IEEE Transactions on Power Systems*, vol. 32, pp. 3315–3316, 2017.

[34] A. Gholami and A. K. Srivastava, "Comparative analysis of ml techniques for data-driven anomaly detection, classification and localization in distribution system," in *2020 52nd North American Power Symposium (NAPS)*, pp. 1–6, 2021.

[35] A. Vosughi, A. Gholami, and A. K. Srivastava, "Denoising and bad data detection in distribution phasor measurements using filtering, clustering and koopman mode analysis," IAS Annual Meeting 2021, Oct 2021.

CHAPTER 6. CONCLUSIONS, CONTRIBUTIONS, AND FUTURE WORK

Considering the massive data availability from sensor deployment within the distribution system, computational techniques need to extract information from distribution phasor measurement units and smart meters. Number of ML techniques have been discussed in literature for abnormal operation detection, anomaly classification, and anomaly localization. This research provides an overview on commonly-used machine learning techniques with a comparative performance analysis. It can be seen from the literature that supervised learning has been more frequently used, specifically ANN for event detection and classification. Also, K-Nearest Neighbor has seemed to be more attractive for the localization purposes. More detailed analysis are needed for the feature selection and classifications for superior performance.

Contributions: The following points summarize the contributions of each developed modules in this work:

- Enhancement of the real-time situational awareness in distribution system through measurement data analytics and integration with applications
- Development of multiple novel architectures for multi-sensor measurement data-

in-data-out fusion and sensor resolution enhancement with introducing EEKF

- Data synchronization and denoising with various statistical and clustering-based techniques
- Detection of bad data followed by mitigation, with an enhanced State Estimation using Weighted Least Square (WLS) method tailored for system with high noise and low observability
- Cyber-physical DS Event analysis for automated real-time detection of events with broader level and detailed level of classification with high accuracy
- Incorporation of D-PMU spatio temporal dependencies through hybrid Ensemble method with location identification
- A novel ORCA scheme to comprehensively identify, locate, and root cause of load outages within the distribution network
- a novel rule-mining technique based on the frequent pattern growth, tailored for outage historical log data with capability of industrial implementation

Research Challenges: Open challenges spans a wide range of domain-nature complexities as well as technique-based limitations. Apart from the large-scale nature of the power system streaming measurements, the highly dynamic behavior of the system makes the anomaly classification and localization a very challenging task.

Furthermore, high requirement on huge labeled dataset of event libraries, need for online training and forecasting, non-stationary operation of the system with presence of concept drift, and massive spatio-temporal dependency of measurements are among the domain challenges which are to be well-recognized and mitigated to come up with an efficient solution approach. Below points summarizes the existing challenges either in the solution approach or inherit-in technical power system challenges:

- huge amount of required training dataset, while limited available labeled domain data
- non-stationary operation causing the event signatures to be dynamically changing over the course of a daily operation
- interlinked nature of system components and high system measurements correlation
- non-linear system behavior with high computational complexity and real-time requirements
- requirement of human intervention for tuning purposes based on the system operating conditions
- setting threshold in ML technique for detection while considering concept drift in anomaly definition over dynamic system operation

Future Research Needs: Anomaly Detection, Classification, and Localization (AD-C-L) as well as root cause analysis is still in fancy and offers challenges and research opportunities for ML/ Data Science researchers. Although there is a wide range of distribution system anomalies that are likely to happen in a daily operation of the system, resulting in different system dynamic behavior, in the context of the anomaly analysis, application in the literature are mainly limited to fault events and power quality events, and the deficiency on analysis of the large-scale uPMU data streams in an online time frame, still remains to be an open area of research work. The future researches in the area of distribution system anomaly detection, classification, and localization includes, but not gets limited to, the following points:

- library inclusion of real-field huge supervised labeled datasets
- integration of domain knowledge with ML approach in terms of a physics-based analysis
- automation of the anomaly detection, classification, and localization with a single-step approach for time-sensitive applications
- performance evaluation using online real-field measurements or Hardware-in-the-loop with actual uPMUs to account for the noises and phasor estimator specifics
- differentiating anomalies caused by cyber event, physical events, human cause

events, sensor data and asset deterioration is a real challenge, specially given data coming from biggest man-made dynamic machine infrastructure ever.