2019-01-01

Artificial Intelligence in the Assessment of Transmission and Distribution Systems Under Natural Disasters Using Machine Learning and Deep Learning Techniques in a Knowledge Discovery Framework

Rossana Villegas
University of Texas at El Paso

ARTIFICIAL INTELLIGENCE IN THE ASSESSMENT OF TRANSMISSION AND DISTRIBUTION SYSTEMS UNDER NATURAL DISASTERS: USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES IN A KNOWLEDGE DISCOVERY FRAMEWORK

ROSSANA VILLEGAS ARGUELLES

Doctoral Program in Electrical and Computer Engineering

APPROVED:

Patricia A. Nava, Ph.D., Chair

Virgilio E. Gonzalez, Ph.D.

John A. Moya, Ph.D.

Jose F. Espiritu Nolasco, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

Dedication

with love for Carla, Ricardo, and Jorge, my beloved children. My mother Josefina, and specially in loving memory of my father, Marcelo. My friend and personal Google Monica, and my constant life counselor.

ARTIFICIAL INTELLIGENCE IN THE ASSESSMENT OF TRANSMISSION
AND DISTRIBUTION SYSTEMS UNDER NATURAL DISASTERS: USING
MACHINE LEARNING AND DEEP LEARNING TECHNIQUES IN A
KNOWLEDGE DISCOVERY FRAMEWORK

by

ROSSANA VILLEGAS ARGUELLES, B.S., M.I.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Department of Electrical and Computer Engineering

THE UNIVERSITY OF TEXAS AT EL PASO

December 2019

ACKNOWLEDGEMENTS

ABSTRACT

Warming trends and increasing temperatures have been observed and reported by federal agencies, such as the National Oceanic and Atmospheric Administration (NOAA). Extreme-weather events, especially hurricanes, tornadoes and winter storms, are among the highly devastating natural disasters responsible for massive and prolonged power outages in Electrical Transmission and Distribution Systems (ETDS). Moreover, the failure rate probability of any system component under extreme-weather tends to increase in the impacted geographic area. This dissertation proposes an Artificial Intelligence (AI) Decision Support System that can predict damage in the ETDS and allow operators to mitigate disastrous extreme weather events. The document reports the results of the exploration of a novel method to integrate two main domains: the critical operation of the ETDS under natural disaster conditions; and data integration based on the sequence of steps in a Knowledge Discovery Framework (KDF). Machine Learning and Deep Learning approaches, including the spectrum of data mining, are incorporated in the KDF and used to perform the estimation, regression, and classification tasks. By means of two scenarios, a winter storm and a major hurricane, the proof of concept of the consolidation of the two domains, AI and ETDS, is demonstrated. The results of the methods are compared, as well as techniques and accuracy of the algorithms. Discussion includes descriptive statistics of the data analysis, conducted to understand each data set, and how they are related to each task. The results reveal a powerful tool, that incorporates disparate ideas and data, and increases the accuracy of predictions and classifications of extreme weather damage in the hypothetical cases presented. This is of importance to the operator decision support in order to solve problems in the area of critical operation of the Transmission and Distribution systems during extreme-weather events.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

This chapter introduces the problem of enhancing the resilience of the Transmission and Distribution Systems under natural disasters. The rationale of the study, the research objectives and the expected outputs of this study are presented in this chapter.

1.1 Natural disasters and their effects

Climate change is increasing the risk of climate-related natural disasters, as stated by [Muna13]. Important indicators include the increment of global temperature, changing patterns in precipitation, storms, and extreme temperatures as reported by Kausky [Kaus14] and Mercado [Merc16]. The increment in the frequency and intensity of extreme weather events has broadened floods, droughts, wildfires and extreme temperatures in the world since 1940, as shown in figure 1.1. These events have affected 232 million people, and caused more than $100 billion dollars in damage worldwide between 2001 and 2010, as Brown [Brow18] states.
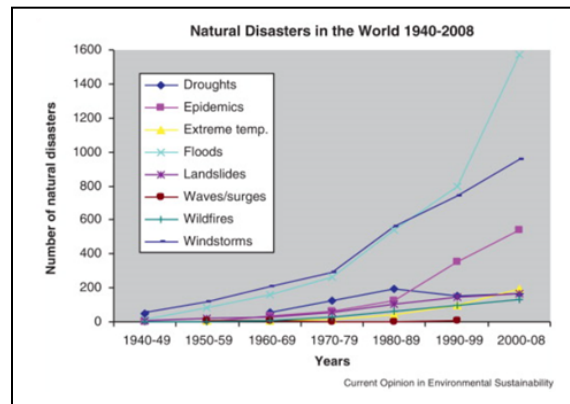
Figure 1.1: Natural disaster in the world since 1940 [Muna13].

1.2 Natural Disasters in the United States

In the past 10 years, up to the time of this study, the following hurricanes have impacted the United States: Ike (category 2-4, September 2008) hit mostly Cuba, Florida, Texas, Louisiana and Arkansas reported by Berg [Berg08], while hurricane Katrina (category 3-5, August 2005) hit

Florida, Louisiana, Mississippi, and Alabama reported by Knab et al. [Knab06], and hurricane Rita (category 3-5, September 2005) hit Arkansas, Florida, Louisiana, Mississippi, and Texas. These hurricanes, collectively, devastated much of the U.S. Atlantic coast from Florida to New England, as reported by Knab et al. [Knab06]. Other catastrophic events include hurricane Charley (category 4, August 2004), which hit Florida, South Carolina, and North Carolina, as reported by Pasch et al. [Pasc11], and hurricane Ivan (category 3, September 2004) which hit Alabama, Florida, Louisiana, and Texas, as reported by Baker [Bake17]. Recently, Hurricane Harvey (category 3, August 2017) and Hurricane Irma (category 5, September 2017), with extremely heavy winds, plowed through the Caribbean. The wind smashed buildings, and downed trees and power lines, as reported by Blake et al. [Blak18] and Masters [Mast16]. Moser et al. [Mose14] states that these disaster events underscore the vulnerability of Caribbean countries and coastal cities. Figure 1.2 shows the 16 billion-dollar weather and climate disasters that impacted the United States (U.S.) during 2017, as reported by NOAA [Noaa17].
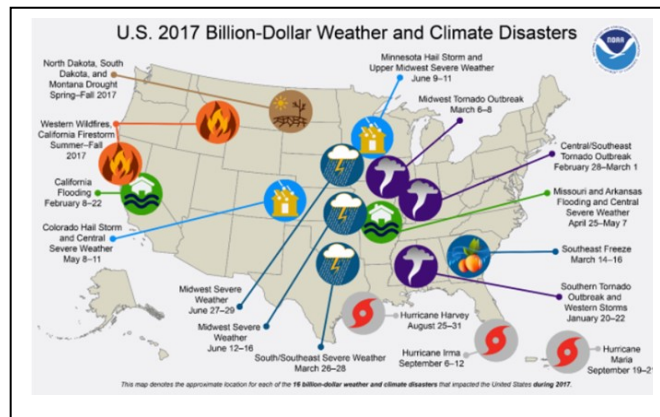
Figure 1.2: Climate disasters impacted the U.S. during 2017 [Noaa17].

The most current reports, at the time of this writing, indicate that the U.S. (in 2018) has suffered billion-dollar and climate disasters, as highlighted in the map shown in figure 1.3. As reported by the NOAA [Noaa18], severe weather is the major cause of disasters during the first months of 2018, with a toll estimated at 36 Deaths and $7.1 Billion in estimated costs. According

to a NOAA research group, these threats could be continued for the rest of the year and years to come, as climate change continues to develop, as stated in section 1.1.
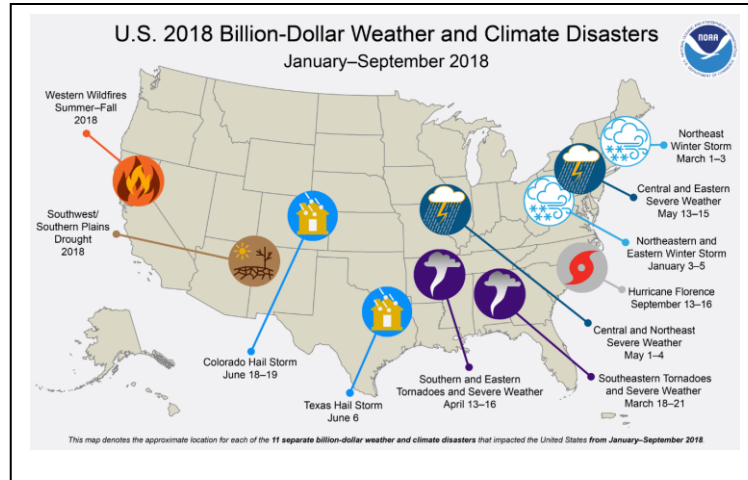
Figure 1.3: Natural disaster impacted in the U.S. during 2018 [Noaa18].

1.3 Electric power systems

The electric power systems or the Electricity Transmission and Distribution System, are generally categorized into three segments: generation, transmission, and distribution. Between each division there exists a crucial link, the electric substation.

1.3.1 ELECTRIC GENERATION SYSTEMS

There are many different ways to generate electricity in a central generation station, or by island modes, such as hydro, coal, oil, renewable, nuclear, and gas turbine. Furthermore, generation can involve a mix of synchronous and asynchronous machines, whose behavior could potentially create a source of power energy disturbance, such as short-circuit faults. Tleis [Tlei18] states that the most common short-circuit faults are weather related, followed by equipment failure. Figure 1.4 shows the existing types of electricity production in the United States through 2018, reported by the U.S. Energy Information Administration (USEIA) [Usei18a]. The power energy from the generators is delivered to the transmission systems with the help of a step-up transformer.

3
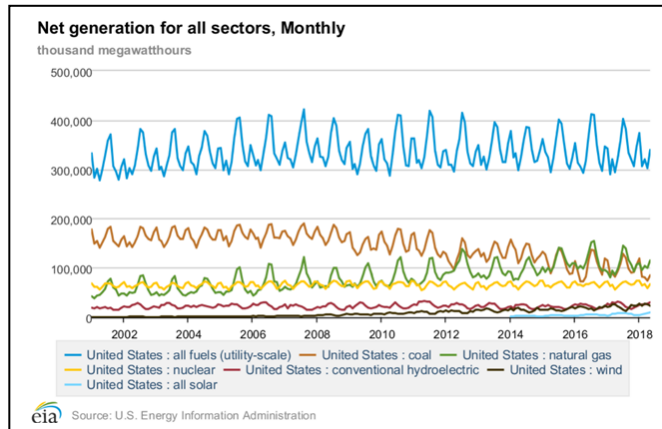
Figure 1.4: United States electricity production in 2018 [Usei18a].

1.3.2 ELECTRIC TRANSMISSION SYSTEMS

Electric power transmission is the bulk transfer of power by high-voltage connections between generation and load substations, to ultimately supply power to end users. Figure 1.5 shows the actual transmission network in the United States at the time of this research study, reported by the USEIA [Usei18b]. Real-time monitoring, with communication technology, is based on Phasor Measurement Units (PMUs) and state estimator sensors, located at key points to remotely monitor where and when the power might go out as studied by Momoh [Momo12].

The main objective of electric power transmission is the ability to transfer energy, reliably and efficiently, between generation sources and distribution load points. Any causes of disruption, like power outage, could alter the transmission system operation. One of the primary causes of transmission line outages is extreme weather conditions, and this is directly dependent on the geographic location of the transmission line as studied by Shen et al. [Shen98].
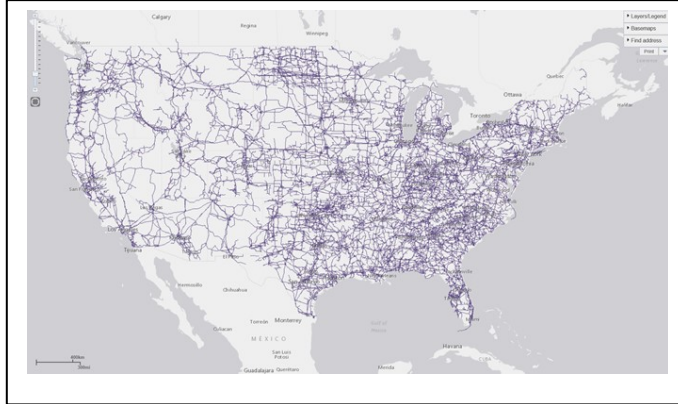
Figure 1.5: United States energy power transmission systems [Usei18b].

1.3.3 ELECTRIC DISTRIBUTION SYSTEMS

The Distribution System is the last step in the transmission of energy to end users. This system is similar in structure to the transmission system, but covers a smaller geographical section. The effect of weather conditions, like icing, is one potential cause of energy fault in the distribution systems, because of the effect of the expansion or contraction of the conductor. Additionally, other factors like wind and ice loading on overhead distribution lines lead to potential faults. The timely action in finding the location of a fault is crucial to minimize the interruption time in Distribution Systems as studied by Feizifar et al. [Feiz13].

1.3.4 ELECTRIC SUBSTATIONS

The Substations are an important part of the power energy system, as the main objective of the substation is to dispatch electric power from generation stations delivering through transmission lines, to the distribution system by step-up and step-down transformers. In addition, the substation contains the most important equipment in power energy systems, which typically are: Transformers, Regulators, Circuit Breakers, Air-Break, Disconnect Switches, Switchboard, Measurement Instruments, Relays and Bus Bars. The operation of a substation, could be carried out by a local operator, or remotely by an automatic control center system, named Supervisory Control and Data Acquisition Systems (SCADA) [Ahme08].

The recent deployment of distributed intelligent devices and Phasor Measurement Systems (PMUs) as studied by Yu [Yufb17] and Nakafuji [Naka17], are enabling system stability by detecting and isolating a disturbance caused by different events such as a short-circuit, equipment failure, temporary damage, or an extreme-weather generated faults. Moreover, these devices generate useful data for wide-areas distribution analysis.

1.4 Power faults classifications

A power energy fault is the unintentional and unwanted creation of a conducting path (short circuit) or an open circuit. It exists as four types of power energy faults: single line-to-ground, line-to-line, double line-to-ground, and balanced three-phase (details of this faults are not presented in this document for purposes of simplicity). On the other hand, the causes of power energy faults can be classified in two broad categories: typical and extreme-weather related [Egbu16]. Both types of power faults are briefly detailed below to give a point of comparison, but only the extreme-weather related power fault will be considered in this study.

1.4.1 TYPICAL POWER FAULT (TPF)

The components of the power energy system could exhibit failures during normal operation of the ETDS, which could happen at any time. As stated by Arabali et al. [Arab16], the most common causes are trees falling onto power lines, wind damage, lightning, line breaks due to excessive ice loading, vandalism, birds shorting out lines, vehicles' collision with towers or poles, etc. Furthermore, authors Kuma et al. [Kuma17] and Demazy et al. [Dema17] affirm that with the recent modernization in the distribution and generation system by the integration of solar and wind energy into the power grid, and the addition of new technologies like Plug-in Hybrid Vehicles (PHEVs), the power demand is increasing, thus causing more energy outages that are more complex to address.

1.4.2 EXTREME-WEATHER POWER FAULT (EPF)

According to the U.S. Energy Information Administration [Usei18b], *Natural Disaster* incidents are hierarchical in number of occurrences: windstorm (hurricane, severe storm), thunderstorm (tornado, lightings), winter storm (ice, snow), temperature extreme, wildfire, and other undefined weather. Moreover, Egbue et al. [Egbu16] sustain that recent earthquakes and tsunamis are increasing and are now included in the list. According to the U.S. Department of Homeland Security FEMA [Fema18] and Kenward et al. [Kenw14], **extreme-weather**, caused by windstorm and thunderstorm, are among the most destructive causes of sustained power faults and massive damage to the ETDS infrastructure. This is also reported by NOAA and shown in Figure 1.6 [Noaa18b].
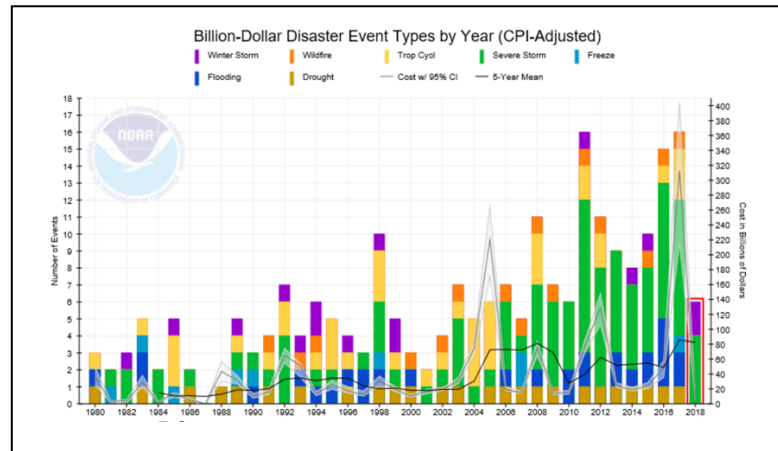
Figure 1.6: Billion-dollar weather and climate disasters: Time series [Noaa18b].

1.5 Electric system modernization and data analytics modeling

Modernization in the ETDS allows the generation and acquisition of big data, by means of new technology devices like Remote Transmission Units (RTUs), Phasor Measurement Units (PMUs) and smart meters in the Advance Metering Infrastructure (AMI). Supervisory Control and Data Acquisition (SCADA) is still being used extensively in the ETDS with upgraded options,

such as the modernization in the new SCADA system architecture, which provides remote communication via circuit breakers and line switches, to help reduce time to restoration as proposed by Agarwal [Agar16].

The rapidly rising levels of computer technology, in addition to the massive data-gathering capabilities of companies like Google and Amazon, develop the ability to process massive amounts of information at an extremely efficient speed. Several new concepts, techniques and new approaches have been emerging as a result of these developments, like *Knowledge Discovery, Big Data and Data Mining*, and in the field of the Artificial Intelligence: *Machine Learning, Artificial Neural Networks,* and *Deep Learning.* A brief description is presented below.

1.5.1 KNOWLEDGE DISCOVERY PROCESS (KDP)

Knowledge Discovery Process, is mostly used to find interesting hidden patterns in data. Cios et al. [Cios05] revealed studies where data mining plays a main role in areas like automation, employed as an efficient strategy for human decision-making, other authors support this theory as expounded in their works [Davi08], [Hutc08], [Erjo08] and [Pere08]. This study includes an in-depth discussion of KDP in Chapters 3 and 5.

1.5.2 BIG DATA (BD)

Big Data (BD) is distinguished from other kinds of data particularly because of its three salient characteristics: volume, velocity, and variety. In addition, it has become popular in many fields that stream data in large quantities, e.g. social media, as established by Anbalagan et al. [Anba16], finance [Fu15], energy [Yang17], weather forecasting [Tiem18], and many other areas that continue to be added to the list. In this study we do not employ BD because of time restrictions, but the ideas will certainly be part of a future study.

1.5.3 DATA MINING (DM)

The increasing amounts of data recorded and stored by government agencies and companies like Google, Amazon, utilities (smart meters, SCADA, PMU, etc.), Facebook and the like, have accumulated vast amounts of big data, as discussed below, which is being utilized at a greater

rate, as more large-scale data problems to need to be solved. Therefore, this has led to the creation of a whole new movement called *data mining*, which differs from classical statistical analysis in their hypothesis-testing. However, most of the methods used in data mining are related to methods fostered in statistics and machine learning (i.e., regression, classification, clustering, and visualization). DM methods focus on much smaller samples of data, due to the dimensionality reduction techniques used by the data mining approach. DM has the purpose to make sense of large amounts of data, mostly supervised, consistent with a specific domain, as studied by Cios et al. [Cios05]. Microsoft has the SQL Server Analysis Services tool (SSAS) [Micr17], to manage multidimensional models, a useful tool for predictive analytics and Machine Learning for Data Mining. Moreover, several studies in this area have been conducted using Data Mining techniques such as Xiao et al. [Xiao17], Yu [Yu16] and Zhao et al. [Zhao07] exposed. DM has been progressing dramatically in the recent years, the main reasons being advancement in storage technology and computational capabilities. Machine Learning (ML) and Deep Learning (DL) can have effective performance, via their algorithms, based on their datasets, regardless of the types of features the data consists of, or how it is represented. Therefore, data mining in a knowledge discovery framework is a good approach to implement in this study, to extract knowledge, and to better understand the data (or understand the nature of the extreme-event in the case of this study).

1.6 Problem statement

The power sector is one of the critical infrastructures that can be adversely affected by any kind of extreme weather, leading to disruption in the ETDS. However, the results of power disruptions by extreme weather depends on the length of exposure and degree of vulnerability in the ETDS system. For example, infrastructure components and geographic location characterizes the degree of vulnerability. Furthermore, natural disasters vary with the scale of the event, and are interconnected by geographic proximity (i.e., severity, geospatial regulator and energy sector) and level of exposure length (i.e., hours and days). During this study, we initially explore the causes of typical power faults, and power faults that occur in extreme weather, specifically *hurricanes*

and *winter storms*, that are among the highly devastating natural disasters responsible for massive and prolonged *power outages*. Finally, we explore Knowledge Discovery Process framework (KDP) and Data Mining techniques i.e., Machine Learning (ML) and Deep Learning Neural Networks (DLNN), finding remarkable progress in recent years in several applications i.e., computer vision, natural-language processing, and generative models. This leads to a methodology capable of consolidating both approaches, modernized measurement devices from the ETDS, and DM techniques in a KDP, under natural disaster conditions.

1.6.1 RESEARCH QUESTIONS

In recent years there has been an explosion in the application of data mining techniques and artificial intelligence in the sub-area of machine and deep learning. This is evidenced by the huge increase in the number of publications that employ ML and DL, for the state-of-the-art applications in computer vision, natural-language processing, and generative models. Similarity, the increasing state-of-the-art performance algorithms, including the most popular deep learning model, due to their effective representation and remarkable performance, are published in the IEEE Transactions on Neural Networks and Learning Systems, as evidence of their effectiveness. At the time of this study, knowledge discovery process (KDP) and data mining techniques, are not used to the point of prediction and classification, in terms of natural disasters in the ETDS, by following the path of a hurricane. The research questions to be addressed in this study, based on the need to exploit artificial intelligence in the ETDS are:

- As a main question in this study, would it be effective to use artificial intelligence techniques in a KDP framework to support the decision making during the assessment of the critical operation in the ETDS by following a hurricane path or winter storm?

- Which *variables* (meteorological, geographical, loads, phasor values), could be used in the architectures of machine learning and neural networks, to be the *most effective parameters* to define the level of weakness of the critical components in

the ETDS, and to predict and classify the *contingent-states*, and *critical-loads* in Energy Transmission and Distribution systems (ETDS) as a consequence of power faults, during the unfolding stages of the hurricane or a winter storm?

- How should these variables be *represented* in order to draw *accurate* prediction and classification modeling in the ETDS, during hurricane or winter storm stages?
- And, which are the *most effective models* (machine learning and neural networks) to derive *accurate* prediction and classification *tasks* by using those variables?

1.6.2 RESEARCH OBJECTIVES

From the observations research questions addressed above, the main objective of this study can be stated as follows:

1. The main objective of this study is to consolidate the domain of the power energy/data analytics by developing a methodology that takes advantage of AI techniques in a KDP framework, as a tool to generate accurate prediction and classification model assessment in the ETDS, under extreme-weather event conditions.

Similarly, other objectives in this study, based on the additional research questions and to test the proposed approach, two scenarios are defined and detailed as follows:

2. Scenario-1: Influence of a winter storm in the electricity market price and demand forecasting in New York City, using Knowledge Discovery framework as a data analysis process. The objectives for this scenario are:

    a) Datasets collection, preparation, and visualization as first steps in the KDP, to clean, normalize, and transform data into optimized datasets, including data analysis to find patterns, connections, and relationships for further modeling.

    b) Model using a Non-linear Autoregression exogeneous as a Data mining technique, then discuss the model performance.

    c) Complete a data analysis using Machine Learning by the use of pre-processing in data preparation for better model accuracy and performance.

11

3. Scenario-2: Power outages assessment in Houston area from the impact of Hurricane Harvey. The objectives for this scenario are:

   d) Datasets collection, preparation, and visualization as first steps in the KDP, to clean, normalize, and transform data into optimized datasets, including data analysis to find patterns, connections, and relationships for further modeling.

   e) Identify *variables* (meteorological, geographical, loads, phasor values), that can be used in the architectures of machine learning and deep neural networks, to be the *most effective modeling* to define the level of weakness of the critical components in the ETDS, and to predict and classify the *contingent-states*, and *critical-loads* in Energy Transmission and Distribution systems (ETDS) as a consequence of power faults, during the unfolding stages of the hurricane, respectively, then, discuss outcomes.

   f) Identify how the variables should be *represented* in order to draw *accurate* prediction and classification modeling in the ETDS, during the unfolding stages of a hurricane, then, discuss outcomes.

   g) Based on the assumptions of prior data analysis, test all approaches, modeling analyses to identify and understand the *most effective models* (machine learning and deep learning), for *accurate* prediction and classification *tasks*, then discuss model performance, and approaches, followed by research directions.

Due to the characteristics of the complex problem statement, several tasks will be tested. Data used will be from benchmark data, as well as data from specialized software simulations.

1.6.3 AIM OF STUDY

The contribution of this research study includes the following:

1) Two major literature surveys. The first provides a broad review of metrics and challenges in the Electricity Transmission and Distribution systems power operation

under extreme-weather events. This survey provides the readers with the information to make knowledgeable decisions about techniques and opportunities available to analyze the ETDS under an extreme-weather event. The second literature review covers the major works in prediction and classification modeling with an emphasis on evaluation metrics and model performance in Machine Learning and Deep Learning. This survey provides the data analytics scientist with a concise summary of the state-of-the-art machine learning and deep learning techniques and how these techniques can be used as a tool to predict and classify the two scenarios selected, as a critical operation of the ETDS during an extreme-weather event. Additionally, a discussion is provided on opportunities in the impact of this novel research field.

2) Creation of knowledge discovery framework methodology to consolidate data mining and modeling techniques with simulation of the critical operation in the ETDS through the challenge of an extreme-weather event. Additionally, given the uncertainty of the data, the point-of-view *Markov-state* based and *Empirical-based* probability sampling techniques are provided. Furthermore, a deterministic sampling is developed in this study. These methods allow researchers to accurately predict and classify the critical operations in the ETDS during the unfolding of extreme-weather events.

3) Finally, this dissertation provides several evaluation procedures for learning algorithms dealing with labeled and unlabeled data. These evaluation methods provide researchers with solid material examples on how to test learning algorithms in power energy systems.

1.6.4 STUDY AREA

From the observations in section 1.2, and to evaluate our approach from previous sections, we present two scenarios: a winter storm in New York, and a hurricane storm, in the Houston area.

Thus, to address Scenario-1, based on a winter storm, the geographical region of New York is selected because (i) it is one of the major cities in the U.S.; (ii) the electrical infrastructure of New York plays a key role in the U.S. economy; and (iii) the electric infrastructure is connected to others major cities in the U.S. including the interconnection with Canada. On the other hand, to address Scenario-2, based on a hurricane storm, the geographical region of Texas is selected because (i) it is one of the largest states in the U.S.; (ii) the electrical infrastructure of Texas has an independent interconnection named the Electric Reliability Council of Texas (ERCOT); and (iii) Texas is the largest energy-producing state and the largest energy-consuming state in the U.S., since the consumptions are by the industrial sector, refineries, and petrochemical plants. Figure 1.7 shows the geographical location of these two geographical regions in the United States.

Figure 1.7: Map of the United States showing the study areas.

1.6.5 SCOPE OF STUDY

The scope of this study includes the natural disaster of two scenarios with vastly different sets of characteristics. Both cases impacted the coast cities of the United States of America (USA), as discussed in section 1.6.4, and are used in this study as a proof of concept. Other geographical regions in the USA and other countries are not considered, due to time constraints of the study, but results may be applicable to other regions of the USA, and of the world. The methodology is generalizable and can be applied in other types of storms.

14

An outline of this dissertation is as follows: Chapter 2 is a background discussion of different concepts in the critical operation of the ETDS system under extreme-weather events. Furthermore, chapter 3 presents a brief review of the Knowledge Discovery Process (KDP). Chapter 4 presents a review of the data mining techniques in AI used as experimental techniques in this study. Chapter 5 presents the detailed development of a generic framework based in KDP, with adaptation techniques from chapter 4, including the deep learning technique applied in the critical operation of the ETDS in an extreme-weather approach. This generic methodology framework is able to provide a roadmap to seek and extract new and useful knowledge in data. The latter is followed by discussion of its implementation using two different scenarios, as previously discussed. Chapter 6 discusses the winter storm. An extension of the approach is discussed and modeled in chapter 7. Chapter 8 discusses conclusions. Finally, Chapter 9 provides the summary conclusion of this dissertation, and future research directions. Appendices A to D are included for completeness, in that they contain notations and supported references.

CHAPTER 2: METRICS OF THE CRITICAL OPERATION OF THE ETDS SYSTEM UNDER EXTREME-WEATHER EVENTS

Knowledge of the difference in behavior during normal operation and extreme-weather event operation in the ETDS is critical in order to understand when the event is occurring. This section presents a brief overview of classical electric power grid configurations and the state-of-the-art metering systems to acquire data from the ETDS to detect a power failure. The current state and availability of the components, power plants, and grid infrastructure across the ETDS is a challenging task. Thus, as a way to tackle the complexity of the ETDS, a probabilistic risk analysis in the ETDS by using Markov analysis is presented in Chapter 7, as a method of experimentation to obtain systematic uncertainties. Additionally, this section discusses power fault metrics and power systems resilience that can be measured to assess the effectiveness of power system operation under the impact of an extreme-weather event.

2.1 Power electric grid configurations

The analysis of power energy in transmission and distribution networks is performed mostly in two major configurations, as stated by Panigrahi [Pani17] and Vorobev [Voro17]: Radial, as a single representation, and Interconnected, as a ring or loop representation. Figures 2.1 and 2.2 depict radial and interconnected configurations, respectively.
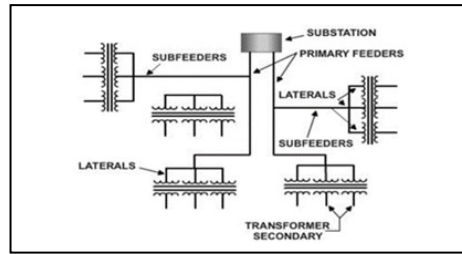
Figure 2.1: ETDS radial configuration, image taken from [Elec18].
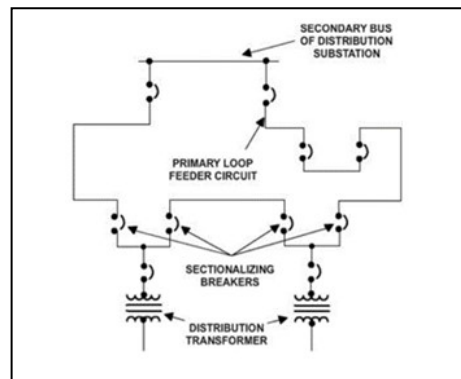
Figure 2.2: ETDS Interconnected configuration, image taken from [Elec18].

On the other hand, many authors use IEEE Reliability Test Systems (RTS) to prove their findings [Kuma14], [Gonz13], and [Peng15]. Figure 2.3 and 2.4 show the IEEE 9-bus modified system and the IEEE 14-bus modified system, respectively. In this study, scenario-2 simulations conducted using IEEE 9-bus modified are presented. IEEE14-bus modified systems (future work) are cited from the author [Deme17] to prove our findings, because in the presence of a hurricane, the power system is vulnerable to severe contingencies propagated along the power system leading to power system instabilities, as explained in chapter 7. Both of the IEEE systems are used for analysis of transient instabilities, exposing the failure to retain the synchronism of the main components of the power system (generators to the rest of the system) after a severe disturbance. On

the other hand, the synthetic Texas electric network (ERCOT) from the authors in [Birc17], intro-
duces the methodology to generate synthetic line topologies in a 2000 bus case using realistic
parameters that satisfy loads, generations, and transmission lines. This case includes realistic geo-
graphical placement of the substations, using real energy and population demand from the selected
study area. Figure 2.5 shows the AC power flow solution (without any manual intervention, e.g.
system operator best decision options detailed next in section 2.2) of the Texas 2000-bus. The
reliability test busses data from the IEEE 9-bus, IEEE 14-bus, and Texas 2000-bus are presented
in Appendix A.2.
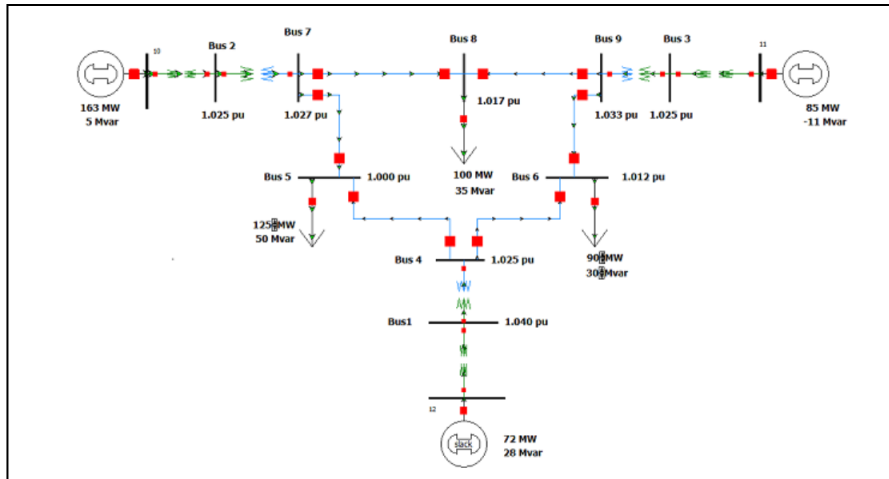
Figure 2.3: IEEE 9-bus reliability system, image taken from [Deme17].
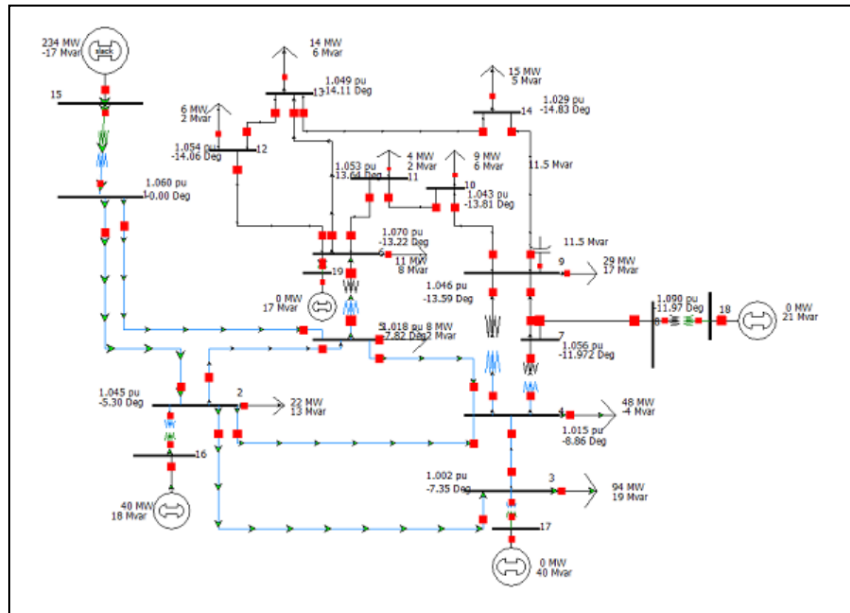
Figure 2.4: IEEE 14-bus reliability system, image taken from [Deme17].
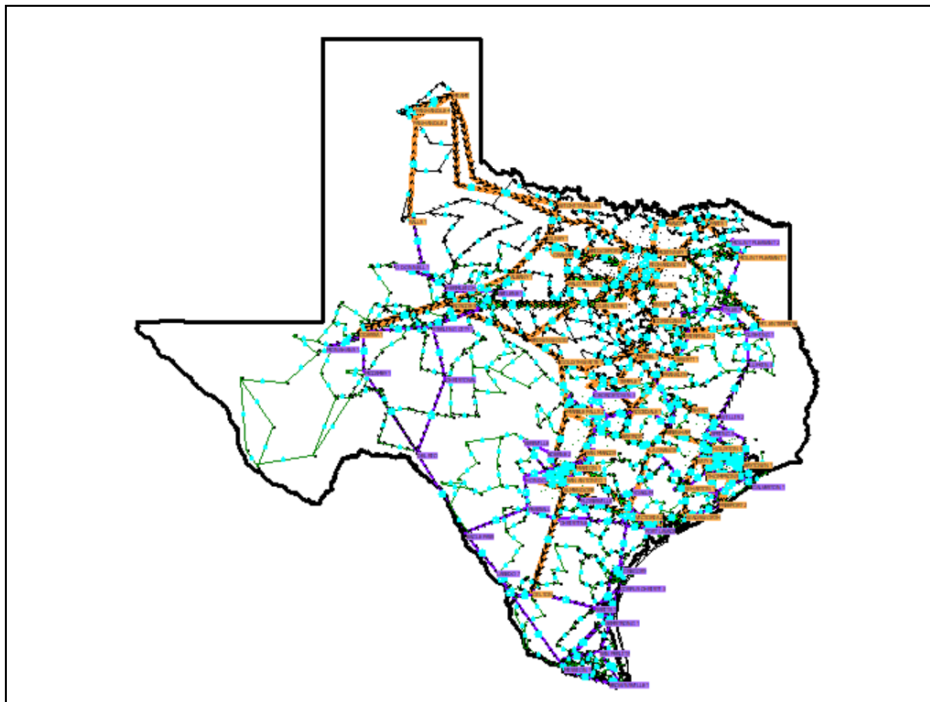
Figure 2.5: Texas 2000-bus reliability test system, image taken from [Birc17].

19

2.3 ETDS state-of-the-art data acquisition

Recently, modernization in the ETDS allows the generation and acquisition of big data, by means of new technology devices like RTUs, *Phasor Measurement Units* (PMUs), and smart meters within the Advanced Metering Infrastructure as discussed in section 1.3.2. and 1.3.4. Furthermore, the recent deployment of distributed intelligent devices and PMUs was used to enhance system stability, by detecting and isolating disturbances caused by different events as a short-circuit, equipment failure, temporary damage, or an extreme-weather generated fault as stated by Yu [Yu17] and Nakafuji [Naka17]. Assessment in the power distribution system components are crucial to strengthening the ETDS system reliability under normal conditions, as well as increasing resilience under extreme-weather conditions.

Data acquisition in the ETDS by the *Wide-Area Measurement Systems* (WAMS) using PMUs is increasing. Today, the number of entities sharing data flow is increasing in number, adding more PMU units and syncrophasor data communication networks along the power grid. Figure 2.6 depicts PMU locations in North America and related Data flow during 2015 reported by Energetic Incorporated [Ener16].
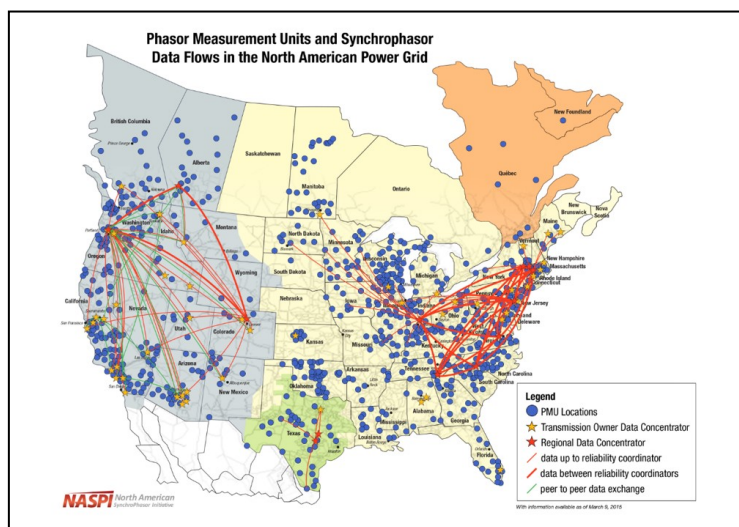
Figure 2.6: PMU locations in North America and related power flow during 2015, image from [Ener16].

PMU is a sophisticated monitoring device that can measure the instantaneous voltages, currents and frequencies as a measure of phasor, frequency, and rate of change of frequency (ROCOF). Equation 2.1 represents the AC signal $x(t)$ with constant frequency and magnitude.

$$x(t) = X_m \cos(wt + \varphi_o) \tag{2.1}$$

Where X_m represents the signal peak value, the angular frequency is $w = 2\pi f$, and φ_0 represents the initial phase of the signal. The complex phasor is represented in equation 2.2 as follows:

$$\bar{X} = \frac{X_m}{\sqrt{2}} e^{j\varphi_0} = \frac{X_m}{\sqrt{2}} (cos\varphi_0 + jsin\varphi_0) = X_r + jX_i \tag{2.2}$$

Where X_r and X_i represent the real and imaginary rectangular components of the complex phasor representation, respectively. The RMS value of the sinusoid is $X_m / \sqrt{2}$ and ROCOF is represented by ROCOF $= \frac{d_f}{d_t}$ [Mont16]. Figure 2.7 shows the phasor representation of sinusoidal signal.
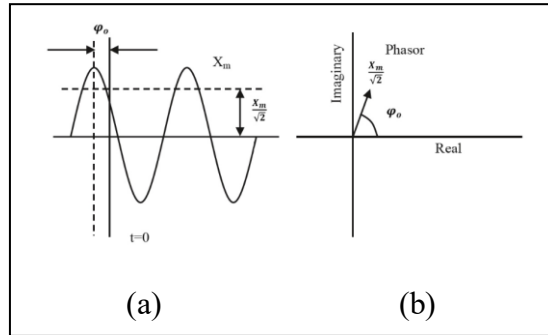
Figure 2.7: Phasor representation of a sinusoidal signal. (a) Sinusoidal waveform signal. (b) Phasor representation.

The phase measurement is estimated via the orientation, based on the global time reference from the Global Positioning System (GPS) situated in a specific location on the grid. In this study, steady-state modeling and ROCOF estimation was not conducted. The data from the PMU is received by a Phasor Data Concentrator (PDC), which concentrates and manages the information of a wide geographical area from multiple PMUs located in optimum locations in the ETDS for fully

observability. Figure 2.8 shows an IEEE 9-bus with PMUs located in predefined buses, including a Phasor Data Concentration (PDC) unit, modified from [Dyna19].
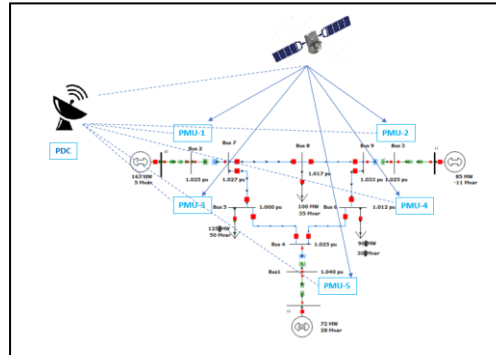
Figure 2.8: IEEE 9-bus with PMUs located in predefined buses, including a Phasor Data Concentration (PDC) unit, modified from [Dyna19].

The PMU is able to detect frequency variations that occur in a wide range of disturbances, e.g. harmonics (10^2-10^3 Hz), inter-harmonics (0-10^3 Hz), network resonances (10^3-10^4 Hz), power swings (0-10 Hz), power faults (10^4-10^5 Hz) and lighting (as fast transients, 10^6 Hz). Phadke [Phad08] states that the PMUs are able to acquire data at high speeds, 100 times faster than SCADA systems. Additionally, Monti [Mont16] confirms that PMUs have their communication systems interconnected to different points (nodes) in the wide area of the ETDS, as mentioned above and depicted in Figure 2.4, allowing utilities to measure the state and health of the electric grid over vast expanses of geography in near real-time. With this approach, a situational awareness is provided to the monitoring systems between the large interconnected systems in the ETDS by giving the availability to track what is happening over time and space, and be able to take appropriate action if necessary.

On the other hand, a *Smart Meter (SM)* measures energy consumption in real time and in more detail than conventional meters, which are only capable of recording the household consumption without reporting any other energy activity. Recently, with the concept of "smart grids," SM usage is growing across America covering 43% of U.S. homes. Other countries are imple-

menting those smart measures devices as well. The SM is also a component of the *Advanced Metering Infrastructure (AMI)* that realizes the transmission of data from the customer houses to the Meter Data Management System (MDMS) to the network, collecting data measurement in time intervals of 15 or 30 minutes. In a bi-directional or two-way communication technology, the MDMS executes a request to the smart meter and then the results are sent back by the use of wireless communication, as stated by Tan et al [Tan17]. This enhances the utilities' ability to integrate the smart meter data into outage management systems, allowing the identification of power outages on distribution networks, without sending crews to search the physical areas.

Of interest here, we can start by implementing PMUs to monitor and track power faults during extreme-weather events, by means of wide-area visualization (topology), and smart meters to detect household interruptions, as well as a demand management during the evolving time period of the extreme-weather event.

2.2 Power fault metrics

Probabilistic and statistical analyses are required to measure power faults in the ETDS, in a standard form. Over the past decades, terminologies like reliability and availability incorporate statistics and probability into the analysis of power faults, which have been the most used measures by several utilities to quantify the performance of the components of the ETDS. Using both metrics can define the dependability of the components, or the system, as stated by Short [Shor04]. This is important because the purpose of the components working together is to deliver energy in a reliable fashion. Climate change, as discussed in section 1.1, increases the frequency, intensity, and duration of extreme-weather events, impacting system reliability over time in the ETDS infrastructure. EWPR analysis is a key approach to localize weak power system components as mentioned in previous section. Increasing the landfall of a hurricane results in the dramatic decrease of reliability measures in the ETDS. As mentioned above, *Reliability* (R) is one of the cost-effective measures in the ETDS. By estimating the system performance, it is capable of warranting continuity, and quality in the electric grid, as this is the apparent goal of the ETDS. Several

methods were proposed by several authors like Short, Kumar, Cheng and Espiritu [Shor04], [Kuma14], [Chen09], [Espi07], for the analysis of reliability in the ETDS systems. Moreover, the technique most frequently used was the Monte Carlo or Sequential Monte Carlo method. These techniques work well, but in complex conditions, such as those resulting from the addition of the variable *extreme-weather*, become a limitation in assessing the impact of the extreme-weather. Authors Gaver et al. [Gave64] and Billinton et al. [Bill68], were the first in separating the distinction between *normal condition power faults* and *under the influence of weather-related failures*, to model the sudden increment in the failure rate of the components, and the probability of overlaying during an extreme-weather events. Other authors like Bhuiyan and Billinton [Bhui94], [Bill02], use this approach based on a two-states model: Normal and Extreme. Other effects of the extreme-weather events are the physical damage of the components in the ETDS from the increasing in intensity, frequency, and duration of such events, as discussed in sections 1.1, 1.3.3, and 1.5.2. Aged components are more likely to be more vulnerable to extreme-weather conditions as stated by Bruch et al. [Bruc11] and Shafieezadeh et al. [Shaf14].

Due to the this assumption, reliability measurements need to be separated into two major metrics: Typical Power Failures Metrics (TPFM) and Extreme-Weather Power Failures Metrics (EWPFM), both are reviewed below to understand distinctions. In addition, critical components in the ETDS system need to be considered, therefore, the component reliability importance metric is also presented.

2.2.1 TYPICAL POWER FAILURES METRICS

Utilities typically use reliability indices in their calculations to categorize service quality. The most common metrics in a reliability analysis are: SAIFI (System Average Interruption Frequency Index), SAIDI (System Average Interruption Duration Index), CAIDI (Costumer Average Interruption Duration Index), CAIFI (Costumer Average Interruption Frequency Index), and ASAI

(the Average Service Availability Index). All of these represent the portion of time that the customer has received electricity during the elapsed period reported and suggested as standard by the IEEE Power and Energy society [Ieee12].

2.2.2 EXTREME-WEATHER POWER FAILURES METRICS

The metrics used in EPFM for reliability and availability differs from the TPF, since the utilities do not include the extreme-weather related power faults metrics as described above. The Storm Average Interruption Index (STAIDI) is used, instead of SAIDI, with momentary interruption using MAIFI as proposed by Brown et al [Brow97]. Moreover, the Potential Storm Event (PSE) is used by the same author to predict STAIDI, described as the probability density function of the duration, and root mean square (RMS) of wind speed during its time interval.

2.2.3 COMPONENT RELIABILITY IMPORTANCE METRICS

It is understandable that some ETDS infrastructure components are more critical than others. The component importance index, then, is used to rank and classify target components which need to be enhanced in the reliability process. Different measures are used in probabilistic risk analysis to establish the importance of an ETDS component. Likewise, the importance of a component depends on two factors: location within the system, and reliability of the component. The most popular measures, including the factors for component importance analyses as Rausand et al. states [Raus04] are: Birnbaum's measure (BM), Improvement potential measure (IPM), Risk achievement worth (RAW), Risk reduction worth (RRW), Critically importance measure (CIM), Fussell-Vesely's measure (FVM).

Espiritu [Espi07] suggests a modification of the component importance measure to be used in Electrical Transmission System (ETS), in which traditional measures as described above cannot be directly used because they are not properly characterized for probability of failure or success for the specific mission. It is then suggested to use individual component sustained outage rates and system unavailability (increase, decrease) instead of a specific probability of failure that the

standard methods use. Other authors, Hilber et al. [Hilb07], propose the use of customer interruption cost as a measure of system performance instead of the use of every load point, in order to identify critical components in the system by reducing the calculation of the whole network.

2.3 Power systems resilience definition and metrics

The Electric Transmission and Distribution Systems has been increasingly at risk mostly by the effect of the extreme weather in recent years. The resilience in the ETDS changes during the high impact of these events. **Resilience** is, at the present, the foremost indicator for power delivery during and after an extreme-weather event and also for the enabling of restoration of the ETDS system. To understand the concept of resilience the following is a broad definition taken from Carlson et al. [Carl12]: "*The ability of a system to anticipate, resist, absorb, respond to, adapt to, and recover from a disturbance.*" Power faults are always present at some time in the ETDS, where the system is mostly prepared under the N-1 criterion (reliability), but in an extreme-weather event (e.g. a hurricane) the N-1 criterion is exceeded, and numerous instantaneous power faults could occur through the components and through the system, caused by the physically compromised exposure of several components in the ETDS systems (i.e., generators, transmission lines, transformers, or other equipment contributing to the overall performance of the ETDS system) through the route of the hurricane. It is necessary to treat all the stages of an extreme-weather event, using the state-of-the-art technology, methodologies, modeling and software to enhance the resilience in the ETDS. Resilience is a feature that is distinguished from the conventional principles of reliability. The distinguishing factors are: 1) Inclusion of all hazards and events (reliability does not include hurricanes, for example); 2) Quantification of all transition times through the states; and 3) Capture of the effects damages to customers (reliability captures only the number of customers impacted), grid operators and staff, and infrastructure [Tech18, Kahn17]. Despite the fact that resilience is an abstract concept and is also difficult to quantify, several metrics have been proposed by several authors [Pant17, Azad17, Azad15, Kahn17]. Therefore, finding a standard procedure is a challenging task, thus, there is no universal procedure to be used. However, a good

explanation and evaluation of resilience is found in the work of Lin et al. [Lin16b], where the authors measure the resilience of a system using qualitative and quantitative evaluation methods. This study will not explore the challenges in the measurement of resilience, but the job of the resilience metric quantification used by Panteli et al. [Pant17], based on *multi-phase resilience trapezoid* (MPRT), is depicted in figure 2.9. This method is taken in this study as a suitable reference. The MPRT defines a set of metrics to capture the performance during the unfolding of an extreme weather event. Moreover, the same authors divide the resilience into two main subjects: *operational resilience*, referring as the ability to guarantee uninterruptible energy supply to customers in the case of extreme weather; and *infrastructure resilience*, referring to the physical ability to mitigate the section of the system that is lost, damaged, or nonfunctional in the case of extreme weather. The motivation for using MPRT as a resilience metric emerges from the contribution to providing a big picture of the resilience level, by including the operational and infrastructure perspective.
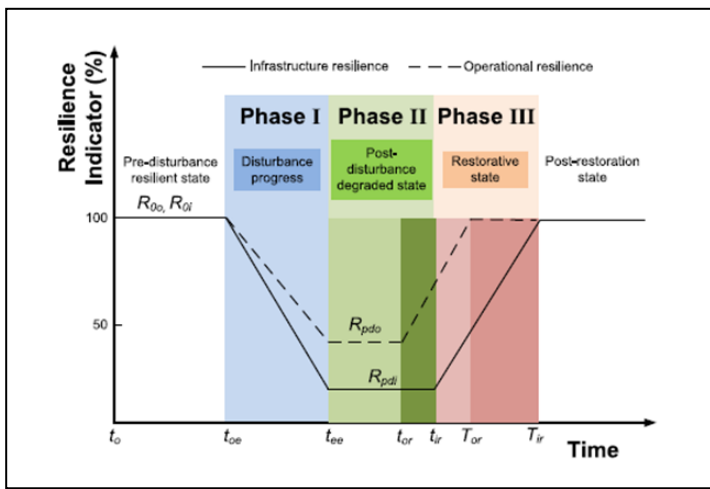
Figure 2.9: Multi-phase resilience trapezoid, image taken from [Pant17].

The pre-disturbance resilience begins at 100 percent, which indicate the normal operation of the power system. At *Phase I*, during time t_{oe} to t_{ee}, the resilience drops through the disturbance progress. At *Phase II*, during the post-disturbance degraded state, the power faults in the ETDS

system remains until the load restoration is initiated, the duration depends on the resilience of the covered region in the path of the hurricane i.e., fragility of the ETDS. *Phase III*, the restoration state, has times t_{or} and t_{ir}, where the operational and infrastructure restoration, respectively, are initialized. The occurrence of t_{or} is faster than t_{ir}, because in real world customers are reconnected first, then flooded substations or collapsed transmission systems are restored later. The three phases of resilience are dependent on many factors, as studied by Adderly [Adde16]. These factors could be, for example, frequency of blackout, magnitude, time of the year, time of day, and geographic location of the extreme-event. Figure 2.10 presents the mathematical expression of the resilience metrics proposed by Panteli et al [Pant17]. The different phases of the resilience trapezoid are calculated using these metrics to capture the performance through the unfolding of the extreme weather event (i.e. hurricane). Moreover, the resilience area defines the timeline during the disturbance and degraded states in both operational and infrastructure resilience, respectively.

Phase	State	Mathematical Expression		Measuring Units	
		Operational	Infrastructure	Operational	Infrastructure
I	Disturbance Progress	Slope: $\dfrac{R_{pdo} - R_{0o}}{t_{ee} - t_{oe}}$ (2.3)	Slope: $\dfrac{R_{pdi} - R_{0i}}{t_{ee} - t_{oe}}$ (2.4)	MW/hours	Number of lines tripped/hours
		MinResilience $R_{0o} - R_{pdo}$ (2.5)	MinResilience $R_{0i} - R_{pdi}$ (2.6)	MW	Number of lines tripped
II	Post-disturbance degraded	MinResilience: $t_{or} - t_{ee}$ (2.7)	MinResilience: $t_{ir} - t_{ee}$ (2.8)	Hours	Hours
III	Recovery	Slope: $\dfrac{R_{0o} - R_{pdo}}{T_{or} - t_{or}}$ (2.9)	Slope: $\dfrac{R_{0i} - R_{pdi}}{T_{ir} - t_{ir}}$ (2.10)	MW x Hours	Number of lines restored/hours
Resilience Area:		ORA: $\displaystyle\int_{t_{oe}}^{T_{or}} R_{op}(t)\,dt$ (2.11)	IRA: $\displaystyle\int_{t_{oe}}^{T_{ir}} R_i(t)\,dt$ (2.12)	MW x Hours	(Number of lines in service) x hours

Figure 2.9: Mathematical expression of resilience metrics, modified from [Pant17].

The proposed resilience metrics from Panteli et al. [Pant17], were successfully tested in their work, and is, therefore, a good approach to quantify the resilience in the area of interest.

2.4 Optimal power flow analysis in transmission and distribution systems

The controlled generating unit as a real power output in a selected area to sustain a given load under minimal total operating costs is named Economic Dispatch (ED). However, the only weakness of the ED is that "it ignores the limits imposed by the devices in the transmission lines" as stated in [Glov12]. Moreover, in reality the transmission lines and transformers can only handle a limit of the volume of power that can be transmitted through it, as the effect of the thermal, voltage, or stability restrictions. The United States, as well other countries, operates under a regulated market, i.e. ERCOT, as mentioned in section 1.6.4, where the cost associated with active transmission lines is an important factor. Enforcing the generation transmission lines to act with the ED is what today we know as Optimal Power Flow (OPF) [Glov12]. The most used power solvers simulators are: Powerworld (used in this study), Matpower, Powerfactory, and Digilent. Scenario-2 of this study is based on the power solver Powerworld ver. 21. The next section briefly discusses the representation of DC power flow and the DC OPF used in this study. Since this study focuses only on the development of KDP using ML and DL, the theory behind electrical circuit analysis and their representations will not mentioned here, but more detail can be accessed in [Glov12].

2.4.1 DC POWER FLOW

An effective power system operation under normal balanced three-phase operation in a steady-state condition entails the four conditions [Glov12] below:

- Generation needs to deliver the needed demand (load) plus losses;
- Bus voltage magnitudes need to remain close to rated values;
- Generators need to operate within the specified real and reactive power limits; and
- Transmission lines and transformers need to be under overloaded limits.

The computation of the voltage magnitude and phase applied in each bus of the power system under a balanced three-phase steady-state condition, is named *power flow problem*. In an *AC power grid*, the n buses and the m transmission lines constitute a complex network. The *AC power flow analysis* in each node i is labeled as one of the following [Glov12]:

a) *Slack bus* (or Swing bus). In the analysis, there exists only one slack node indexed as node 1 (in many cases), with voltage typically 1.0 with 0 angle per unit as input data. Also, the slack node computes the active power P_1 and the reactive power Q_1 in the power flow solver.

b) *Load (PQ) bus*. The active power P_i and the reactive power Q_i are input data, which need to be known for the power solver, where V_i and θ_i are computed with the program.

c) *Voltage Controlled (PV) bus*. The active power P_i and the voltage magnitude $|V_i|$ are input data, which need to be known, and the reactive power Q_i also with the angle θ need to be computed with the power flow solver.

AC power flow is a nonlinear system, with respect to their voltages. DC power flow provides a linearized approximation of the active power flow presented in the AC model. Many authors use *DC power flow approximation* instead of using AC power flow in their analysis for simplicity purposes [Wang17], [Sang19]. The DC linearization needs to follow the conditions [Glov12] below:

a) The difference between the voltage phase angles of every coupled neighboring bus needs to be small, such that $\theta_{ik} \approx \theta_{ik}$ and $\cos\theta_{ik} \approx 1$.

b) The active power losses need to be negligible, thus, $Y_{bus} \approx iB$ (B is the admittance matrix, imaginary part), which in the computation, the line resistance is neglected.

c) The voltage magnitudes' $|V_i|$ variation need to be small, and it is assumed that $|V_i| = 1\ \forall i$.

Consequently, under the above assumptions and given the active power P_i at each bus I, the angle of the buses can be estimated by $\tilde{\theta}_i$ using the *DC power flow approximation*. Equation 2.3 and Equation 2.4 are the DC power flow formulation equations.

$$P_i = \sum_{k=1 \ k\neq 1}^{n} P_{ik}^{(DC)} = \sum_{k=1 \ k\neq 1}^{n} B_{ik}(\widetilde{\theta_i} - \widetilde{\theta_k}) \qquad (2.3)$$

Or reduced in a matrix form with the real power balanced to a complete linear problem,

$$\tilde{P} = -B\tilde{\partial} \qquad (2.4)$$

Where,

B is the imaginary part of the Y_{bus} $\tilde{P} = [\tilde{P}_t, P_i, P_i]^T$, and $\tilde{\partial} = [\tilde{\partial}_1, ..., \tilde{\partial}_n]^T$ are the matrices estimated by assuming that the phase angle at the slack bus (first entry) is 0. While the DC power flow lines are lossless, therefore, Equation 2.5 can be used to estimate the phase angle on the buses by solving Equation 2.4.

$$\widetilde{P_t} + \sum_{i=2}^{n} P_i = 0 \qquad (2.5)$$

2.4.2 DC OPTIMAL POWER FLOW

As mentioned in section 2.2, the OPF plays an important role in planning and system operations derived by the action of the TSO. Different optimization techniques exist to solve the OPF problem, the most common are *Linear Programming (LP)*, *Dynamic Programing (DP)*, *Newton-Rapson (NR)*. Powerworld OPF solver uses LR technique. All of these methods have their advantages that will not be discussed in this study. The goal of the OPF, as discussed in section 2.2, is to minimize the lost load during an event in the ETDS system. Under this assumption an approximation of the objective function in a *DC Optimal power flow,* can be represented by the Equation 2.6, taken by [Moha19].

Minimize $\quad \sum_s \{ \pi_s \sum_t [\sum_g (c_g \, PG_{(s,g,t)} + c_{(g)}^{NL} \, u_{(s,g,t)} + c_{(g)}^{SU} \, v_{s,g,t} + c_g^{SD} \, x_{(s,g,t)}) +$

$\sum_n (c_{(n)}^{(lsh)} \, p_{(s,n,t)}^{lsh}) + \sum_g (c_{(g)}^{og} \, P_{(s,g,t)}^{og})]\} \qquad (2.6)$

Where *generation cost* is represented by the first term, which includes: power generation cost (no load cost, start-up cost, and shut-down cost), *load shedding cost* is represented by the

second term, and *overgeneration cost* is represented by the third term of Equation 2.6 above. Equation 2.6 has constraints which, as a plain reference without mathematical expressions, can be summarized below:

- Generation ramping limits (maximum and minimum),

- Minimum Up Time and Down Time constraints of generators,

- Nodal injected power calculation (when load shedding and over generation are allowed),

- Power Flow of Lines (ensuring the flow on the lines that should be monitored stay within limits),

- Calculate the power flow for such lines accounting for changes in topology,

- Power balance (allowing load shedding and over-generation),

- Ramp-Up and Ramp-Down limitation over generation limits, and

- Other constraints, i.e. generator status added by the operator.

2.5 Extreme-weather probabilistic risk analysis (EWPRA)

The use of an EWPRA is a useful technique to localize weak spots in the ETDS. Moreover, extreme-weather risk analysis aims to prepare contingencies in the ETDS for rapidly recovering, and to prevent or mitigate the impact of similar events. HAZUS simulator, created by the Federal Emergency Management Agency (FEMA), is used to estimate damages and loss of buildings and essential facilities from Earthquakes, Floods, Hurricanes, and Tsunamis and to visualize such events. HAZUS was used in this study to visualize the hurricane event. Moreover, Stochastic methods like Monte Carlo, Sequential Monte Carlo, and several other methods are widely used as stated by the authors Brown et al. [Brow97], Alvehag et al. [Alve08], and Balijepalli et al. [Bali05] to assess the risk analysis in the infrastructure due to extreme-weather. Furthermore, HURDAT2 is a historical dataset collected since 2005 to the present, and regularly upgraded with data collected by the Oceanographic and Meteorological Laboratory at National Oceanic and atmospheric Administration (NOAA). HURDAT2 contains broad six-hourly information about hurricanes, which

includes data like the maximum wind speed, location, central pressure, etc. HURDAT2 [Hurr18] is a comma-delimited text format dataset (CSV), and is commonly used in literature for analyses in atmospheric predictions and other applications, like HAZUS, in our case. Furthermore, HAZUS uses Geographical Information System (GIS) technology to estimate physical, economical, and social impacts of disasters which can identify high-risk locations due to the extreme event as previously mentioned. Additionally, HAZUS can use customized data input in the model to generate new scenarios, but only scenarios offered by the HAZUS software can be created. The geoprocessing data from HAZUS needs to be merged into ArcGIS software, specifically ArcGIS ver. 1.5.1. ArcGIS version 1.7.x and ArcGIS Pro can manage Deep Learning in their platform as a smooth means. Deep Learning concepts are explained in further chapters. In this study, deep learning is not evaluated under those versions, only supervised and unsupervised machine learning is used to evaluate the vulnerability in the ETDS under Scenario-2. Scenario-2 and Machine Learning algorithms are presented and discussed in further chapters, respectively. All ArcGIS versions including that required by HAZUS, requires a subscription to allow the access to the ArcGIS platform. Some versions are costly, especially the version including Deep Learning. This study uses a student subscription of ArcGIS to access only the HAZUS requirement to evaluate the performance and validate the accuracy of the proposed model, specifically in Scenario-2. The reason that DL under GIS systems is not compared is that DL algorithms under those platforms are not available with the software. ArcGIS software, ENVI, QGIS and other platforms can be used to create and manage GIS information. Specifically, ArcGIS software is administered by the Environmental Systems Research Institute (ESRI), created in 1969. Many databases and projects under ESRI were developed by NOAA, NASA, IEA, etc., which generated lot of geodatabases, raster data (mosaic data structure for managing multidimensional raster and imagery data), and vector data (shapefiles). Geoprocessing data from the platforms mentioned above are mostly shared on the web. Therefore, geoprocessing data from authors' findings shared on ArcGIS and Geoplatform is used in this study (Scenario-2), since the purpose of this study is not an analysis in the prediction of hurricanes, winter storms, or any other extreme event. In this manner, the hurricane path modeling

is only under the assumption of HAZUS software ver.4.2 [Fema19]. Analyses of the Supervised and Unsupervised Machine Learning can be accomplished with photos taken before and after the hurricane, in order to visualize the damaged location as a geographical asset using satellite image or other technology, i.e. Light Detection and Ranging (LiDAR). ENVI software offers a good approach to those assets.

2.6 Conclusion

The vulnerability analysis in the ETDS during an extreme-weather event needs to be deepened by the examination of the components. The components' failure rates need to be detailed in a fragility curve, according to the potential damage of the wind under the extreme-weather event [Li14], to estimate the power failures in the ETDS as a first rule. A second rule, on outage assumptions can be deterministic. In the case of using the deterministic rule, it is necessary to corroborate real outage history by the utility companies in the geographic area of study, or by means of observations using remote sensing techniques, with digital image analysis (image in Tagged Image File Format (TIFF) with ML techniques. The outcomes can be taken as an action from the TSO, by pondering current and future topology states in the ETDS. On the other hand, failures in the ETDS can result as input data representing the system states [Wang17], or scenarios [Sang19] in the OPF solver. DC OPF can be a simplified method to solve the optimization problem in the ETDS, including the spatial-temporal translation of the extreme-weather. In the next section, a novel methodology is proposed to aid the resilience in the ETDS by predicting load, detecting critical loads, and as a tool to help operators to make better decisions.

CHAPTER 3: KNOWLEDGE DISCOVERY PROCESS FRAMEWORK

The use of raw data in data analytics presents several issues in data quality. Raw data often contains outliers, data errors, noise, and possible hidden trends. Additionally, massive data generation can cause scalability problems, as well as insufficient feature description, fashioning impractical datasets for Modeling. From this assumption, the data driven problems can be solved following a road map of process called Knowledge Discovery Process (KDP). Thus, this chapter presents a KDP multistep framework implemented to solve the problem in the datasets, as results of raw data acquisition from the operation of the Electricity Transmission and Distribution Systems (ETDS) under extreme-weather events. These latter KDP processes are a source of inspiration of many data scientists that have worked with data mining projects, and are adapted in chapter 5 for the purpose of this study.

3.1 Knowledge discovery process framework

Knowledge Discovery Process consists of multiple steps that are executed in a sequence. As mentioned in section 1.5.1, its main purpose is to seek new knowledge by finding interesting hidden patterns in data from a certain domain. Gullo [Gull15] proposes five steps, as shown in Figure 3.1. This figure was briefly modified from [Gull15] to show the steps, which are described below.
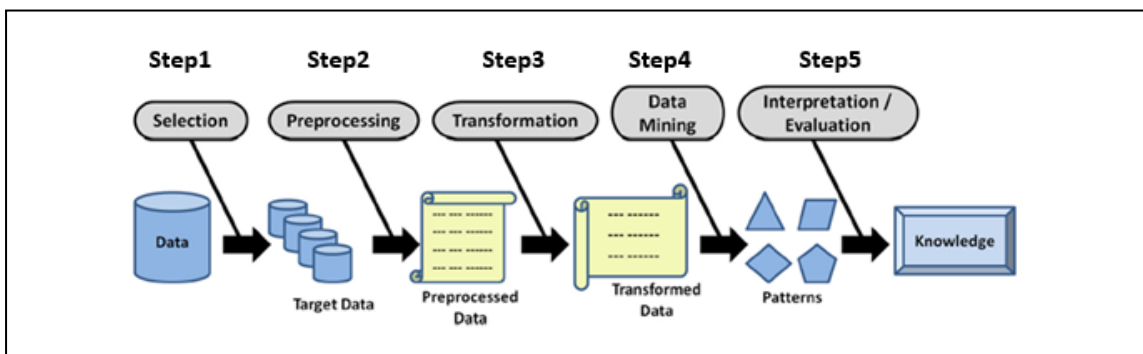
Figure 3.1: Knowledge discovery process, briefly modified from [Gull15].

Step 1.- **Selection** is the assessment of the situation in which discovery has to be achieved, by generating target data (i.e. collection of initial data, selection of subset of variables or data samples, exploration of data, and verification of data quality);

Step 2.- **Preprocessing** is the cleansing of data by completion of diverse operations, like removal of outliers, dealing with noise and missing values in data, and accounting for time-sequence information.

Step 3.- **Transformation** consists of using techniques for selection and extraction algorithms (reducing dimensionality), by derivation of new attributes (i.e. discretization), and summarization of data (data granularization), in order to meet suitable representation (final data set) to be used as input, to be fed in the data mining tools in the next step;

Step 4.- **Data Mining** involves the use of several methods to extract interesting patterns (knowledge) from preprocessed data by choosing specific data mining methods or tasks (i.e. classification, clustering, regression, etc.), proper algorithm(s) for completion of the assigned mission, and suitable representation of the output results;

Step 5.-**Interpretation/evaluation** is supported by interpretation of the visualization of patterns, models, or data, and by the understanding the results. Often, replication into the previous step is necessary.

Furthermore, many authors are using KDP and Data Mining in their studies, like material science and engineering and many other areas. Other authors such as AbuOmar, Fayyad, Gull, Cheng, Thomas, and Bandaru [Abuo13, Fayy96, Gull15, Chen18, Thom17, Band17a], and [Band17b] are in favor of KDP to manage large amounts of raw data, validating the efficiency of the KDPs. A discussion of data mining techniques is presented next.

CHAPTER 4: EXPERIMENTAL DATA MINING TECHNIQUES

In the introduction (Chapter 1), some data mining characteristics were briefly mentioned, and in the previous section (Chapter 3) the knowledge discovery process framework was discussed. In this chapter, the subject of Data Mining techniques is furthered expanded. Data Mining forms the core step in KDP, and some of the most used algorithms in Machine Learning and Deep Learning Neural Networks are also discussed in this chapter, including their characteristics and a discussion of their state-of-the-art programming frameworks and libraries. Their interesting performance approaches in prediction and classification are further coupled and adapted to the proposed framework in Chapter 5, to further solve the problems in Chapter 6.

4.1 Data Mining techniques

One of many definitions collected by the author [Gull15] of "data mining", is "*automated exploration and analysis of large quantities of data in order to discover meaningful patterns.*" The patterns and trends can be collected and defined as a *data mining model*. Thus, data mining is a method to discover actionable information from large sets of data. Data mining uses mathematical exploration to derive patterns and trends that exist in data, mostly as unsupervised data in some domains, which is less expensive as supervised data, that deal with known input corresponding to known output (expert needed to determine this relation). Typically, these patterns cannot be discovered by traditional data exploration because the relationships are too complex, or because there is too much data. The methods of constructing data mining models are discussed in section 1.5.3 are reviewed next.

4.1.1 MACHINE LEARNING

Machine Learning evolved out of the subfield known artificial intelligence. *Artificial Intelligence*, defined by Shapiro [Shap87] as *"The study of ways in which computers can be made to perform cognitive tasks, at which, at present, people are better."* The same author includes expert

tasks in AI as representation and inferences in order to diagnose diseases, design computer systems, etc., by processing relevant knowledge and creating search-based problem-solving as methods to take advantage of "knowledge." Therefore, AI is designed to solve problems. Since 1842 AI has been evolved over decades as cited by lady Lovelace [Love10], and first launched in 1950 by Alan Turing [Shap87].

On the other hand, Murphy [Murp12] defines *machine learning* as the *"set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty."* Similarly, Mitchel [Mitc97] states that a "*computer program is said to learn from experience E with respect to some class of tasks T, and performance measures P, if its performance at tasks in T, as measured by P, improves with experience E."* Kelleher [Kell15] coincides with the previous authors by defining machine learning as an automated process to extract patterns from data.

Murphy classifies Machine Learning in two major types of learning: Predictive or *Supervised* and Descriptive or *Unsupervised*, and one minor type, *Reinforcemen*t Learning. *Supervised Learning* (SL), utilizes an $x \rightarrow y$ mapping from input x to output y, where the representation in a *D-dimensional* vector of numbers representing *features, attributes* or *covariates*, SL is represented in equation 4.1.

$$D = \{(x_i, y_i)\}_{i=1}^{N} \tag{4.1}$$

where *D* is the *training set* and *N* is the number of training examples. x_i(input) may contain complex structured recipients that are regularly stored in an (*N* x *D) design matrix*, such as an image, a sentence, a time series, a graph, etc. The output is typically of two types: *Classification* (pattern recognition) or *regression*, both tasks are used in this study and detailed in chapter 5. In SL, *classification* has the goal to predict a class label, which is a selection from a predetermined list of possibilities.

In simple words, SL uses historical data, previously labeled (dependent variable), and a teaching algorithm to define a decision surface, thus resulting in prediction of the target value in a generalized way.

Unsupervised Learning (UL), is of the form z → y, where only the inputs are given. *D-dimensional* vector represents this technique in equation 4.2, where x_i and N represents the same case as equation 4.1.

$$D = \{(x_i)\}_{i=1}^{N} \tag{4.2}$$

The trainer does not exist in unsupervised learning. Instead, the algorithm must learn, by itself, to make sense of the data.

Reinforcement Learning (RL), involves an algorithm that interacts with the environment to learn how to act or behave in a certain state or situation, when the occasional signal is rewarded or punished. Therefore, a feedback loop exists between the learning system and its involvement employing a set of inputs and some outputs which are graded as a training data [Good16]. Equation 4.3 represent this technique.

$$D = \{(x_i, y_{ie}, y_g)\}_{i=1}^{N} \tag{4.3}$$

where, y_{ie} is some output interaction with environment, and y_g a grade received as optimum output. where x_i and N represents the same case as equation 4.1. RL is not discussed in this study for simplicity and time constraints.

ML algorithms extract knowledge from an example (i.e. data), learning by the example, and generalizing this learning properly to deal with new examples [Beng11]. Machine Learning (ML) is used to generate predictive models that can be exploited to forecast an event (i.e. hurricane scale, demand forecast) or classify a task. Flach [Flac12] focuses on the model as a central part of machine learning, figure 4.1 shows this concept.
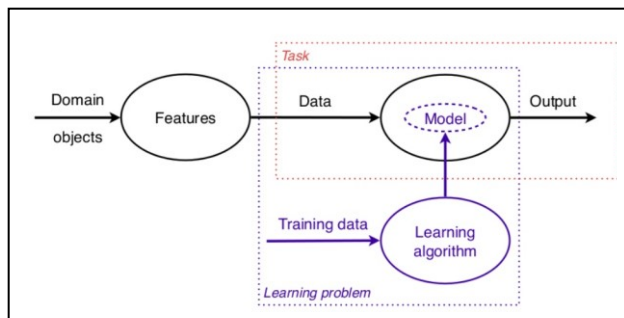
Figure 4.1: Main components of Machine Learning [Flac12].

Many techniques have been used to solve ML problems. The most common are depicted in Figure 4.2 [Math18b].
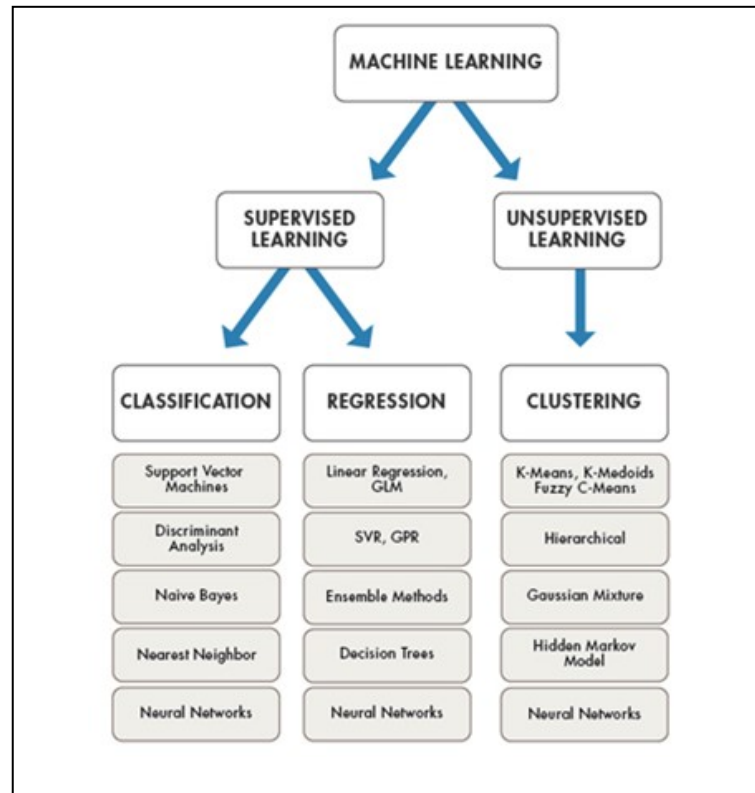
Figure 4.2: Machine learning assortment models, divided by supervised and unsupervised characteristic [Math18b].

4.1.2 NEURAL NETWORKS

Neural Networks are one type of model for machine learning as shown in figure 4.2. Deep Learning a subfield of Machine Learning, which was inspired by the analogy of *biological neural networks*. These networks have been used by many scientists, in many applications such as computer vision, natural-language processing, and generative models, with a rise in usage because of the dramatic improvement of the computational resources. A brief description of this kind of architecture is described next, in order to understand the architecture of the neural networks.

Since 1943, the McCulloch-Pitts model [Mccu43] was a seminal work, as the model resembles the structures in the human brain. The neuron is the main unit of the human brain, which contains approximately 100 Billion neurons with about 6,000 connections from each neuron to other neurons [Budu17]. The aim of the neurons is to optimize the information received from other neurons, processing this information in a distinctive way. The basic artificial neural network (ANN) structure consists of basic units called *nodes*, which simulate the operation of a neuron within a neural network. The operation is similar to its biological counterparts, where the activation occurs when the sum of the total input signals exceeds the activation threshold. The nodes transmit the signals between them by mean of *connections*, which simulate the operation of biological synapses. These signals, sent by a neuron, passed from one to the next, can behave as a filter. Each neural network has its own number of nodes to receive the input signals from the outside, where the first group of nodes represents the first layer, called the *input layer*. The second group is an intermediate position the neural network, and is referred to as a *hidden layer*. The last layer is called the *output layer*, which sends the results directly to the outside [Nell18]. In a broad point of view, the basic operation of a neural network is the flow of data that enters the neural network from left to right, by processing the data in a more or less complex way depending of its structure, and finalizing in an output result. Figure 4.3 describes the function of the structured biological neuron, and Figure 4.6 shows a deep neural network.
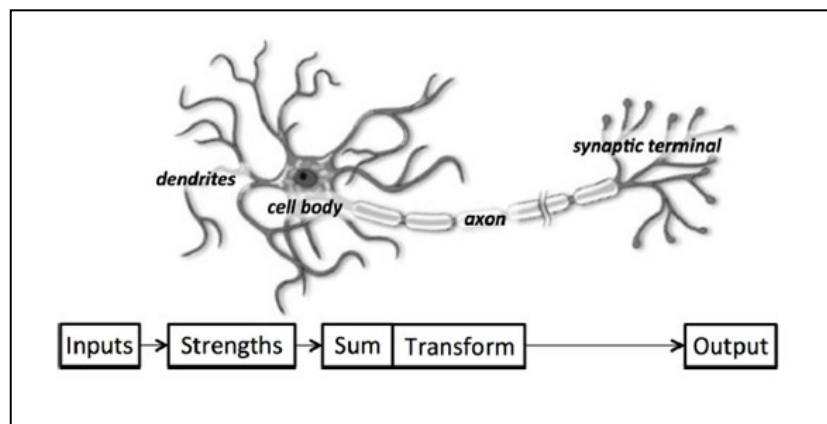
Figure 4.3. Biological neuron and its resemblance to an artificial neuron [Budu17].

41

The relationship between neural networks and machine learning is that we have a neural network (the process of defining the model) instead of a model, and the learning rule instead of machine learning [Kim17]. Figure 4.4 shows both processes.
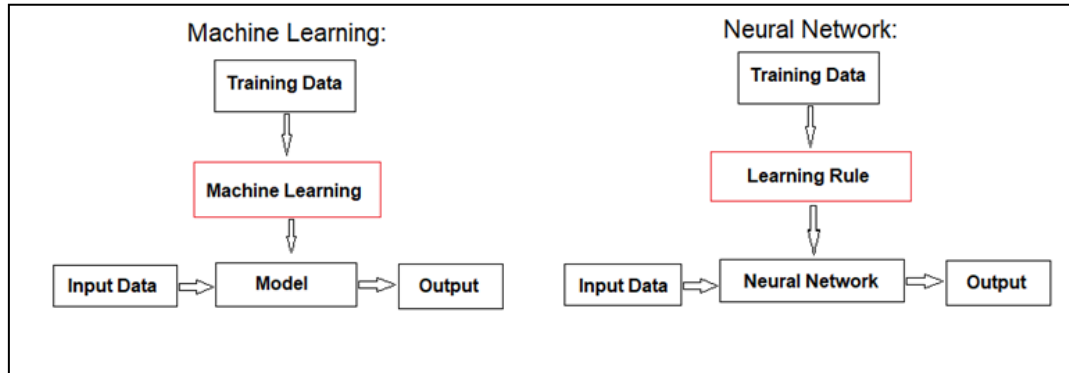
Figure 4.4. Relationship between Machine Learning and Neural Networks [Kim17].

The artificial neural network, is represented by Candel [Cand17] a single neuron as shown in Figure 4.5. The nonlinear activation function is represented by **f,** and the neuron's activation threshold is represented by **b** (bias).
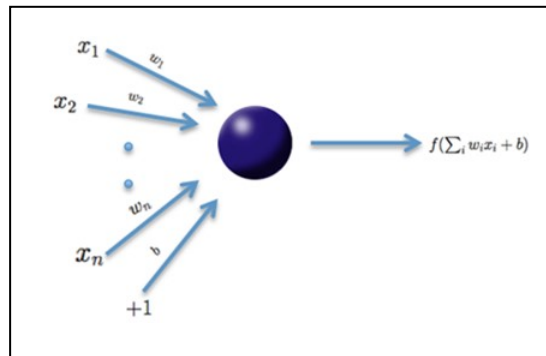
Figure 4.5: Representation of a single neural network [Cand17].

The same author represents the basic unit modeled by the Equation 4.4, where $f(\alpha)$ is the output signal transmitted by the linked neuron.

$$\alpha = \sum_{i=1}^{n} w_i x_i + b \qquad (4.4)$$

4.1.2.1 Connection weights, Biases, and Neuron Activation Function

Connection weights as discussed before are mostly represented by a "*w*" as a mathematical representation of a neural network, in which lines/arrows connected from one point to another point in the direction of information flow: input signal given a weight on the connections as a coefficients scale to amplify or minimize the input [Patt17].

Biases are generally noted as a '*b*", which are scalar values added to the input to ensure that at least a few nodes per layer are activated regardless of the signal strength throughout the learning process, allowing the action in the event of low signal by giving a try of new interpretation or behaviors to the network [Patt17].

Activation function is the function which governs the artificial neuron's behavior, by transforming the inputs, weights, and biases into a convenient range (0 to 1 or 1 to -1) to the next node layer which acts as an input. This transmission of that input is known as a *forward propagation*. The activation of the artificial neuron occurs when the neuron passes a nonzero value to another artificial neuron [Patt17]. There are three main of neuron activations in Neural Networks that allow nonlinearity in the data, these are: (1) **Logistic Sigmoid Function.** The S-shaped property of this function allows it to work with probability distribution in deep learning, and is commonly used [Good16]. (2) **Hyperbolic Tangent (tanh) Function**. This is an S-shaped nonlinearity, similar to the sigmoid function, but is rescaled and shifted with symmetry around 0 (ranged from -1 to 1), allowing faster convergence of the training algorithm [Weis18]; and (3) **Rectified Linear Unit (ReLU)**. This neuron activation function has the shape of a hockey-stick, which is one of the most popular for authors because it can be used in many tasks, especially in computer vision [Budu17].

4.1.3 THE LEARNING PROCESS

The learning process of a neural network is named the learning phase, as discussed before the weighs of the synapses "*w*" is slightly modified in each predetermined cycle of operation in the neural network. Each learning cycle is named as an "epoch." The most well-known algorithm,

is the *back-propagation algorithm* which is a gradient-based algorithm with numerous variations (depending on data), and the *Levenberg-Marquardt algorithm*, which is commonly more efficient but with a higher computational cost (memory). In order to perform learning in the neural network, the result must be evaluated. It is necessary to use the appropriate input data to achieve some confidence by use of a random mechanism like *simple hold-out validation* or *k-fold validation*, to separate the experimental data into tree nonoverlapping and independent data sets [Izen08]. Simple hold-out validation, sets apart a fraction of the data to form the test set. Training (or learning) is accomplished with the remaining data, so evaluation takes place on the test set. K-fold on the other hand, splits the data into *K* partitions of equal size [Chol18]. The splitting of the available data into three sets to perform the model are: 1) *Training or Learning data set*, 2) *Validation data set*, and 3) *Test data set*. In the **training set** for each input value an output value is expected, where all the optimization activities are guided by the performance index. The **validation set** is used for model selection, by comparing the output values produced by the neural network with the expected ones, and by monitoring the learning process which could present a different tendency. The **test set** is used to assess the performance of a completely specified final model, by the analysis of another set of inputs whose results are known in a supervised learning, and some performance measures are computed. The evaluation of the differences between the values obtained and the ones expected defines the ability of the neural network to solve the problem, this value is mostly presented as the accuracy of the neural network [Nell18].

The developing of a learning model involves the tuning of the neural network configuration, by choosing the numbers of layers or the size of layers in the network, this is called *hyperparameters* of the model. The process of tuning is by using the feedback signal, which is the performance of the model on the validation data [Chol18]. ML and DL have progressed and are introducing additional learning algorithms. To wit, at the beginning of this study, the number is so great that not all can be discussed in this document.

4.1.4 DEEP LEARNING NEURAL NETWORKS

ANN Learning dates from 1940's, but since 2006, it has become popular because the advent of computational power coupled with the increment of massive data. Most recently, a variation referred to Deep Learning has been adopted. In addition, DLNNs are a variation of ANNs known by several different names, such as *feedforward deep network, feedforward neural network*, and *multilayer perceptron (MLP)* [Good16]. Deep Learning Neural Networks (DLNN) are the embodiment of Artificial Neural Networks (ANN) using Deep Learning Algorithms, as seen in section 4.1.2. DLNN have a distinct structure from the previous neural network, as it contains several layers of interconnected single neuron units, resulting in multiple hidden layers. The name "deep" was inspired by the use of deep structures, which resemble the biological brain [Good16]. Subsequently, a family of algorithms known as neural networks has recently seen a revival under the name "*Deep Learning.*" Deep Learning Neural Networks is a Learning technique which is getting a lot of attention recently by authors like Babri [Babr96], Villmann [Vill17], and Angelov [Ange17], used to approximate some function for *classification* and *regression*, since its models can achieve state-of-the-art accuracy with substantial computing power reduction in training time. One objective of the work in this document is to incorporate this approach.

Deep Learning models have the capability to learn directly from the data set, where the training process is achieved by using a large set of labeled data and neural network architectures containing many layers. DLNN structure model begins with an input layer, followed by multiple nonlinear layers which include the bias units in each layer, and includes an output layer (linear regression or classification layer). Figure 4.6 shows a relatively simple DL neural network known as multilayer perceptron (MLP) also known as (Vanilla) feed-forward neural network [Mull17], showing weights and biases where the superscript n, w, and θ refers to the first or second layers [Koiv18]. MLP can be seen as the generalization of a linear model that perform several stages of processing as a way to derive a solution (classification or regression).
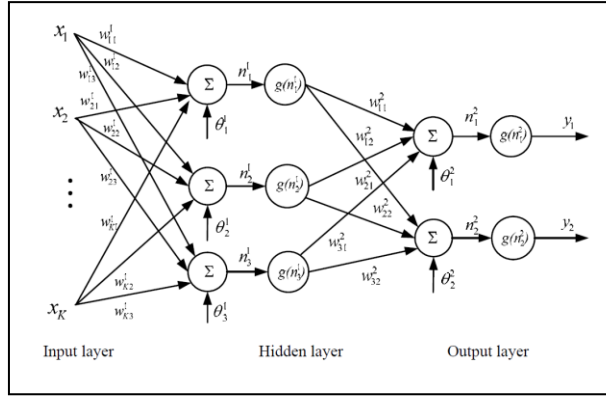
Figure 4.6. Deep Learning Neural Network or MLP representation [Koiv18].

The learning process occurs when the weights are adjusted by minimizing the error in the labeled data, representing the objective function (sometimes referred to as the cost function or error function). This learning process is represented as equation 4.5 [Cand17]:

$$L \ (W, B \mid j), \tag{4.5}$$

In this study we used the structure of the MLP neural network, presented in chapter 7. Thus, for better understanding Figure 4.7 shows a single node in an MLP neural network. Here, the *neuron* is represented by *i* with a summer and a ***nonlinear activation function g***. The ***weights w_{ki}*** are multiplier with the ***inputs*** x_k , **k=1 ,…, K** in the neuron, and summed all together with the ***bias θ_i***, next, n_i is the ***input*** to the ***activation function g***.
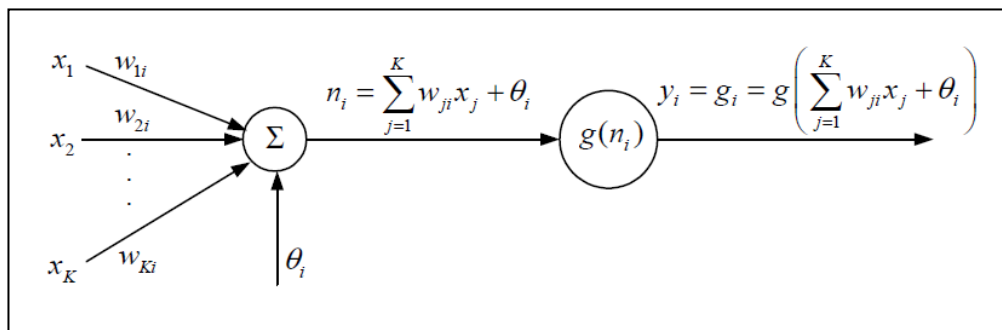
Figure 4.7. Single node of an MLP neural network [Koiv18].

The output in **node i** is represented by the *output y_i*. Equation 4.6 shows this representation. Subsequently, Equation 4.7 represents the connections between the several parallel and series nodes which at the final results the structure form of the MLP neural network [koiv18].

$$y_i = g_i = g[\sum_{j=1}^{K} w_{ji} x_j + \theta_i] \qquad (4.6)$$

In the equation 4.7 below y_i represents the **output** of the **MLP neural network**.

$$y_i = g[\sum_{j=1}^{3} w_{ji}^2 g(n_j^1) + \theta_j^2) = g(\sum_{j=1}^{3} w_{ji}^2 g[\sum_{k=1}^{K} w_{ki} x_k + \theta_j^1] + \theta_j^2 \qquad (4.7)$$

The conclusion is a nonlinear parameterized neural network that maps the data from the input space $x \in R^K$ to output space $y \in R^m$. This particular structure has an arrangement of n = 3 layers. including the parameters of the weights w_{ji}^k, biases θ_j^k, and the activation function g, which is assumed to be the same in each of all the layers. Thus, the design for the utmost performance of an MLP neural network or other types of neural networks, is a data fitting problem. Starting with a given input-output data set of (x_i, y_i), $i = 1, ..., N$ the objective is to determine the parameters of (w_{ji}^k, θ_j^k) respectively, during the learning process. The learning process follows several needed steps as discussed in the section above, and this study follows several steps in approaching the problem in the next chapters. The basis of DLNN is the employment of several layers of nonlinear processing units arranged in a successive manner, where the output from a previous layer is, in turn, the input to the subsequent layer. The simple recipe of a deep learning algorithm is the combination of a dataset, a cost function, an optimization procedure, and a model [Good16]. The number of hidden layers can be settled by trial and error method [Cios07]. Although other methods can be used to decide the number of hidden layers, those will not be covered in this dissertation for simplicity. Thus, by adding more layers, a deep network can characterize complex learning functions (i.e. language, vision, etc.), resulting in the mapping of the input (raw data) to the output (predictions) directly from the data. In simple worlds: "*a chain of simple, continuous geometric transformations mapping one vector space into another*" [Chol18].

Research groups like Toronto, Microsoft, Google and IBM, as presented in [Hint12], observed the capability of deep learning in acoustic modeling. Similarly, deep learning has been demonstrated to be an excellent solution for managing applications such as speech recognition

reported by authors Cui et al. [Cui16], Mateju et al. [Mate15], Liao et al. [Liao13], and Dahl et al. [Dahl12], natural language progressing reported by Galea et al. [Gale18], and computer vision analyses reported by Yang et al. [Yang17b]. Other fields derive benefits from the use of deep learning, such as the management of big data applications, as proposed by Wang et al. [Wang16b], where a multilevel deep learning model is proposed for the analysis of system stability and emergency management.

4.1.5.1 Deep Learning Neural Networks Structure

The neural network spotlighted in the previous section was a simple one. In fact, there are many types of neural networks, used for different tasks. The following is the taxonomy of deep learning represented by [Goll16]
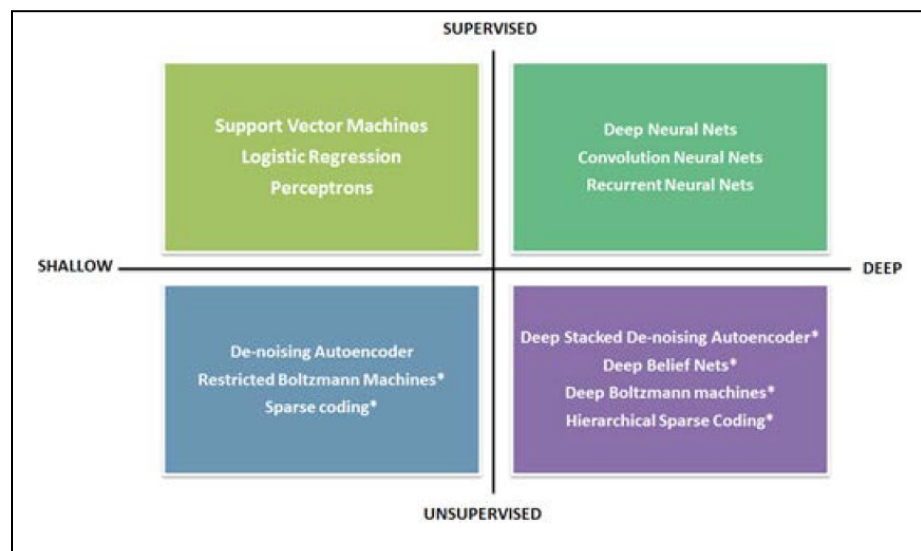
Figure 4.8. Deep Learning Neural Network taxonomy, taken from [Goll16].

A more complex MLP is represented by LeNet architecture, which was the "first architecture" for Convolutional Neural Networks (CNN) created by Yann LeCun [Lecu98]. It consists of two sets of convolutional, activation, and pooling layers, followed by a fully-connected layer, activation, another fully-connected, and finally a softmax classifier. Figure 4.9 shows a CNN used with the original LeNet model [Lisa15].
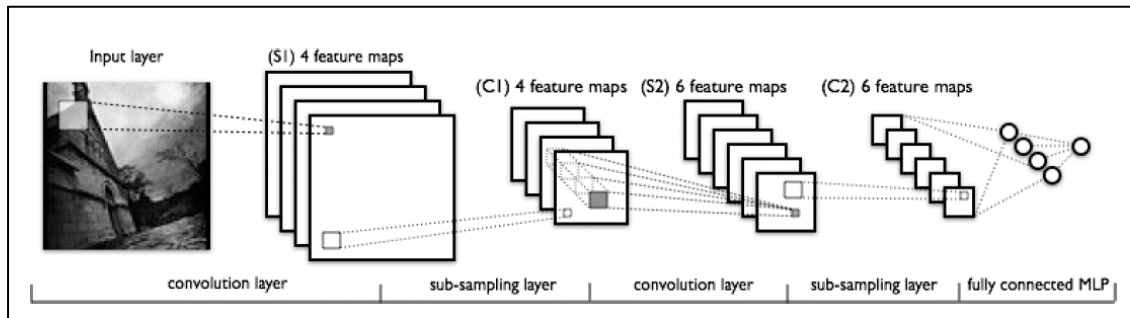
Figure 4.9. CNN used with LeNet model architecture, taken from [Lisa15].

4.1.6 MAIN CHALLENGES IN MACHINE LEARNING AND NEURAL NETWORKS

Two main challenges in learning techniques are: Optimization and Generalization. These challenges are briefly discussed next.

Optimization. This encapsulates all problems of obtaining the best item or best value among a set of alternatives by minimizing (or maximizing) an *objective function* (or criterion) via an iterative process. This function can be linear, quadratic, or more complex. Stochastic gradient descent (SGD) is one popular technique used for the optimization problem [Shal14]. Thus, as stated by Goodfellow [Good16] "*Optimization* refer to the process of adjusting the model to get the best performance possible on the training data".

The critical points that are neither maxima nor minima, are referred as *saddle points*. Figure 4.10 depicts the gradient descent optimization technique, with different types of critical points and multiple local minima. Image taken from Goodfellow [Good16].
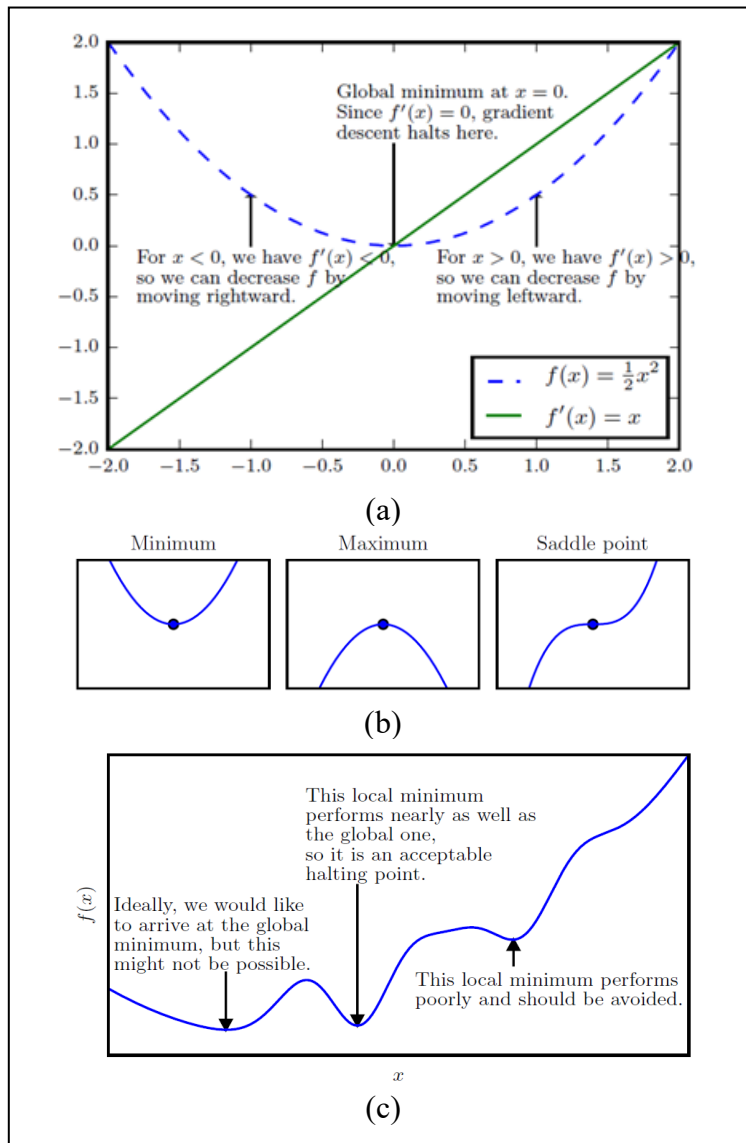
Figure 4.10. Optimization based. (a) Gradient descent algorithm, (c) Types of critical points, (b) Approximate minimization for multidimensional function [Good16].

Generalization is when an algorithm performs well on new or previously unseen (or un-observed) data inputs. This is exhibited when the machine learning model is trained with a set of training values. During this process, training errors are identified as *generalization error* or *test error*, and used to reduce the overall error. When the model begins to degrade after a number of iterations (epochs) on the training data, it said that the model is starting to overfit.

Two factors define the performance of a machine learning algorithm [Good16]: 1) Small error training (underfitting), and 2) Small gap between training and test error (overfitting). Figure 4.11, shows these two factors [Good16].
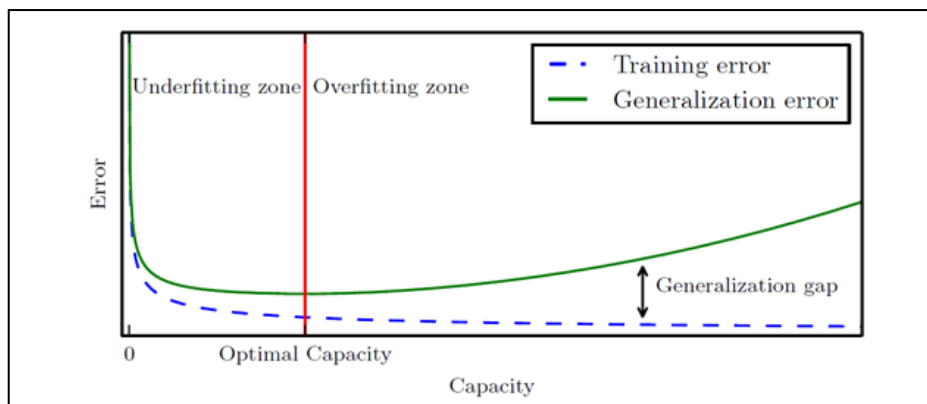
Figure 4.11. Relation between underfit, overfit and optimal capacity [Good16].

A technique to test overfitting is *validation* of a part of the training data (by dividing the training data from the validation set), and used to observe the performance of the model. There are many other techniques used by researchers [Salm15], such as *cross-validation* (randomly selection of the validation dataset) [Murp12]; confusion matrix based on: true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), and false positive rate (FPR), which convert the raw numbers from the confusion matrix into percentages; Precision, Recall, and F_1 measure, which can be calculated directly from the confusion matrix [Kell15].

The process of preventing overfitting is called *Regularization*. There are many training error techniques, such as *Regularization* L1, and L2 also called *weight decay* [Good16]. Authors like Shen [Shen18] and Yu [Yu10] use dropout and pretraining, respectively, to help in the model optimization and generalization. The following list is the most adopted methods for generalization in the literature [Kepl17], including the ones mentioned before:

- Adding more training data (*Data Set Augmentation*),
- Choosing a specialized model (*Bagging* or *bootstrap aggregating*),
- Lowering model complexity (*Norm penalties* as constraint optimization),

- *Pretraining,*

- *Early stopping,*

- *Weight decay, and*

- *Dropout.*

4.1.6 DATA PREPROCESSING IN DATA MINING

This section is dedicated to the importance of the preprocessing in data for machine learning and deep learning neural networks, and as discussed in section 1.5.3 and section 3.1 KDP workflow data preprocessing or data preparation as many authors call it, is the second step before the model development in data mining. Data preprocessing is also known, by many authors, as *feature engineering*. The unsupervised transformation of the dataset are the algorithms that create a new representation of the data which might be easier for machine learning or deep learning neural networks. This can define any data transformation or analytics that extracts information from a raw dataset which may be useful in a modeling context. Data preprocessing also includes vectorization, normalization, handling missing values and feature extraction as mentioned before [Chol18]. Some of the algorithms have the task of *dimensionality reduction*, which takes a high-dimensional representation of the data, containing many features to a new way of representing the data, which could summarize the essential characteristics with fewer features than before. Other algorithms, on the other hand, have the task of partitioning the data into distinct groups of similar items. For example, the *clustering algorithm* carries out this particular task. Figure 4.12 shows this concept taken by Thom [Thom17].
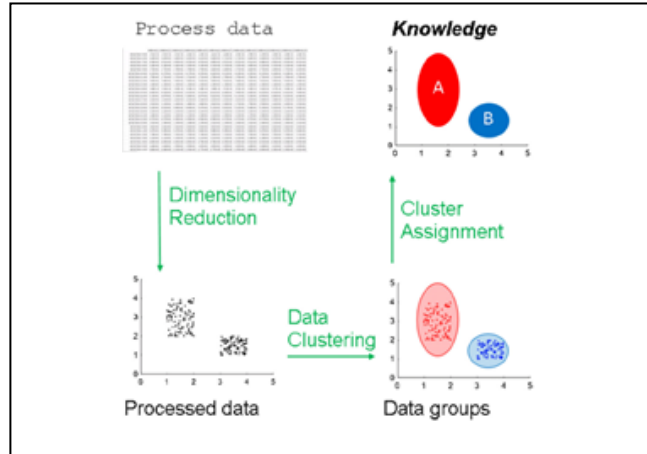
Figure 4.12. Data Mining algorithm approach in a KDP [Thom17].

In the case of deep learning neural networks, extracting the features completes automatically, so the net can learn all features in one pass rather than being adjusted by human intervention. This approach allows the use of less resources and solves a problem with less data [Chol18].

In conclusion, data mining is not a best method that fits all data, but the selection depends on the size, type, and perception of the project and how it will be used. This could be in many cases, by trial and error method. A list of its analysis purposes [Mcki18] is shown below:

- *Classification*: SVM, nearest neighbors, random forest, logistic regression, etc.;
- *Regression*: Lasso, ridge regression, etc.;
- *Clustering*: k-means, spectral clustering, etc.;
- *Dimensionality reduction*: PCA, feature selection, matrix factorization, etc.;
- *Model selection*: Grid search, cross-validation, metrics;
- *Preprocessing*: Feature extraction, normalization.

4.2 Conclusion

Data Mining Algorithms are implemented by the use of Scikit-learn toolkit module for Python machine learning [Thom17], which contains an extensive selection of standard supervised and unsupervised machine learning methods. Regardless of the importance of feature engineering

in Machine Learning, this research work does not cover the analysis of the evaluation of features in data, only the discussion of feature selection to provide a measure of importance for each feature, for simplicity. On the other hand, deep learning neural networks makes problem-solving much easier by automating feature engineering and, in the case of the need to manage larger databases, also, the multi-layer approach allows the learning of the model at the same time by the joint representation of the layers. At the time of this study, the above methods, algorithms, neural network architectures, and tools, suggest a better performance on many problems. Thus, from these observations, this work uses the approach discussed previously.

CHAPTER 5: GENERIC EXPERIMENTAL FRAMEWORK MODEL

This chapter discusses the development of an experimental KDP framework model adapted to solve the main objective discussed in section 1.6.2 of this dissertation, based on two domains:1) The critical operation of the ETDS systems under an extreme-weather event with concepts discussed previously in chapter 2; and 2) the KDP framework with concepts discussed previously in Chapter 3 and 4. From the previous background concepts, a comprehensive generic framework based on KDP is proposed. Figure 5.1 shows the suggested methodology in this dissertation study to incorporate both domains.
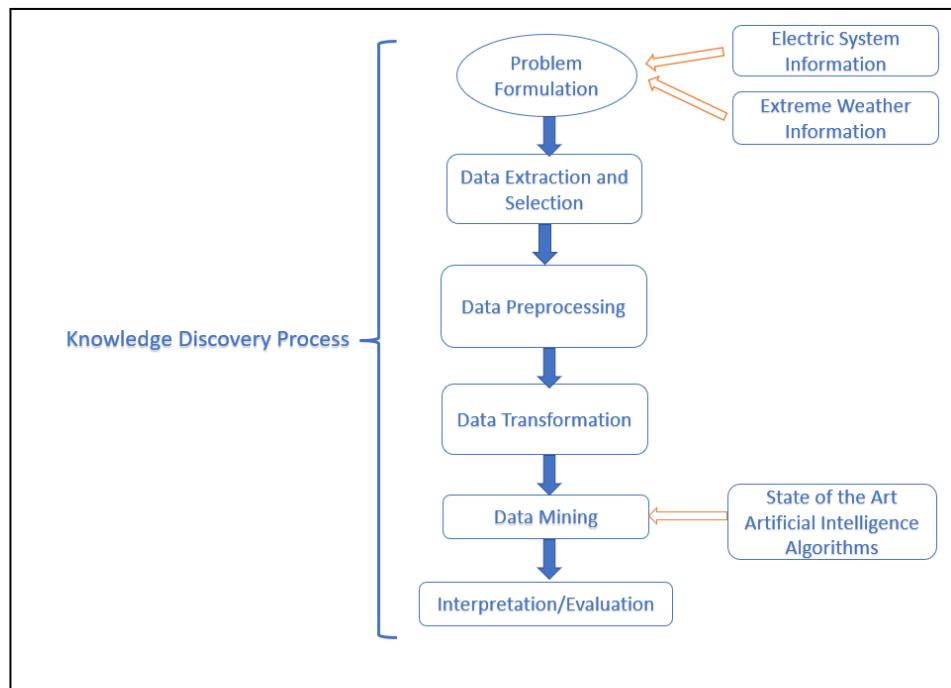
Figure 5.1: Generic framework based in KDP to incorporate both domains.

CHAPTER 6: SCENARIOS-1 EXPERIMENTAL FRAMEWORK MODEL

This chapter discusses the development of a KDP framework from the assumption in chapter 5, adapted to solve the problem discussed previously in section 1.6.2. Discussion in Chapters 3, 4 and 5 serve as the foundation of the work described in this chapter. Here, a novel methodology framework to assess the critical operation of the ETDS under an extreme-weather event is detailed. The goal of ***Scenario-1*** is to verify the feasibility of the proposed method, focused on two data analyses of a short-term electricity market price and demand consumption from New York city. The **two analyses** are based on: **1)** *De-noised wavelets and NARX neural network*; and **2)** *A complete machine learning data pre-processing analysis, and evaluation of standard algorithms by improving the results on the Machine Learning model*. The experimental framework is presented first for both analyses by the following development, analysis, and observation during each step of the proposed framework.

6.1 Experimental Framework model Scenario-1 analysis-1.

Figure 6.1 shown the methodology used to incorporate both domains in Scenario-1. Only one model was used in this scenario to assess the impact of an extreme-weather event in the ETDS based in a KDP framework.
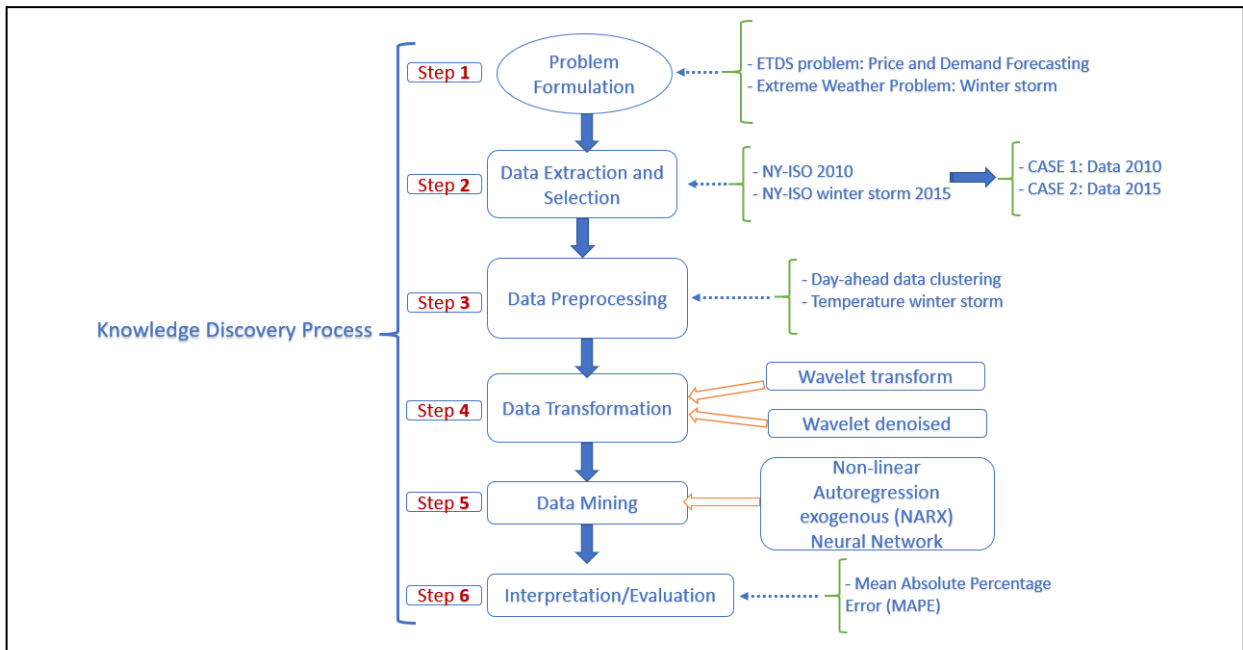
Figure 6.1. Experimental framework model Scenario-1 analysis 1.

6.1.1 SCENARIO-1 (S1) ANALYSIS -1 AND 2: WINTER STORM IMPACT IN NEW YORK AREA

The characteristic of this scenario is focused in the influence of a **winter storm** in the electricity market price forecasting in New York City.

Task:

a) Short-term price and demand prediction under winter storm. The main objective is to develop a price and demand prediction during a winter storm in New York city, which occurred on February 2015 as Case study-2; and a price and demand prediction of a previous year (2010) as Case study-1.

6.1.2 STEP 1: PROBLEM FORMULATION

In this step, both analyses were focused on a winter storm, reported during the year of 2015 in the geographical region of New York, as discussed previously in section 1.6.4 of this dissertation.

Due to the implication of the non-linearity, high volatility, and seasonality factors in the electricity price forecasting, and the demand uncertainty during an extreme-weather event, these factors need to be addressed with accurate forecasting models to minimize cost of wholesale power by serving proper loads during the critical operation of the system under stress [Vill19]. Thus, scenario-1 has two targeted analyses. Analysis-1 includes two case studies to verify the feasibility of the proposed experimental KDP framework shown in Figure 6.1. Analysis-2 is presented in section 6.2, which includes a predictive regression analysis.

6.1.3 STEP 2: DATA EXTRACTION AND SELECTION

In this step, data extraction and selection are explained next.

6.1.3.1 Data Extraction

Data extraction is from the NY-ISO databases, which have vast amounts of data, containing 8760 entries, with high quality numerical values for prediction purposes and represents the system perfectly with the best data sources. The second step of the KDP is data extraction, thus, open-sources of electricity data information were selected from utilities, electricity market, and the United States Energy Information Administration availability. On the other hand, the meteorological events are selected based on the geographical path of their passages of the storm information availability. Data collection was acquired from two data sources: *benchmark data* obtained from hurricane or winter storm based in historical computer modeling simulation, and *hypothetical data* obtained from power flow and microgrid analysis simulation based on power system specialized software. Appendix A shows a data acquisition flow for the data used in this scenario.

6.1.3.2 Data Selection

The data collection is based on information from the New York Independent Service Provider (NY-ISO) web site [NYIS19], and temperatures data is based on information from the Network for Environment and Weather Applications web site [NEWA19]. Both load and price data are based on a time sequence from the two case studies carried out in preparation for this scenario. The time horizon for point of prediction is shown in Figure 6.2 and Figure 6.3. The electricity price and electricity demand consumption for 2010 and 2015 was used for this purpose in [Vill19]. Three evident fluctuations are observed in the months of January, August, and October in 2010; and two fluctuations in winter and summer during 2015. Fluctuation occurring during 2010 are attributed to the occurrence of high temperatures mostly during the months of June-August [Weat19]. Similarly, high extreme temperature fluctuations during the month of August in the same year directs a considerable variation in the electricity consumption levels. On the other hand, a winter storm occurred in New York area in February of 2015 and the high extreme fluctuation during the month of August in the same year, directs fluctuations in the electricity consumption levels.
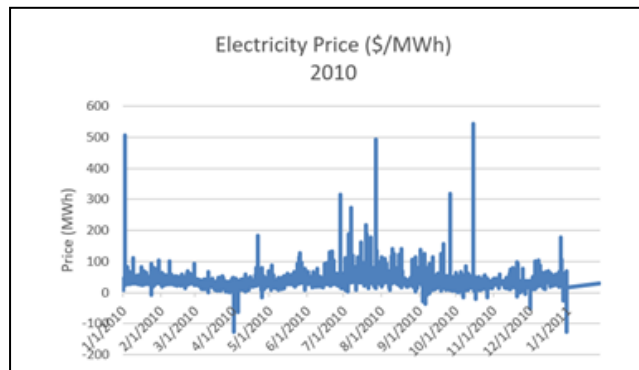
Figure 6.2. NY-ISO electricity price in 2010.
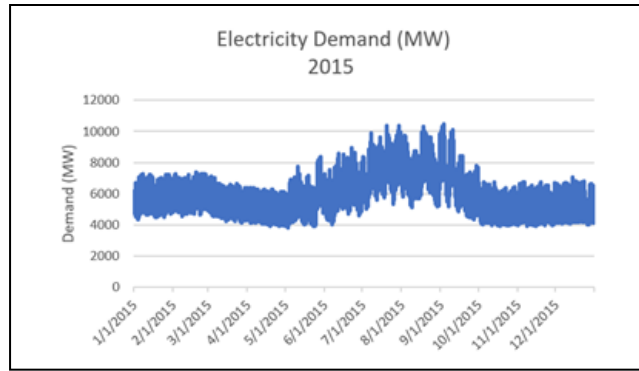
Figure 6.3. NY-ISO electricity demand in 2015.

6.1.4 STEP 3: DATA PRE-PROCESSING

In this step, missing values and outliers was removed. Similarly, data was prepared in tabular format, suitable to be used for the methods of analysis in step 4 and 5. Data from previous steps was used in two case studies: *Case 1*, was a price forecasting case study using data from one year (01- 01-2010 to 12-31-2010), which is a short-term forecasting based mainly on an hour-ahead prediction. Price, in this study, is based on market clearing price (MCP), which is the price established by the ISO, in this case NY-ISO. MCP establishment is based on three factors, supply and demand, transmission congestion, and market rules. *Case 2*, was a demand forecasting using 168 data points. In this study, the demand forecasting takes into account a winter storm from a cold wave that occurred during February of 2015 in New York city. This case study was used as proof of an extreme-weather demand forecasting application. Appendix A.1.1 shows a workflow of data preprocessing with a portion of the data sets for both cases.

6.1.4.1 Feature Engineering

Assuming the various factors that affect electricity consumption and electricity price, we employ feature engineering to identify the factors that affect both of them, but without taking into account economic factors, pollution or air quality factors like author Gou et al. [Gou18] report, for simplicity of this study. We split the data only into two categories: 1) Date-related factors (weekday, weekend, and Festive day), and 2) Weather-related factors (daily temperature, winter storm day report).

60

6.1.5 STEP 4: DATA TRANSFORMATION

In this step, data was transformed from time domain to a wavelet signal. Wavelet denoising was used to normalize price non-linearity to make different samples seen more similar to each other. The selection of the correct wavelet representation of the original signal is important for effectively achieving of the desired results. Wavelet transform technique consists of the time-frequency decomposition. This technique allows the identification of hidden trends in the signal by the processing of the time series main frequency component, and the abstraction of local information. The wavelet decomposition is based on two types of filters: 1) The low-pass filters, which correspond to the *approximated series* "a_n", and 2) The high-pass filters, which correspond to the *detailed series* "d_n". Figure 6.4 shows an example of a wavelet Daubechies type 3 with the low and high frequency representation, containing three coefficient levels decomposition, illustration taken from the work of Nazaripouya and lightly modified for better understanding [Naza16].
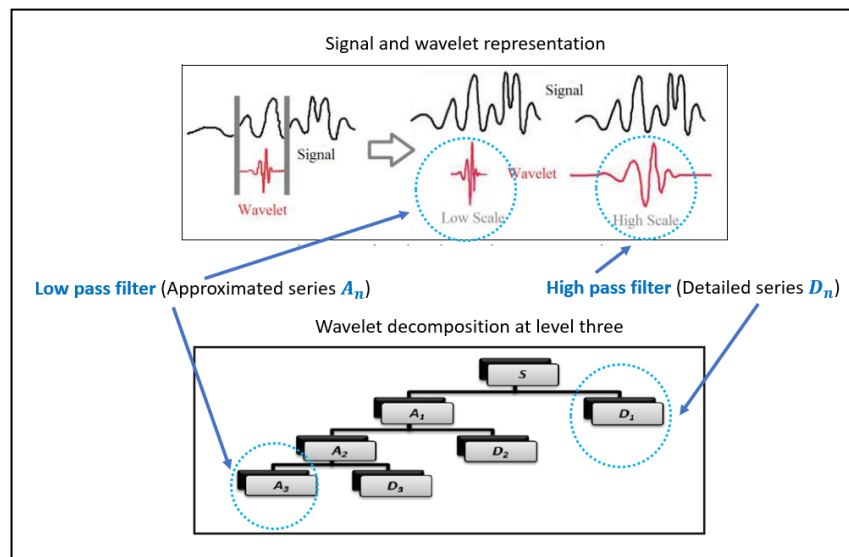
Figure 6.4. Example of a wavelet Daubechies type 3 with the low and high frequency representation, image lightly modified from [Naza16].

In *Case-1,* because of the nature of electricity pricing data was analyzed as a signal, the chosen wavelet was the Daubechies type 3 with 7 coefficient levels. Figure 6.5 and Figure 6.6

show the wavelet decomposition (approximation and coefficients) for fall and winter seasons, respectively.
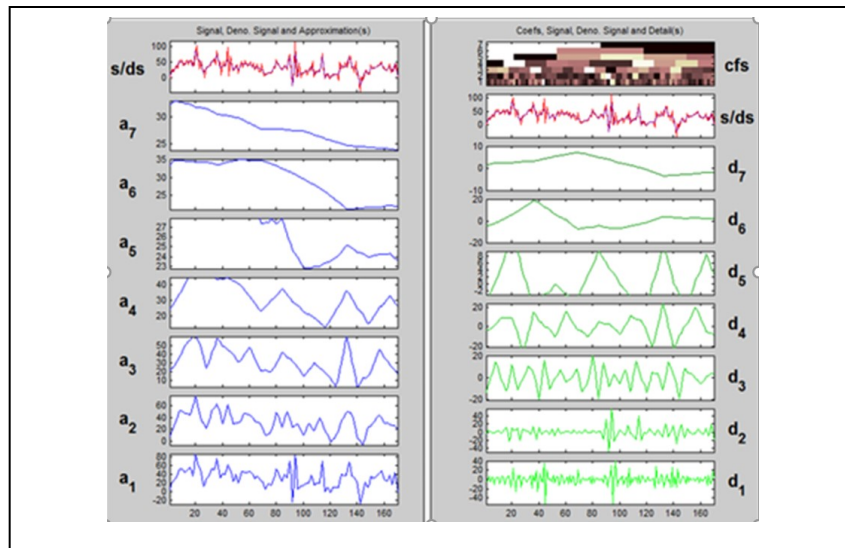
Figure 6.5. Fall electricity price wavelet decomposition signal: approximations (blue), coefficients (green).
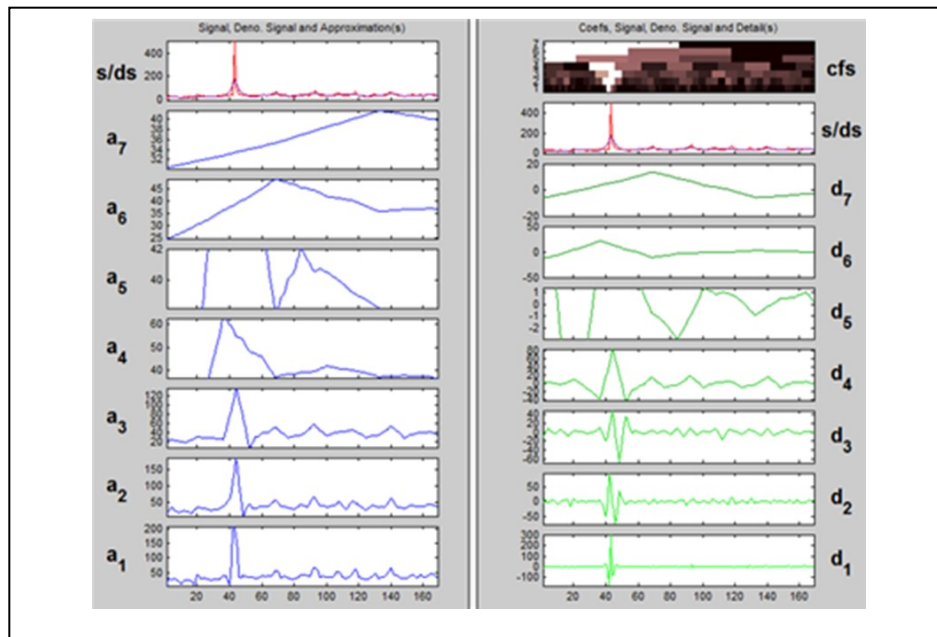
Figure 6.6. Winter electricity price wavelet decomposition signal: approximations (blue), coefficients (green).

The wavelet de-noising setting and selection procedure can be described with the following steps:

1) Use a wavelet transform to reconstruct the signal (price) from a noisy signal (original price). In this study, the chosen level was 7.

2) Select the appropriate threshold limit at each level to remove the noise. This study uses soft thresholding to smooth the signal, preventing loss of important features.

3) Use an inverse wavelet transform of the thresholder wavelet coefficient to obtain the de-noised signal.

4) Finally, electricity price emerges from the previous step as preprocessed data, ready to be the input on the Neural Network.

In Case-2, because of the nature during a **winter storm** the electricity demand data was analyzed as a signal, the chosen wavelet was the Daubechies type 3 with 3 coefficient levels. Figure 6.7 shows the wavelet decomposition (approximation and coefficients) for the winter seasons.

The wavelet de-noising setting and selection procedure can be described with the same steps as Case-1. All the parameters were normalized by the same procedure but with different data (Demand).
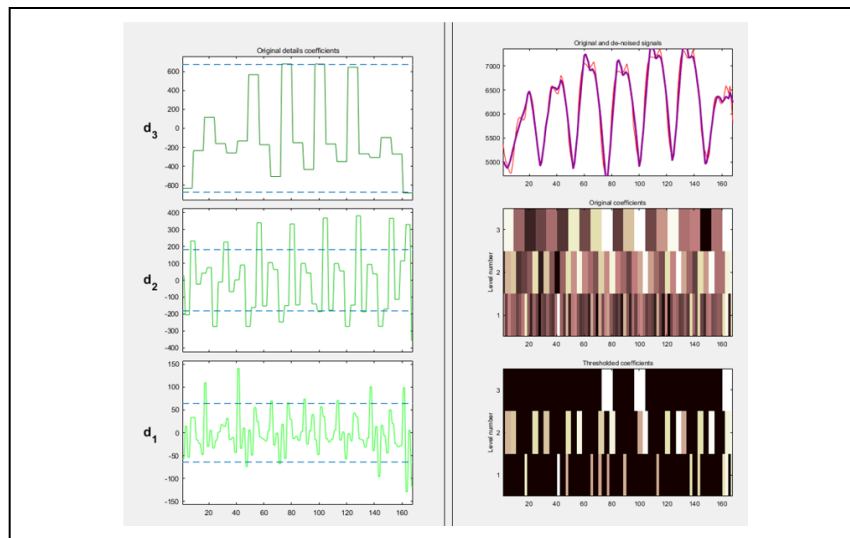
Figure 6.4. Winter electricity demand wavelet decomposition signal: approximations (blue), co-efficients (green).

6.1.6 STEP 5: DATA MINING

In this step, the data mining model setting and training was developed. A supervised learning Non-linear Auto-regression eXogeneous Neural Network (NARX) was used, with feedback connections to several layers of the network, and one-layer delayed feedback.

6.1.6.1 Supervised Learning-based NARX neural network for electricity consumption forecasting

A block diagram of the proposed short-term electricity market price and demand forecasting based in de-noised wavelets and NARX neural networks framework is shown in Figure 6.7.

The implementation of the NARX model in this study uses a multidimensional input as shown in Figure 6.7, employing the feature engineering from step 3: actual price (P_t), actual demand (D_t), week-ahead price temperature (TP_{t-168}), week-ahead demand temperature (TD_{t-168}), week-ahead (168 minutes) electricity price (P_{t-168}), and week-ahead (168 minutes) electricity demand (D_{t-168}). Output value of the NARX is the forecasted electricity price (P_{t+168}) and the forecasted electricity demand (D_{t+168}) which is the estimated output of the nonlinear dynamic system.
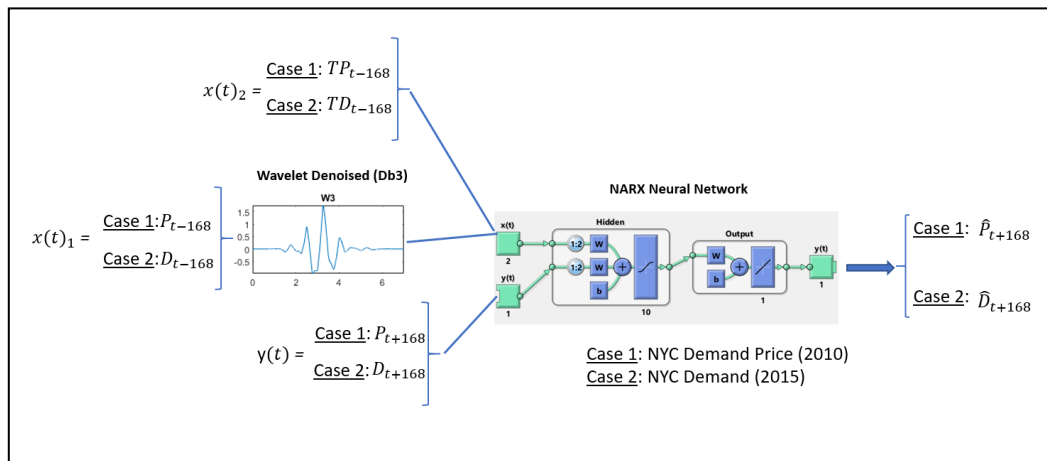
Figure 6.7. Proposed short-term electricity market price and *demand forecasting* based in *denoised wavelets and NARX neural networks framework.*

The neural network for the two cases was implemented in the same manner with 3 inputs, 1 output, and 10 hidden neurons. The training method used was the Levenberg-Marquardt.

6.1.6.2 Supervised Learning-based NARX neural network typical representation

The NARX model is represented by Siegelmann [Sieg97] in equation 6.1.

$$y(t) = f(u(t - n_{u)}, \dots, u(t - 1), u(t), y(t - n_y), \dots, y(t - 1) \qquad (6.1)$$

where function *f* is a nonlinear function, *u(t)* and *y(t)* represent the input and output of the network at time *t*. Additionally, n_u and n_y are the input and output order, respectively.

6.1.7 STEP 6: INTERPRETATION / EVALUATION

In this step, the interpretation is shown first and then the evaluation of the de-noised wavelet for Case-1 and Case-2.

6.1.7.1 Interpretation

All seasons values in Case-1 and winter season values are obtained in the same manner. The threshold selection rule was soft in both cases, with rigorous SURE for case-1 and fixed for case-2. Figure 6.8 and Figure 6.9 show the original fall and winter electricity price forecasting and the de-noised wavelet signal, respectively for case-1. For case-2 the original winter electricity demand forecasting and the de-noised wavelet signal is shown in Figure 6.10.
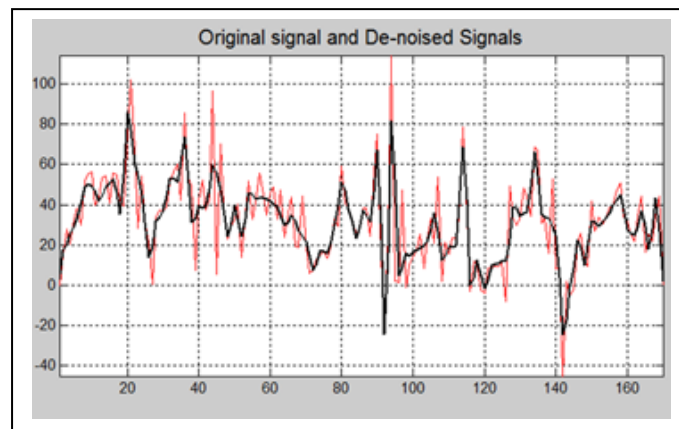
Figure 6.8. *Fall electricity price* showing original signal (red) and de-noised signal (black).
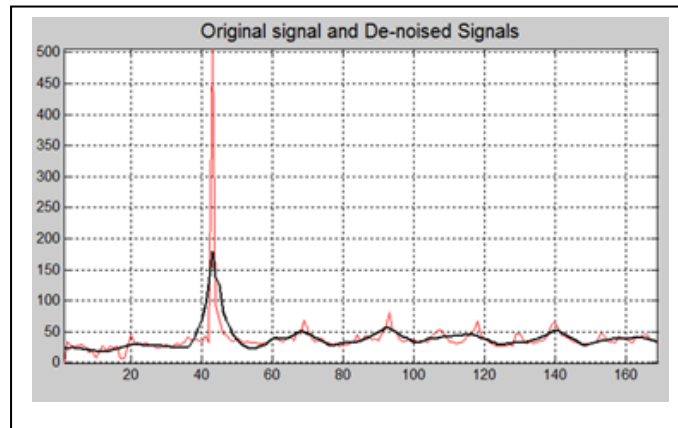
Figure 6.9. *Winter electricity price* showing original signal (red) and de-noised signal (black).
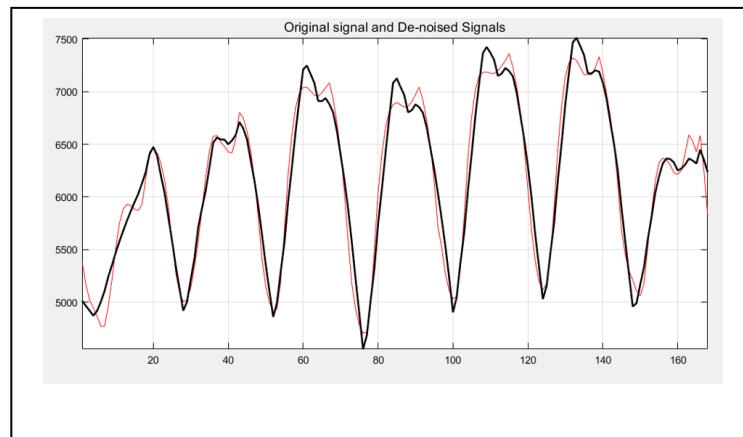
Figure 6.10. NY-ISO *electricity demand* in **2015**.

Case-1: A comparison of the electricity price statistical characteristics is shown in Table 6.1. Specifically, original signal and de-noised signal values from the fall season are shown for comparison. Table 6.2 contains similar information but for winter season.

Table 6.1: Fall season statistical results from original and de-noised data signal from Case-1.

Fall	Original Signal	De-noised Signal
Mean	31.83	31.66
Median	32.81	31.94
Max value	114.4	85.75
Min value	-45.34	-24.73
Stardard deviation	22.36	18.22

Table 6.2: Winter season statistical results from original and de-noised data signal from Case-1.

Winter	Original Signal	De-noised Signal
Mean	39.07	39.05
Median	34.29	36.21
Max value	505.9	178.9
Min value	-3.13E-10	17.53
Stardard deviation	37.95	19.8

Case-2: A comparison of the electricity demand statistical characteristics is shown in Table 6.3. Specifically, original signal and de-noised signal values from the winter season are shown for comparison.

Table 6.3: Winter season statistical results from original and de-noised data signal from Case-1.

Fall	*Original Signal*	*De-noised Signal*
Mean	6293	6160
Median	6223	6267
Max value	7284	7512
Min value	4643	4552
Stardard deviation	833	756.5

6.1.7.2 Evaluation

The three case studies were analyzed for all the Scenarios, which are presented in the following paragraphs.

NN Scenarios studies

The NN is modified in three main configurations as explained below. Each has the same input, hidden and output constraints. The actual price (P_t), actual demand (D_t), week-ahead price temperature (TP_{t-168}), week-ahead demand temperature (TD_{t-168}), week-ahead price (P_{t-168}), and weak-ahead demand (D_{t-168}) are provided to the input layer in the neural network. shown in Figure 6.7. The output of the neural network is the week-ahead forecasted price ($\widehat{P_t}$) for case 1, and for case 2, it is the week-ahead forecasted demand ($\widehat{D_t}$), also shown in Figure 7.

The typical performance metric for evaluating forecasting methods are the Mean Absolute Percentage Error (MAPE), which can be conducted using Equation 6.2.

$$MAPE\ (\%) = \frac{100}{N} \sum_{t=1}^{N} \left| \frac{P_t - \widehat{P_t}}{P_t} \right| \tag{6.2}$$

where P_t and $\widehat{P_t}$ are the actual price and the forecasted price, respectively, and N is the number of samples for case 1, and for case 2 the electricity demand was applied in equation (1).

CASE-1:

1) *Scenario-1 [NN NARX only]*

The NARX NN is utilized to forecast the electricity price in an NY-ISO market in 2010. The historical price data is divided into four seasons, as described in the previous section. Since the selected data for output is week-ahead (168 mins), this data is used as the training target.

2) *Scenario-2 [De-noised Wavelet before NN NARX (DW-NN NARX)]*

This case utilizes the proposed model, where the de-noised wavelet technique is applied first to actual electricity price (P_t), taking into account the fact that it is a noisy signal. De-noised wavelet process results are the input of NARX NN.

3) *Scenario-3 [NN NARX after de-noised Wavelet (NN NARX-DW)]*

In this scenario, the de-noised wavelet process is applied after the electricity price forecasting by the NN NARX network.

Comparison Analysis and Results for Case-1

The comparison between the results in the three scenarios is detailed below. Table 6.4 shows the comparison between Scenarios 1, 2 and 3.

Table 6.4: <u>CASE-1</u>: MAPE (%) results for Scenario-1, Scenario-2 AND Scenario-3.

SEASON	*MAPE (%)* *NARX NN*	*MAPE (%)* *DW-NARX NN*	*MAPE (%)* *NARX NN-DW*
Spring	13.58	14.83	10.72
Fall	62.28	46.15	53.47
Summer	11.44	12.79	21.55
Winter	**23.09**	**21.63**	**18.05**

Examination of the three scenarios' result for the spring season, shows that the NARX NN-DW model presents better performance with 10.72 % (MAPE). The average MAPE obtained with the three methods (scenarios) result in 27.6 %, 23.85 % and 25.95 %, respectively. These results

exhibit a better performance using the de-noising wavelet input as a pre-processed step, by handling non-linearity feature in the price forecasting.

CASE-2:

1) Scenario-1 [NN NARX only]

The NARX NN is utilized to forecast the electricity demand in an NY-ISO market in 2015. The historical price data is divided into four seasons, as described in the previous section but only winter season was taken into account (winter storm case). The data for output is week-ahead (168 mins), this data is used as the training target as explained in Case-1.

2) Scenario-2 [De-noised Wavelet before NN NARX (DW-NN NARX)]

This case utilizes the proposed model, where the de-noised wavelet technique is applied first to actual electricity demand (D_t), taking into account the fact that it is a noisy signal. De-noised wavelet process results are the input of NARX NN.

Comparison Analysis and Results

The comparison between the results in the three scenarios is detailed next.

1) Scenario-1 and Scenario-2 results

Table 6.5 shows the comparison between Scenarios 1 and 2.

Table 6.5: <u>CASE-2</u>: MAPE (%) results for Scenario-1, Scenario-2, and Scenario-3.

SEASON	*MAPE (%) NARX NN*	*MAPE (%) DW-NARX NN*
Winter	**0.039**	**0.032**

Examination of the two scenarios' result for the winter season, shows that the ***DW-NARX NN*** model presented the best performance with ***0.032 % (MAPE).*** This result exhibits a better performance using the de-noising wavelet as input, to handle non-linearity feature in the processed signal and in the demand forecasting signal in the case of a winter storm.

6.1.1 Conclusion Scenario-1 (S1) Analysis -1

The electricity price and demand forecasting can be enhanced using a de-noised wavelet as a prepossessing technique. Similarly, data mining allows better electricity price and demand forecasting in an extreme-weather event. The proposed model shows accurate electricity price and in a nonlinearity winter storm case, where the *MAPE* comparisons between the cases clearly shows that the proposed signal treatment with the de-noising wavelet is well suited to forecast the electricity price and demand. In the next section, machine learning as data mining technique to manage the nonlinearity in the electricity price and demand under an extreme-event are presented.

6.2 Experimental Framework model Scenario-1 analysis-2.

Figure 6.2 shows the methodology used to incorporate both domains in Scenario-1. As discussed in section 6.1, only one model in the KDP framework was used in scenario-1 to assess the impact of an extreme-weather event in the ETDS. However, to prove the generalization of the proposed KDP framework model, Scenario-2 used several models, as presented in the next chapter. Analysis-2 implies different data transformation and data mining techniques as shown in Figure 6.2. As discussed in the previous analysis, step-1 and 2 are assumed to be the same since the development of this analysis is under the winter storm that hit New York city during 2015.
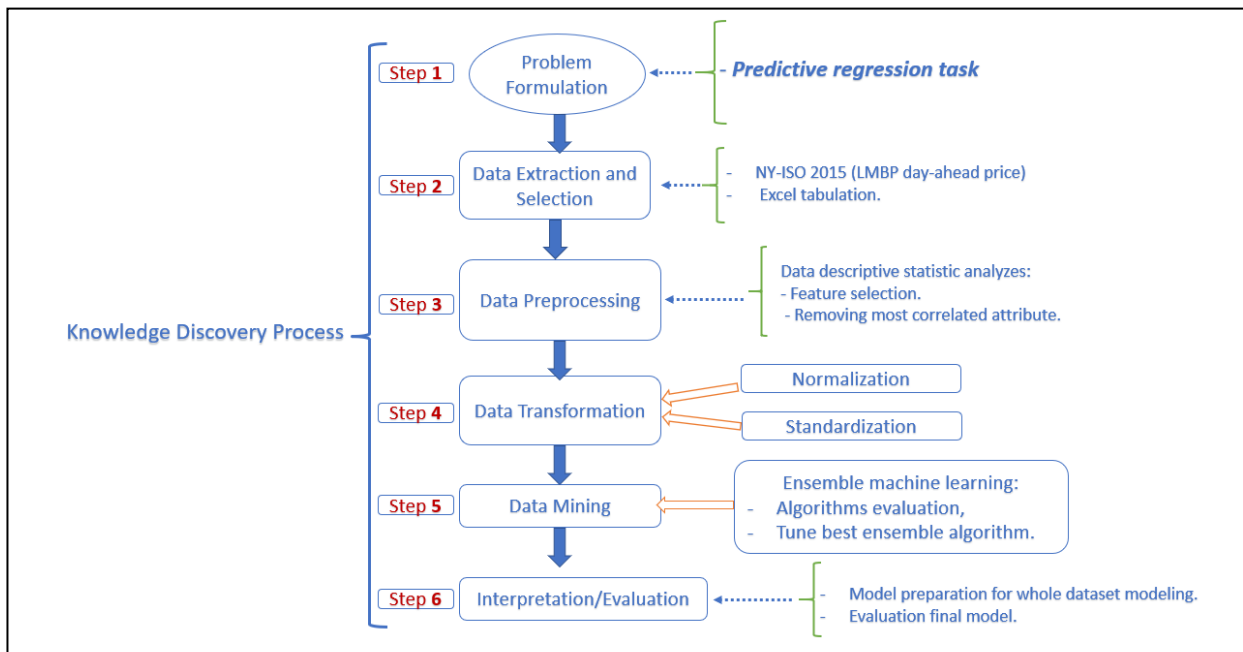
Figure 6.2. Experimental framework model Scenario-1 analysis-2.

6.2.1 STEP 3: DATA PRE-PROCESSING

Missing values and outliers were removed similarly as analysis-1, in which, the same da-taset (excel tabular format) was used to perform this analysis and shown in Appendix A. However, to better understand the data features, selection techniques like descriptive statistics and visualiza-tion was used in this step. The objective is to apply the python software tool to descriptive statistics. The modeling technique is divided into 9 main steps. Appendix A.1.2.

The task selected in this analysis is predictive regression, thus, the libraries chosen are in Appendix D. The dataset contains a total of 8759 data points with a time horizon of one year (*January – December' 2015*), including information on temperature, demand, price and hour value. The pairwise correlation applying the ***Pearson's correlation coefficient*** between variables is shown in the correlation matrix on Figure 6.4 with the corresponding values shown in Table 6.6, respectively.

72
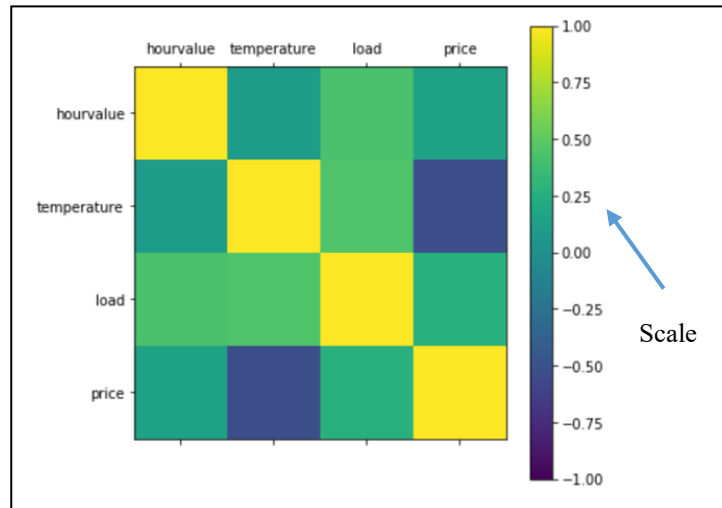
Figure 6.4. *Correlation matrix* of data attributes in Scenario-1 analysis-2.

Table 6.6: Pearson's correlation coefficient *Scenario-1, analysis-2*.

	hourvalue	tempera-ture	load	price
hourvalue	1.00	0.11	0.43	0.15
tempera-ture	0.11	1.00	0.45	-0.52
load	0.43	0.45	1.00	0.27
price	0.15	-0.52	0.27	1.00
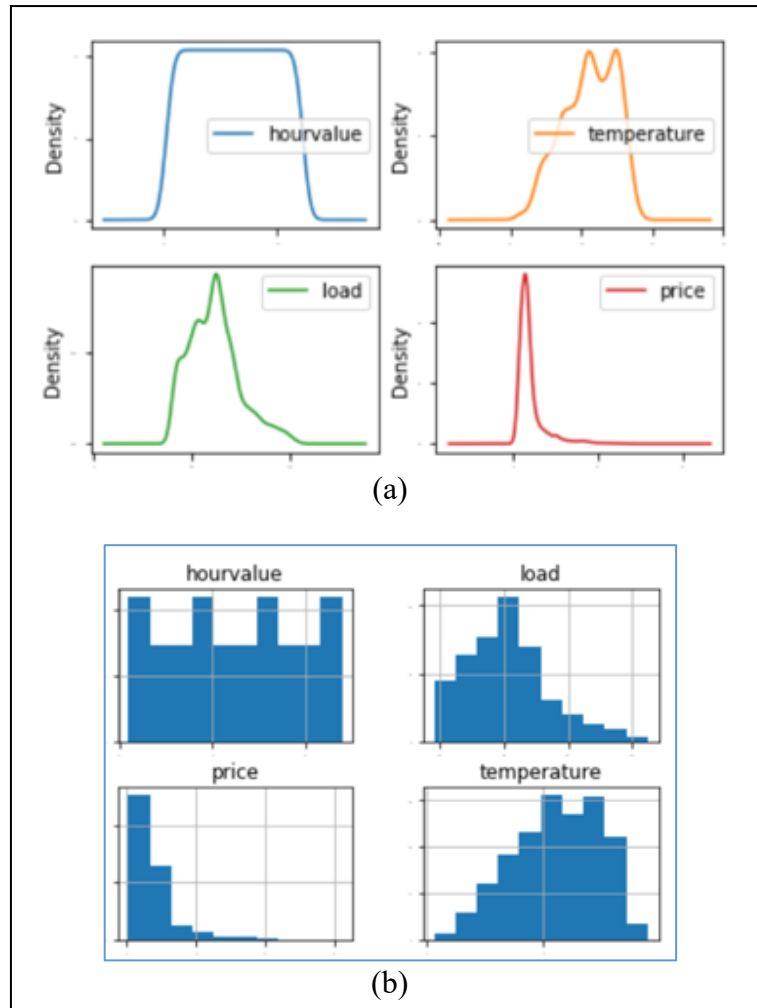
Figure 6.5. Descriptive statistical analysis winter storm dataset. a) Correlations between variables b) Histogram of each set of data points.

The observed correlation between temperature-price pair shows a negative correlation indicating the adverse consequence in the electricity price by increasing the values when the temperature decreases, i.e. winter storm. "*Temperature*" correlation between "*load*" and "*dayvalue*" *(weekday or weekend)* present better correlation. Moreover, the most affecting variable is price as shown in the density curves and histograms in the Figure 6.4a and Figure 6.4b, respectively. In this study, data deep analysis based on histogram and density prediction is not included, for simplicity purposes. On one hand, the "price" attribute shows an *exponential distribution*. On the other

hand, the remaining variables "hourvalue," "load," and "temperature" show a more *gaussian distribution*. In analysis-1, those singularities are observed, but are more evident here, with the assistance of data descriptive statistics.

6.2.2 STEP 4: DATA TRANSFORMATION

In this analysis, data was normalized and standardized during the process of data mining. The reason is to acquire a better selection of the type of standardized and normalized method, according to the behavior of the datasets observed in step 3 in this analysis to reach better results as discussed in chapter 4.

6.2.3 STEP 5: DATA MINING:

The whole dataset was divided in two groups: ***training set*** with the 80% of the data used in this analysis, and ***validation set*** and the 20 % representing the rest of the dataset. The division percentage of each set depend on the total size of the dataset. As evidenced in chapter 4, a good number is 80% for training and 20% for validation, as used in this analysis. The validation set is discussed in the next step. The training set is applied during the process of machine learning modeling implied in this analysis. Additionally, this set was specified as a random seed ensuring that the data was divided in sets of seven, randomly to ensure generalization and prevent overfitting as discussed in chapter 4. The problem formulation from step 1 and the observation in the data behavior in subsequent steps shows two types of data: linear and non-linear, which can be modeled as a *regression model*. Thus, a *machine learning regression prediction* is proposed here, to solve the *price* prediction problem during a winter storm. Several machine learning algorithms for regression analyses discussed in chapter 4 were performed, for extended explanation refer to [Mull16]. The libraries explored are shown in Appendix D 2.1. The results are shown in next step, under interpretation sections.

6.2.4 STEP 6: INTERPRETATION / EVALUATION

In this step, the interpretation is shown first and then the evaluation of the machine learning algorithms follows.

6.2.4.1 Interpretation

Machine learning algorithms performance can be estimated before final modeling, by using a method of validation. There exist several methods to do this. Under the assumptions observed in step 3, the score method named *cross validation* is good approach. The *k-fold cross validation* is a good approach to estimate the performance of the experimental machine learning performed in this analysis. Splitting the data in *K* partitions of equal size, where *i* partition represents the evaluation partition of a remaining *K-1* partition for training. The final score is obtained with the average of the *K* scores obtained [Choll18]. The set was split into 10 k-parts named *k-fold* by several authors [Choll18], [Mull17] to estimate the accuracy of the models. For better performance of the model, the *K-fold* needs to be applied several times. The results are expressed in terms of *mean-square-error* which represents the magnitude of the error during the prediction. Table 6.7 shows the results of the preliminary evaluation of the experimental machine learning regression predictive algorithms (MLRPA). The meaning of the six different algorithms is listed below.

> ***Linear Algorithms:***
>
> *LR*: Linear Regression
>
> *LASSO*: Least Absolute Shrinkage and Selection Operator
>
> *EN*: Elastic Net.
>
> ***Non-Linear Algorithms:***
>
> *KNN*: K-Nearest Neighbors.
>
> *CART*: Classification and Regression Trees.
>
> *SVR*: Support Vector Regression Machines.

Table 6.7: Mean-Square-Error and Standard deviation of the preliminary MLRPA algorithms from Scenario-1, analysis-2.

	LR	LASSO	EN	KNN	CART	SVR
MSE	403.677513	403.712163	403.693367	292.121420	336.249711	1051.81143
SD	32.637604	31.672115	31.710627	34.894623	51.560274	95.499186

The observed estimation accuracy of the models shows a tight distribution between the algorithms *LR*, *LASSO*, and *EN*. Since data is not yet standardized probably *KNN* and *SVR* algorithms are far from the rest. Next, standardization of the datasets, including the scaling in the dataset are applied as a data transformation to avoid error in the performance. Table 6.8 shows the results of the second evaluation of the experimental **Machine Learning Regression Predictive Algorithms (MLRPA)** applying standardization and scalation on the datasets.

Table 6.8: Mean-Square-Error and Standard deviation of the second MLRPA algorithms test. Standardization and scalation on datasets.

	LR	LASSO	EN	KNN	CART	SVR
MSE	403.677513	408.659687	543.317283	217.084419	327.868396	316.458567
SD	31.637604	34.145117	51.876025	28.050340	49.544664	37.687113

The observed estimation accuracy of the models shows KNN with the lowest error. To improve the performance of this wining (best) algorithm it is necessary to perform a tuning in the hyperparameters to achieve better results. Thus, different hyperparameters are applied in this analysis to find the optimum. The values used are *n_neighbors*: 1,3,5,7,9,11,13,15,17,19,21. Table 6.9 shows the different parameters for the *n_neighbors*.

Table 6.9: Mean-Square-Error and Standard deviation of the different tuning hyperparameters used in KNN algorithm.

N_neighbors:	MSE	SD
1	341.789383	48.572649
3	239.811523	34.469458
5	217.283144	28.063716
7	205.888775	26.469730
9	203.673945	29.790342
11	**201.384079**	**29.075794**
13	202.071956	26.285250

15	203.709268	26.608291
17	205.180865	26.671462
19	206.430158	27.390771
21	207.468043	27.644827

The observed estimation accuracy of the different hyperparameters shows the 11 *N_neighbors* as the best with 201.38 of MSE. Moreover, there is evidence that combined multiple machine learning techniques are proven most effective for better performance. The multiple combinations of machine learning is named ***ensemble methods*** [Mull16]. In this study, four different ML algorithms were combined to prove this assumption. Table 6.10 shows the results of the evaluation of the experimental machine learning assembled models. The meaning of the four different algorithms is listed below.

Assembled Algorithms:

AB: ADA Boost Regressor.

GBM: Gradient Boosting Regressor.

RF: Random Forest Regressor.

ET: Extra Trees Regressor.

Table 6.10: Mean-Square-Error and Standard deviation of the different ensemble algorithms.

Ensemble algorithm:	MSE	SD
AB	364.003558	38.524053
GBM	200.884489	23.230503
RF	209.222170	31.994737
ET	230.672360	38.038168

The observed estimation accuracy of the different ensembles shows ***GBM*** as the best with 200.88 of MSE. Thus, the GBM algorithm needs to be tuned, as discussed before, to test different

hyperparameters. The values used were *n_estimators*: 50, 100, 150, 200, 250, 300, 350, 400. Table 6.11 shows the different parameters for the *n_estimators*.

Table 6.11: *Mean-Square-Error and Standard deviation* of the different tuning hyperparameters used in *GBM algorithm*.

N_estimators:	MSE	SD
50	207.939351	24.373456
100	200,977390	23.066173
150	197.413899	23.714248
200	**195.835571**	**24.312499**
250	196.335381	26.018445
300	196.599757	25.488279
350	196.583230	25.156887
400	196.399710	25.686521

The observed estimation accuracy of the different hyperparameters shows the *200 estimators* with *195.84* as the best for the *GBM algorithm*.

6.2.4.2 Evaluation

This final step presents the final configuration of the model as "***the final model***," in this analysis the **Gradient Boosting Regressor** demonstrates having the least MSE. Thus, the final model needs to be trained with the entire training dataset. Prior to this, the data need to be standardized and scaled to better modeling performance as discussed above. The final ***MSE result*** of the ***GBM is estimated as 194.11357825***.

6.3 Conclusion

In this chapter we demonstrate a novel method to solve one task in the Electricity and Transmission Systems during an extreme-event. Data Mining Algorithms was implemented by the use of MATLAB and Python software to solve a *Supervised Learning Non-linear Auto-regression*

eXogeneous Neural Network (NARX) Neural Network and Regression Predictive Machine Learning. Moreover, all the experimental data in this chapter was acquired of truthful sources as a base of data mining in the ETDS during a winter storm. In the next chapter GIS techniques and Deep Learning Neural Network techniques are explored to solve abroad problem during the unfolding of a devastating hurricane.

CHAPTER 7: SCENARIOS-2 AND ITS EXPERIMENTAL FRAMEWORK MODEL

A number of uncertainties exist in the ETDS during the unfolding of an extreme-weather event. Observation from the previous chapter indicates that the proposed KDP methodology offers a better result for price and demand forecasting in the case of a winter storm. Moreover, in chapter 2, the optimal operation of the ETDS can be approximated using the *DC Optimal Power Flow (DCOPF)* by including all the system constraints. Thus, a complex methodology is needed to overcome an extreme-weather event, and at the same time to enhance the resilience in the ETDS. Thus, this chapter presents a solution for such problem including an extension of the method proposed in the previous chapter.

7.1 Experimental Framework model Scenario-2

It is pointed out that under the stressed ETDS system, involves multiple problem-solving tasks, e.g. prediction, classification, etc. including many datasets. A solution for such problem is the reformulated model proposed in Scenario-2. Three models were used for assessing the impact of an extreme-event in the ETDS based on the ***KDP framework*** presented on chapter 5. The reformulated model is a ***Multi-Domain Triple Rehearsal Dynamic System*** **(MDTRDS)**, in which, three different interactives dynamic models are included: a) *Extreme-weather model*, b) *Component model*, and c) *System model*. Figure 7.1 shows the experimental MDTRDS generic model.
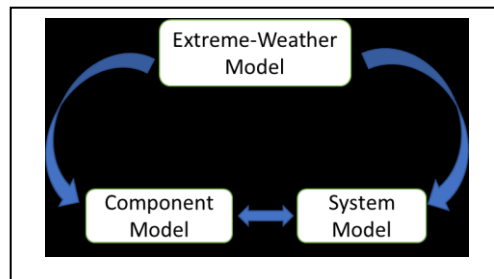
Figure 7.1: Generic Triple Rehearsal Dynamic System **(MDTRDS)** model.

7.2 Scenario-2: Hurricane Harvey (Texas – 2017)

This section is devoted to the analysis of Scenario-2. Step 1 from the KDP (problem formulation) is presented first. Taking into account the three phases of the multiphase resilience trapezoid discussed in chapter 2, scenario-2 looks at phases 1 and 2 during the unfolding of the hurricane (also shown in figure 2.9 on section 2.3), which correspond to the window time in scenario-2. Thus, the three models have the same problem formulation (step 1). A brief synoptic history of Hurricane Harvey is presented next [Hazu18]. A more detailed account can be found in Appendix B.1, which includes information from NOAA and the National weather services [Blak18].

7.2.1. HURRICANE HARVEY.

Formed: August 17, 2017   **Dissipated:** August 31, 2017

Peak Winds: 130 mph (Category 4)   **Lowest Pressure:** 938 mb

Stage 1:	*Stage 2:*
Landfall 1: August 26, 2017 – Rockport, TX	*Landfall 2:* August 30, 2017 – Camron, LA
Landfall Winds: 130 mph (**Category 4**)	*Landfall Winds:* 45 mph (**Tropical Storm**)

Important Observation: After landfall, Harvey continued north and then north-northeast, weakening to a tropical depression over central Louisiana on August 30. The last NHC Advisory was issued on August 31, 2017, at 0300 Coordinated Universal Time (UTC).

7.2.2. STEP 1: PROBLEM FORMULATION (ALL MODELS TRDS).

As an observation, the critical operation of the ETDS needs to be assessed due to the implications during an extreme-weather event. One of the required analyses is to take into account the path of the hurricane that could affect a specific section of the electric grid. Under the assumption of the existing hazard, one question arises: how vulnerable is the equipment in the ETDS that are in the path of the hurricane? The analysis of Scenario-2 was focused in a hurricane storm reported during the year of 2017 in the geographical region of Houston named *"Harvey."* As discussed previously in section 1.6.4 of this dissertation, the electrical infrastructure of Texas is the

one of the largest in the United States and is also one of the major producers in the supply of energy.

With this motivation and the important observation from the synoptic history presented above, a power failure assessment in the Houston area impacted by Hurricane Harvey is presented with four targeted major objectives, to verify the feasibility of the proposed TRDS model by using techniques from chapter 4.

Task:

a) *Objective explored: **Critical Hurricane path analysis (T-A)**.* Based on historical environmental parameters i.e., approximate landing position, maximum sustained surface wind speed, temperature, *Saffir-Simpson category,* central pressure and affected cities (geographic location), identification of the critical hurricane path footprint in the electrical transmission and distribution network.

b) *Objective explored: **ETDS components vulnerability analysis (T-B)**.* Based in the previous analysis (Subsection A), the probability of failure of the critical components by considering their physical strength i.e., aging, demand stress (i.e., classification of critical components by ranking the demand), and environment deterioration under the recurrent normal and extreme-weather influence.

c) *Objective explored: **Critical system operation modeling (T-C)**.* ETDS Optimal power flow prediction under a sequential power fault through the storm event (disconnection cases). The tests were used as a base to estimate the contingent-states by ranking the final results.

7.2.3 DETAILED EXPERIMENTAL FRAMEWORK MODEL SCENARIO-2

To solve the targeting tasks from step 1, a more detailed of the MDTRDS model for Scenario-2 is presented in Figure 7.2.
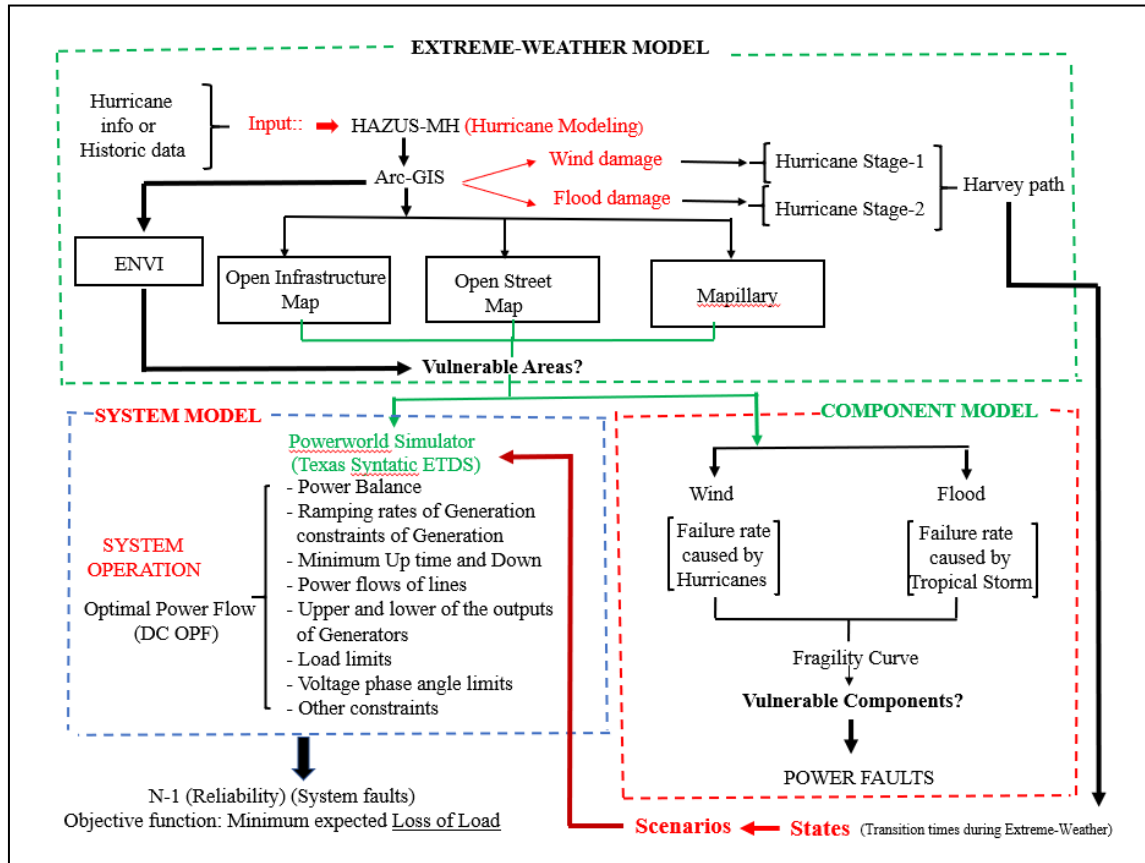
Figure 7.2: Detailed Multi-Domain Triple Rehearsal System (MDTRS) model for Scenario-2.

7.2.4 MDTRS TRANSLATION STATES ON THE HURRICANE PATH.

A hurricane moves along a trajectory or path. Figure 7.3 shows the trajectory of Hurricane Harvey (2017) from the creation stage to the extinction stage [Nasa19].
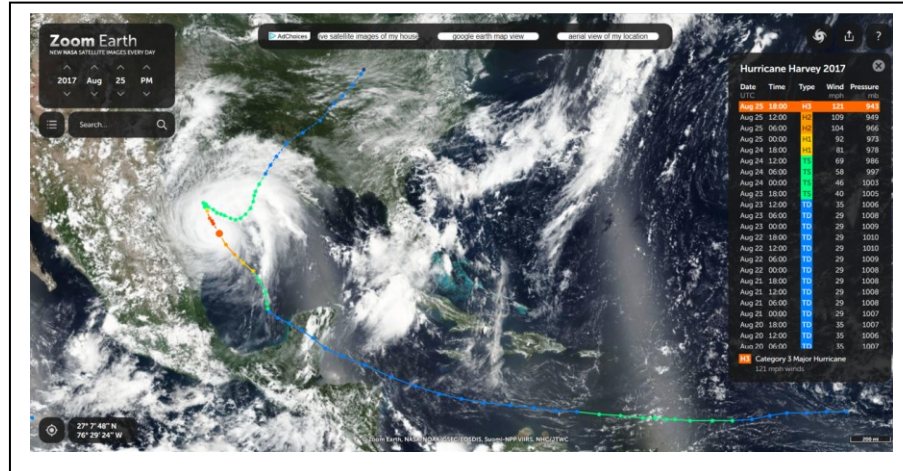
Figure 7.3: Hurricane Harvey trajectory from the creation to the extinction stage, image taken from NASA Zoom Earth application [Nasa19].

As the important observation from section 7.2 and the observed from figure 7.3 above, Hurricane Harvey makes double landing triggering two major stages: 1) Wind damage, *hurricane category 4*; and 2) Flood damage, *tropical storm*. Hurricane scales can be referenced in Appendix B.1.

Thus, translated states need to be considered. Figure 7.4 shows the translated states of the MDTRS following the path of the hurricane with two unfold major stages: stage-1 (Hurricane category 4), and stage-2 (Tropical storm).
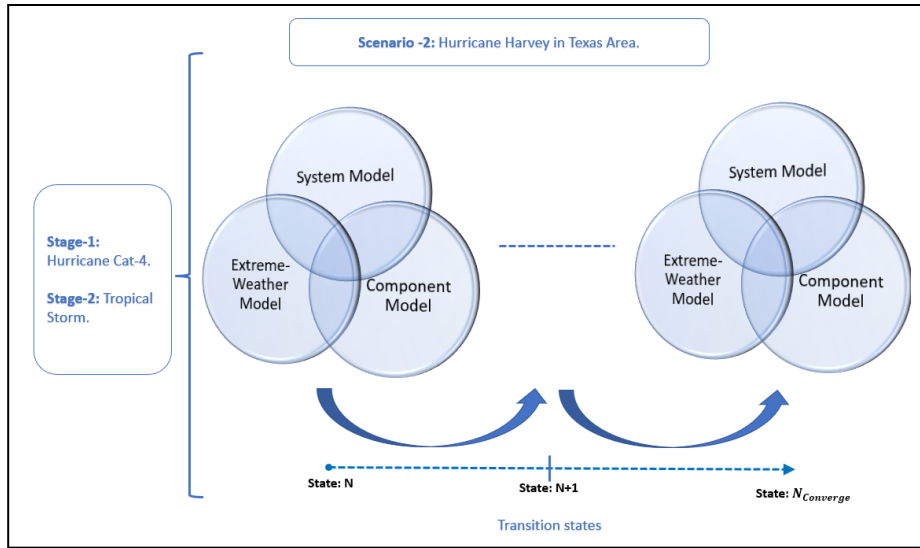
Figure 7.4: Translated states of the MDTRS following the path of the hurricane.

7.3 Extreme-weather model.

Extreme-weather model, refer the stage during the unfold of hurricane Harvey. Under this analysis, the trajectory of the hurricane was taken into account to solve target **T-A** from section 7.2.1. This model is shown in Figure 7.5. Each step is described below.
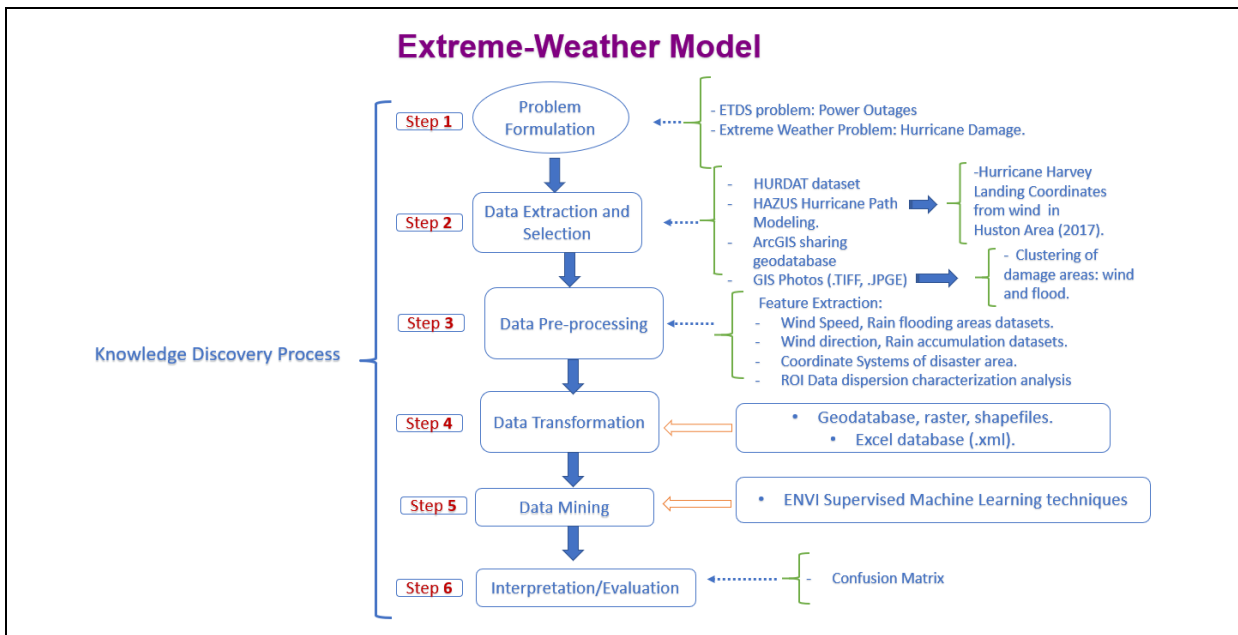
Figure 7.5: Extreme-weather model showing the detailed steps.

7.3.1 STEP 2. DATA EXTRACTION AND SELECTION

In this step, data extraction and selection are explained.

7.3.1.1 Data Extraction

Data extraction was taken from several sources. Appendix C shows the data flow extraction for this analysis. ***Dataset HURDAT***, as discussed in section 2.5, contains the following data: Landing position, approaching angle, translation velocity, central pressure, maximum wind speed, radius of maximum wind, and gust peak.

7.3.1.2 Data Selection

The data collection is based on information from NOAA, NASA, HURDAT web sites. Wind damage analysis under HAZUS simulation was taking account the cities that hit Hurricane Harvey based on a time horizon on the date and hour of the landfall date and hour used as a point of prediction levels. Figure 7.6 show the landfall time and date [Noaa19b].
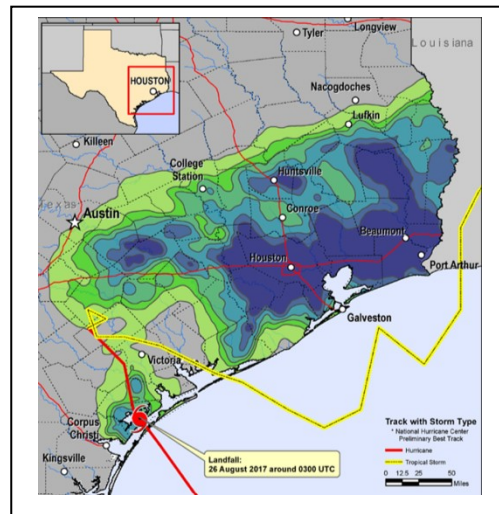
Figure 7.6: ***Hurricane Harvey landfall date and time***, image taking from [Noaa19].

Figure 7.7 shows the results of the ***HAZUS simulation*** based in ***ArcGIS 1.5.1 software***, showing the geographically wind profiles during Hurricane Harvey.
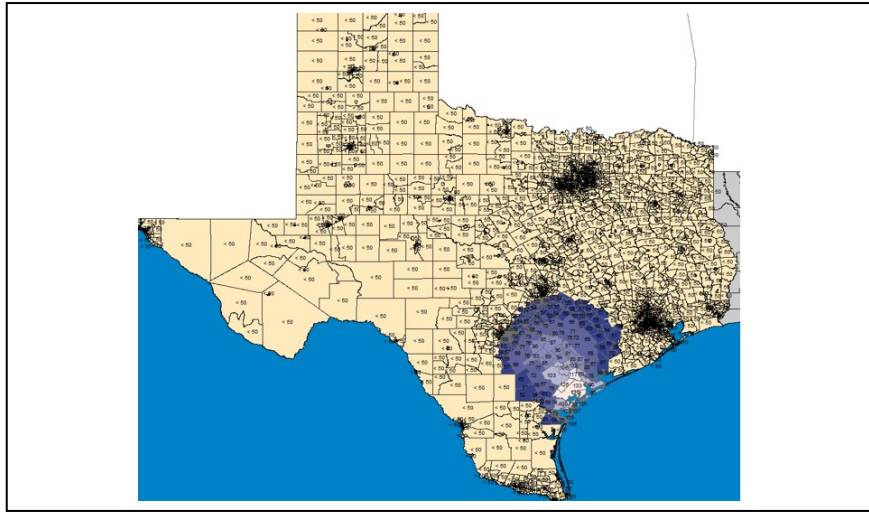
Figure 7.7: Hazus wind damage simulation using ArcGIS 1.5.1 version software.

Figure 7.8 shows a section of the electricity transmission system most affected during the path of Hurricane Harvey [Open19].
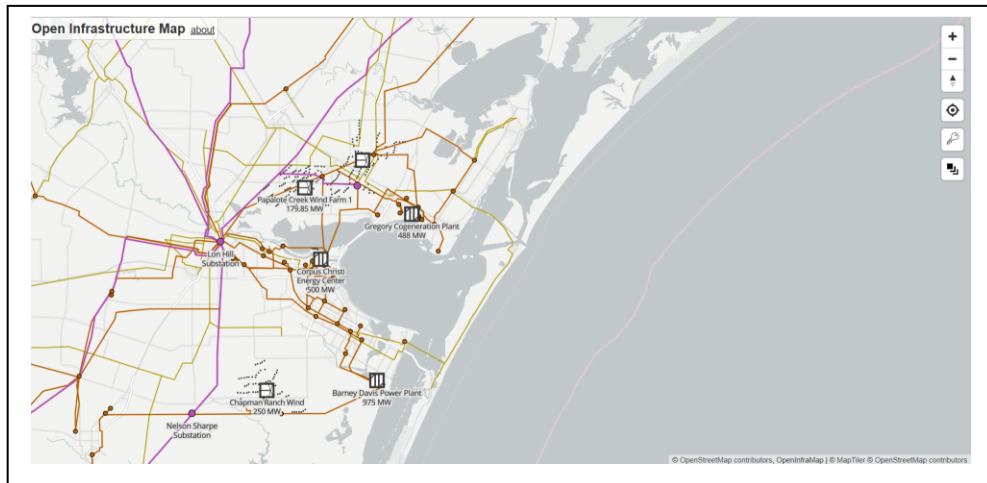
Figure 7.8: Aransas pass and surrounding areas, showing the most affected Electricity transmission system during the path of Hurricane Harvey [Open19].

The data from open engines like Open Infrastructure and Open Street was used to visualize the components affected during the extreme-event. Figure 7.9 and Figure 7.10 shows the electricity

transmission components most affected during the path of Hurricane Harvey in the area of Aransas pass city.

(a)

(b)
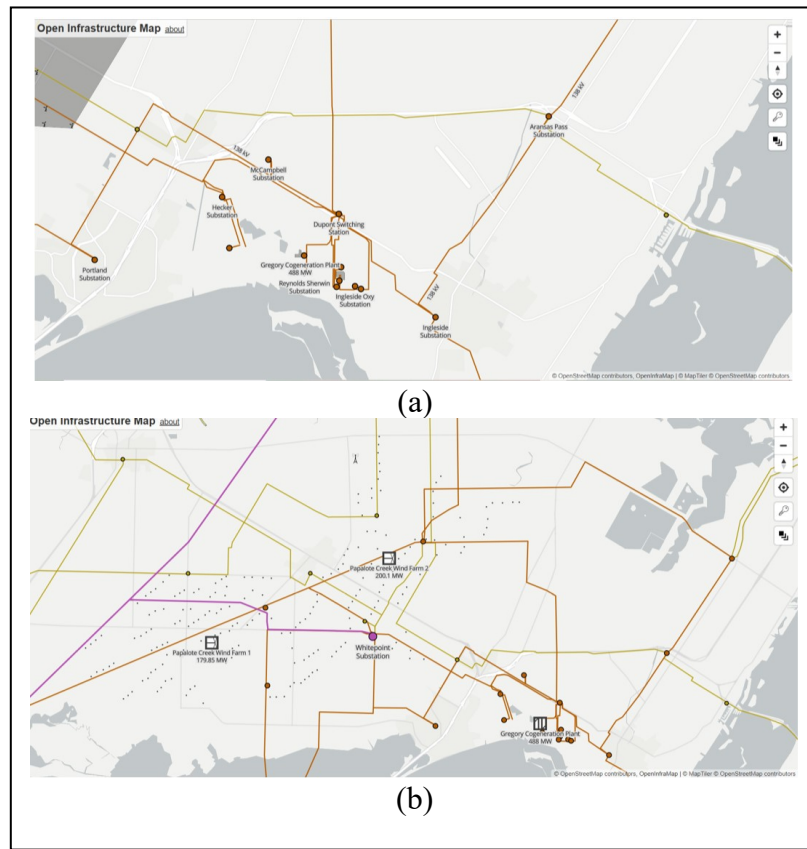
Figure 7.10: Electricity transmission components most affected during the path of Hurricane Harvey, Aransas pass city [Open19].

Several electricity transmission poles were affected during the unfold of the extreme-event. Figure 7.11 shows the affected area in poles vision level. Figure 7.12 show an evidence of poles damaged reported from AEP company [Aept18].
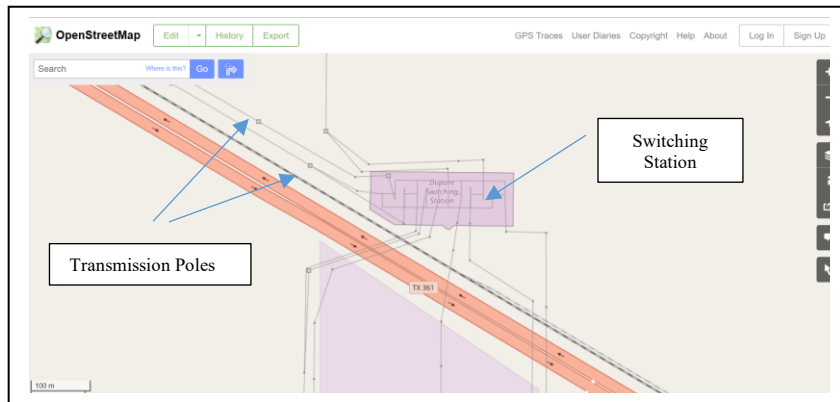
Figure 7.11: Electricity transmission poles located in Aransas pass city [Opst19].
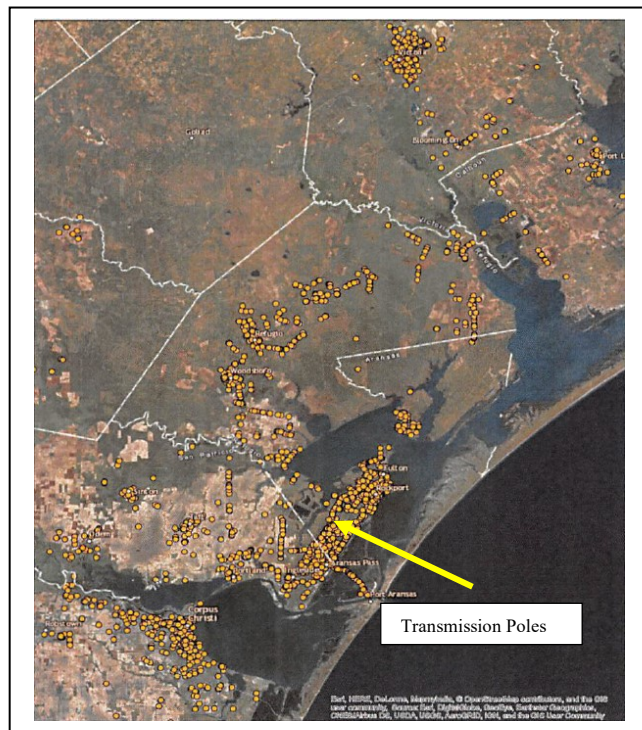
Figure 7.12: Electricity transmission ***poles damaged*** in Aransas pass city [Aept18].

The Houston area was hit with a *flood* during *Hurricane Harvey*. Barker reservoir located in Harris county and the Rosenberg city located in Fort Bend county, was two of the most affected from flooding, caused by the extreme-weather event. As shown in above Figures 7.2 and 7.6 Hurricane Harvey turned into a tropical storm on the second day of landfall on August 27 of 2017. Figures 7.13a and Figure 7.13b shows two screen captures of the aerial images of Rosenberg before and after the extreme-weather event. Figures 7.14a and Figure 7.14b show two screen captures of

the aerial images of Barker reservoir before and after the extreme-weather event. Both pair of images were used in this analysis.
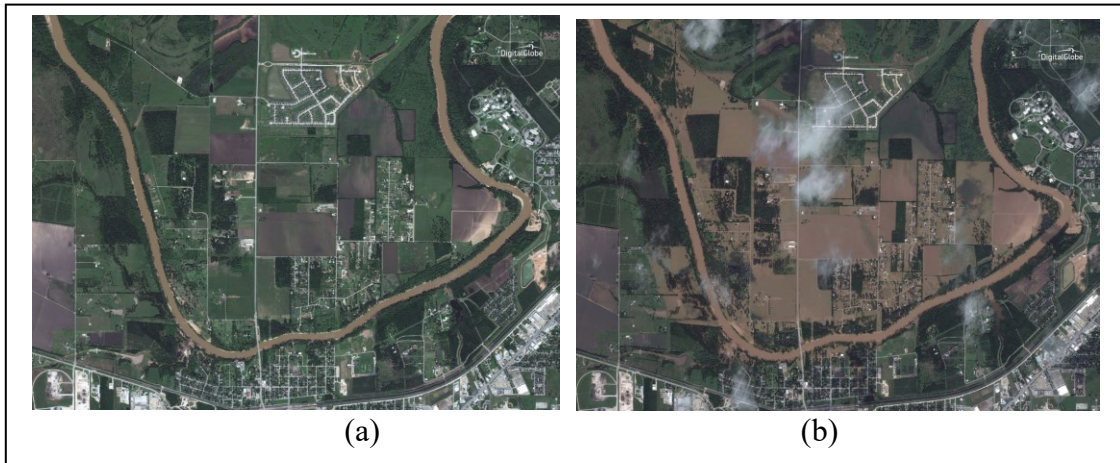
Figure 7.13: Screen captures of the aerial images of Rosenberg. a) *Rosenberg* area **before** flood damage, and b) *Rosenberg* area **after** flood damages [Goog19].

Figure 7.14: Screen captures of the aerial images of Barker reservoir. a) *Barker reservoir* area **before** flood damage, and b) *Barker reservoir* area **after** flood damages [Goog19].

7.4.2 STEP 3. DATA PRE-PROCESSING

Supervised Machine Learning on images need to be preprocessed as digital images format such JPG, GIF, TIFF, and RAW. The processing is similar to techniques' in remote sensing. In this analysis the format used was the *Tagged Image File Format (TIFF)*. TIFF format is able to

save information on the four bands in a single file. The bands consist of Red (R), Green (G), Blue(B), and near infrared (NIR). The software used in this analysis was ENVI 5.3 version. The *Region of Interest (ROI)* needs to be selected as training data and testing data for supervised Machine Learning classification. The images from Figures 7.13 and Figure 7.14 were taken as the selected data for this analysis. Thus, the ROI from both images need to be pre-processed. The *Rosenberg selected ROI's* were: Grass_test, Street_test, Water_test for the testing dataset. Grass_train, Street_train, and Water_train were used for the training dataset. The *Barker reservoir selected ROI's* were: Water_test, Wet_Street_test, Dry_Street_test for the testing dataset. Water_train, Wet_Street_train, Dry_Street_train were used for the training dataset. Figure 7.15a shows an image capture of the *ROI selection process* for *Rosenberg*, and Figure 7.15b shows an image capture of the *ROI selection process* for *Barker reservoir*.

Figure 7.15: ***ROI's selected***. a) *Rosenberg ROI's*, and b) *Barker reservoir ROI's*.

7.4.3 STEP 4. DATA TRANSFORMATION

Point plots or scatter plots can be useful way of examining the relationship between the *ROIs*. In this exploratory data analysis, a group of variables in the ***ROIs*** was analyzed with scatter plots. Figure 7.16a shows the scatter plot of *ROIs* in *Rosenberg*, and Figure 7.16b shows the scatter plot of *ROIs* in the *Barker reservoir*. The observed points show the separation in each *ROIs*, meaning the features selected are well defined from previous step.
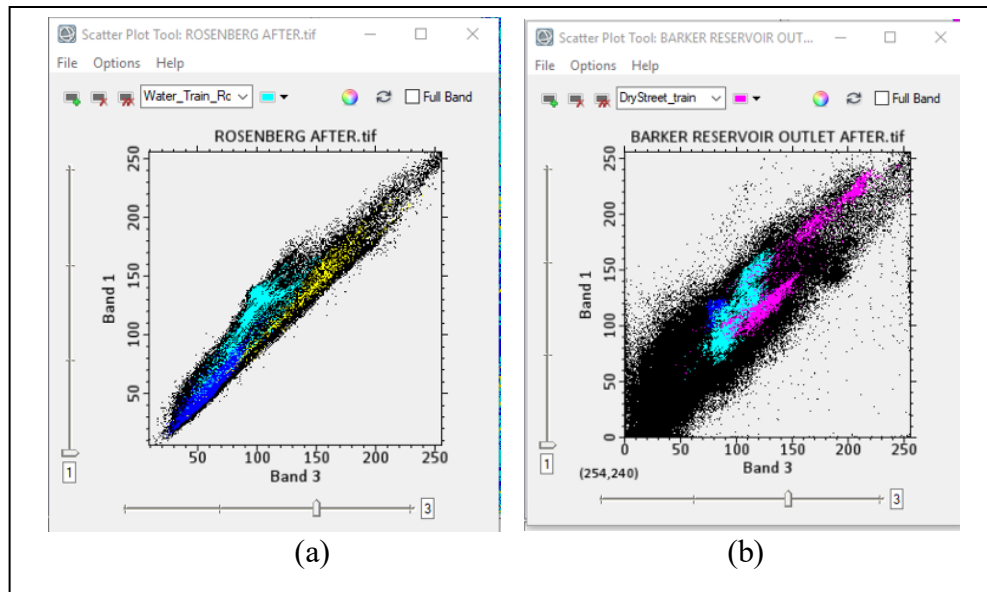
Figure 7.16: *ROI* scatter plots. a) *Rosenberg ROI* scatter plot, and b) *Barker reservoir ROI* scatter plot.

7.4.4 STEP 5. DATA MINING

Supervised Machine Learning Classification is implemented in this step for both of the selected images. The experimental algorithm was *"Maximum Likelihood Classification."* Moreover, techniques like **Sieve** and **Clump classes** was applied to clean the isolated pixels and "salt and pepper" effect. The final image of both ML image classifications are presented in Figure 7.17a and Figure 7.17b for Rosenberg and Barker Reservoir, respectively.

Figure 7.17: Final image classification. a) Rosenberg, and b) Barker reservoir.

7.4.5 STEP 6. INTERPRETATION/EVALUATION

In this step, the experimental algorithm evaluation results were obtained via a confusion matrix. Figure 7.18 shows the ***confusion matrix results for Rosenberg***, and Figure 7.19 show the ***confusion matrix results for the Barker reservoir.***

Class Confusion Matrix — □ ×

File

Confusion Matrix: C:\Users\Rox\Documents\ENVI Supervised Clasification\Rosenberg Analysis\Results\1-Classificatio

Overall Accuracy = (9325/10857) 85.8893%
Kappa Coefficient = 0.7423

Ground Truth (Pixels)

Class	Water_test_Ro	Street_test_Ro	Grass_test_Ro	Total
Unclassified	0	0	0	0
Water_Train_R	5859	0	0	5859
Street_Tran_R	1495	417	19	1931
Grass_Train_R	0	18	3049	3067
Total	7354	435	3068	10857

Ground Truth (Percent)

Class	Water_test_Ro	Street_test_Ro	Grass_test_Ro	Total
Unclassified	0.00	0.00	0.00	0.00
Water_Train_R	79.67	0.00	0.00	53.97
Street_Tran_R	20.33	95.86	0.62	17.79
Grass_Train_R	0.00	4.14	99.38	28.25
Total	100.00	100.00	100.00	100.00

Class	Commission (Percent)	Omission (Percent)	Commission (Pixels)	Omission (Pixels)
Water_Train_R	0.00	20.33	0/5859	1495/7354
Street_Tran_R	78.40	4.14	1514/1931	18/435
Grass_Train_R	0.59	0.62	18/3067	19/3068

Class	Prod. Acc. (Percent)	User Acc. (Percent)	Prod. Acc. (Pixels)	User Acc. (Pixels)
Water_Train_R	79.67	100.00	5859/7354	5859/5859
Street_Tran_R	95.86	21.60	417/435	417/1931
Grass_Train_R	99.38	99.41	3049/3068	3049/3067

Figure 7.18: ***Confusion matrix results*** of *Rosenberg*.

Class Confusion Matrix — □ ×

File

Confusion Matrix: C:\Users\Rox\Documents\ENVI Supervised Clasification\Barker Reseivor Clas

Overall Accuracy = (1699/1748) 97.1968%
Kappa Coefficient = 0.9399

Ground Truth (Pixels)

Class	DryStreet_tes	WetStreet_tes	Total
Unclassified	0	0	0
DryStreet_tra	1075	34	1109
WetStreet_tra	15	624	639
Total	1090	658	1748

Ground Truth (Percent)

Class	DryStreet_tes	WetStreet_tes	Total
Unclassified	0.00	0.00	0.00
DryStreet_tra	98.62	5.17	63.44
WetStreet_tra	1.38	94.83	36.56
Total	100.00	100.00	100.00

Class	Commission (Percent)	Omission (Percent)	Commission (Pixels)	Omission (Pixels)
DryStreet_tra	3.07	1.38	34/1109	15/1090
WetStreet_tra	2.35	5.17	15/639	34/658

Class	Prod. Acc. (Percent)	User Acc. (Percent)	Prod. Acc. (Pixels)	User Acc. (Pixels)
DryStreet_tra	98.62	96.93	1075/1090	1075/1109
WetStreet_tra	94.83	97.65	624/658	624/639

Figure 7.19: ***Confusion matrix results*** of *Barker reservoir*.

A visual comparison of the three images, before-event, after-event, and classification results are presented in a single image. Specifically, the three different images are superimposed to provide a visual comparison, with the "results" portion being the classification of areas (water, street, etc.) in this extreme-weather model analysis. In the case of the Rosenberg analysis, shown

in Figure 7.20, the image is divided in three, as described above, showing firstly, the after-damage image, secondly the before-damage image, and thirdly the supervised Machine Learning classification. In the case of the Barker analysis, shown in Figure 7.21, the image was divided in three as well.
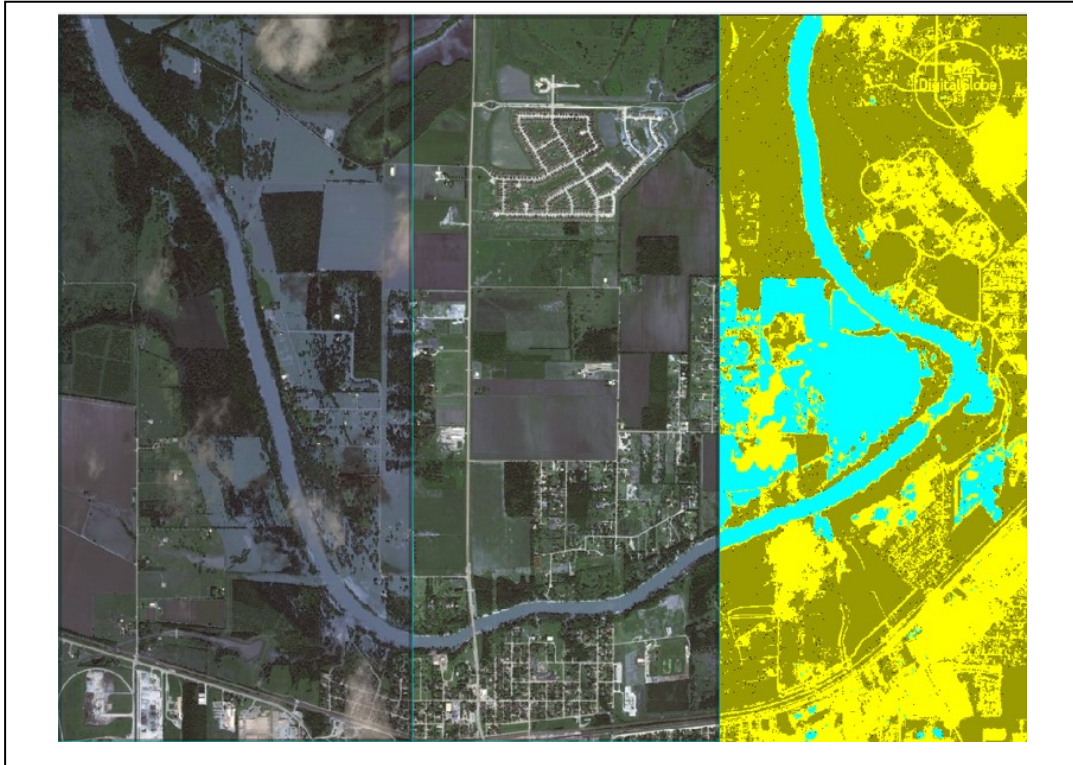
Figure 7.20: Images comparison of *Rosenberg*. (Before, after, and classified).

Figure 7.21: Images comparison of *Barker reservoir*. (Before, after, and classified).

7.5 Component model.

Several components are damaged during the unfolding of an extreme-event. The components that fail are those exposed in the path of a hurricane, but the numbers of failures are uncertain. Several authors use the fragility curve to show the failure probability against the weather intensity. A generic failure rate of components caused by a hurricane can be expressed by Equation 7.1[Li14].

$$\lambda_{wind}[w(t)] = \left[1 + \alpha \left(w^2(t)\Big/w_{sf}^2 - 1\right)\right] \lambda_{norm} \qquad (7.1)$$

where *w(t)* represents the wind speed at time *t*; λ_{sf} is the extreme-weather wind speed based in table B.1 from Appendix B; λ_{norm} represents the failure rate under normal weather conditions; α is the electric component parameter.

Component model refers to the component analysis during the unfolding of the extreme-weather event. Under this analysis, the trajectory from the hurricane is taking into account to solve the *T-B* from section 7.2.1, to estimate the probability of the component failure exposed by the extreme-weather event. The proposed component model is shown in Figure 7.22. Each step is described below.
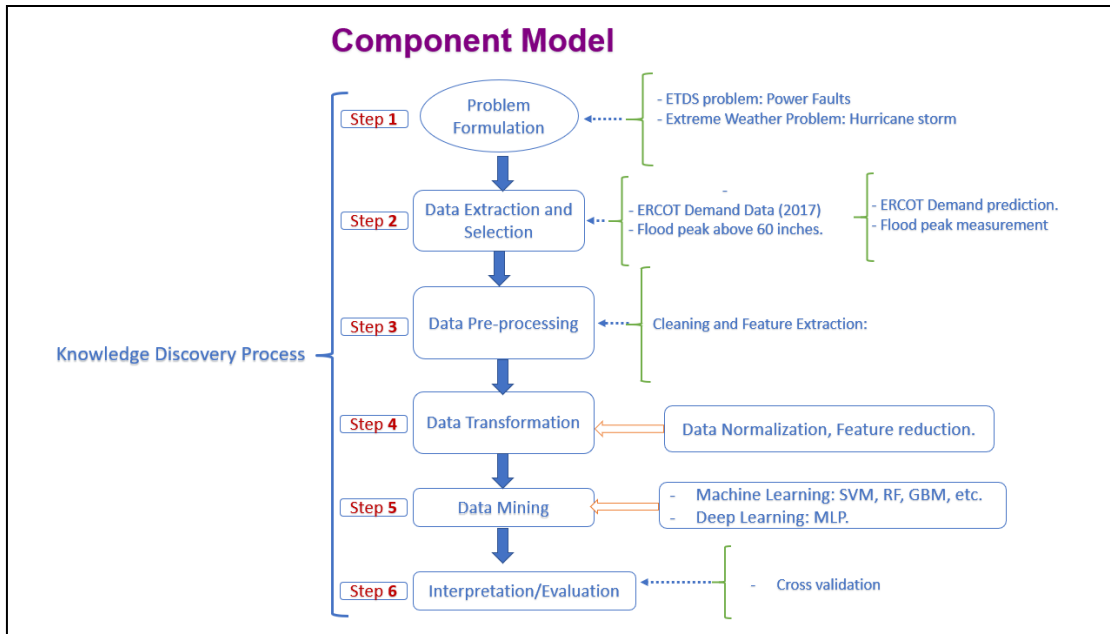
Figure 7.22: Component model showing the detailed steps.

7.5.1 STEP 2. DATA EXTRACTION AND SELECTION

In this step, data extraction and selection are explained.

7.5.1.1 Data Extraction

Data extraction was from several sources taking from the *results* of the extreme-weather model.

7.5.1.2 Data Selection

The data collection is based on information from ERCOT web site [Erco19]. Load data is based on a time sequence from the Hurricane Harvey event, carried out in preparation for this

scenario. The time horizon for point of prediction is one year from January to December of 2017. The **data collection for peak flood** analysis is based on the United States Geological Survey (USGS) data report [Usgs18]. The data flow is presented in Appendix C.2. Figure 7.23 shows the total rainfall and its fragments during Hurricane Harvey [Blac18], taken into account for this analysis.
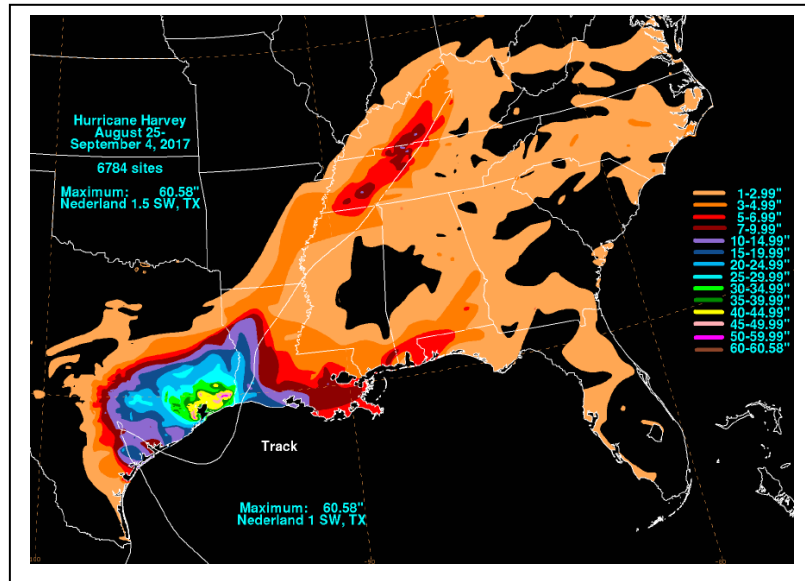
Figure 7.23: Harvey observed total rainfall (inches) and its fragments, image taken from [Blac18].

7.5.2 STEP 3. DATA PRE-PROCESSING

Missing values and outliers were removed similarly as analysis-1, in which, the same dataset (excel tabular format) was used to this analysis.

Descriptive statistics and visualization were used in this step for the experimentation of machine learning, but was omitted for the experimentation of deep learning neural network algorithms for simplicity purposes. Thus, the objective for ML experimentation is the application of the python software tool methodology to descriptive statistics. The same technique was applied from the Scenario-2 analyis-2 presented in Appendix A.1.2.

The dataset used in ML experimentation contains a total of *8785 data points* with a time horizon of one year (*January – December' 2017*), including *ERCOT* data information of the eight weather sectors of the hourly electricity demand. The pairwise correlation applying the *Pearson's correlation coefficient between* variables is shown in the correlation matrix on Figure 7.19 with the corresponding values shown in Table 7.1, respectively. The description of the variables used are presented below, each variable represents the weather region zones ERCOT clustering.

ncent: North central region including cities like Dallas-Fort Worth.

scent: South central region including cities like Austin and San Antonio.

coast: Coast region including cities like Houston.

east: East region including cities like Tyler.

north: North region including cities like Wichita Falls.

west: West region including cities like Abilene.

fwest: Far West region including cities like Midland.

south: South region including cities like Corpus Christi.

ERCOT: Total regions.
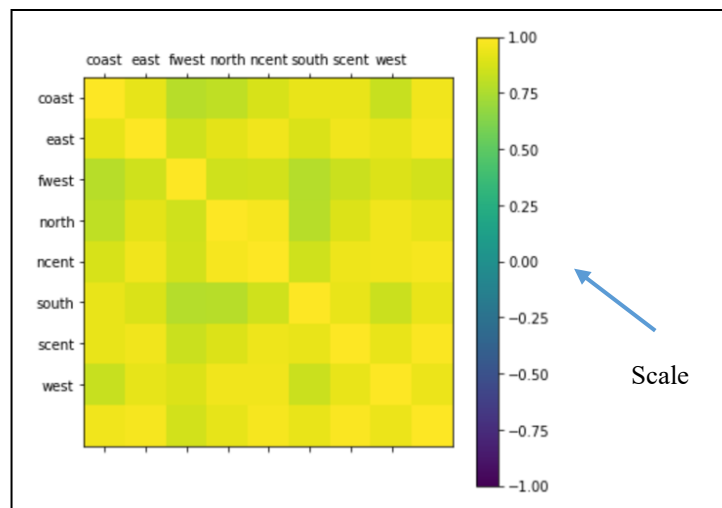
Figure 7.19. Correlation matrix of data attributes in ML experimented algorithms.

Table 7.1: Pearson's correlation coefficient of data attributes in ML experimentation.

	coast	east	fwest	north	ncent	south	scent	west	ERCOT
coast	1.00	0.92	0.78	0.81	0.88	0.94	0.94	0.83	0.96
east	0.92	1.00	0.85	0.92	0.96	0.89	0.96	0.92	0.98
fwest	0.78	0.85	1.00	0.85	0.86	0.78	0.84	0.89	0.86
north	0.81	0.92	0.85	1.00	0.97	0.79	0.90	0.95	0.93
ncent	0.88	0.96	0.86	0.97	1.00	0.86	0.95	0.95	0.97
south	0.94	0.89	0.78	0.79	0.86	1.00	0,93	0.84	0.94
scent	0.94	0.96	0.84	0.90	0.95	0.93	1.00	0.93	0.99
west	0.83	0.92	0.89	0.95	0.95	0.84	0.93	1.00	0.94
ERCOT	0.96	0.98	0.86	0.93	0.97	0.94	0.99	0.94	1.00

The observed correlation between all the pairs variables shows positive correlation indicating the relation between all the divided weather regions. The less correlated regions were the pair *fwest* with the *coast*, probable of this behavior is the fact that ERCOT electricity production are concentrated in the coast (wind generation). Figure 7.20 shows the density function of all the regions.
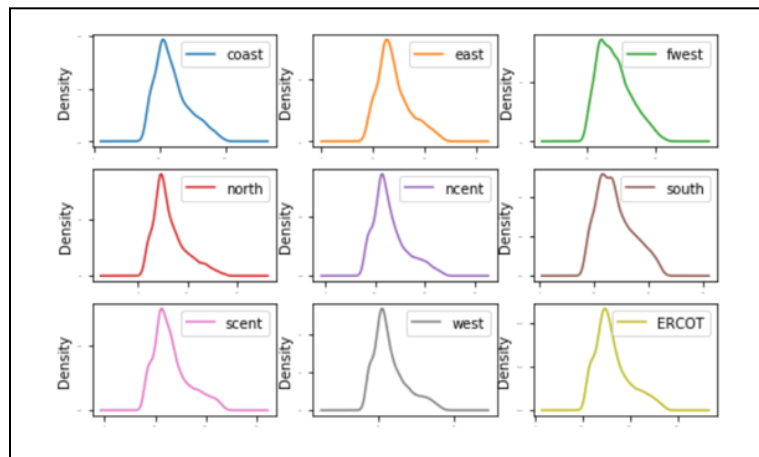
Figure 7.20. Density function of the hourly electricity dataset.

Most of the attributes of all regions show an *exponential distribution* with a left tendency. This singularity shows electricity demand consumption larger at the beginning of the year. We can conclude that electricity demand in Texas was not significatively impacted when Hurricane Harvey was hitting Texas during the end of August and beginning on September of 2017.

7.5.3 STEP 4: DATA TRANSFORMATION

In this analysis, data was normalized and standardized during the process of data mining with the same motivation as analysis-1.

7.5.4 STEP 5: DATA MINING:

The whole dataset was divided in two groups: the ***training set*** consists of 80% of the data used in this analysis, and the ***validation set*** consists of 20%, representing the rest of the dataset. The validation set is discussed in the next step. The training set is applied during the process of machine learning modeling implied in this analysis. Additionally, this set was specified as a random seed ensuring that the data was divided in a set of 7 to ensure generalization and prevent overfitting in data.

7.6 Experimented Data mining algorithms

The analysis is split in order to experiment with Machine Learning algorithms and multi-layer perceptron algorithm in a Deep Learning Neural Network.

7.6.1 EXPERIMENTED MACHINE LEARNING ALGORITHMS:

The problem formulation from step 1 and the observation in the data behavior shows two types of data: linear and non-linear, which can be modeled as a *regression model*. Thus, a *machine learning regression prediction* is proposed here, to solve the *electricity demand* prediction problem during an extreme-event. Several machine learning algorithms for regression analysis were discussed in chapter 4. For extended explanation, refer to [Mull16]. The libraries explored are shown in Appendix D 2.4. The results of this experimentation are reported in the interpretation section, which follows.

7.6.2 STEP 6: INTERPRETATION / EVALUATION (MACHINE LEARNING ALGORITHMS)

The interpretation is shown first and then the evaluation of the machine learning algorithms.

7.6.2.1 Interpretation (Machine Learning algorithms)

The *k-fold cross validation* was used in this experimental machine learning. Splitting the data in *K* partitions of equal size, where *i* partition represents the evaluation partition of a remaining *K-1* partition for training. The set was split with the same number of *k-folds* as scenario-1 analysis-2: 10 k-parts. The results are expressed in terms of *mean-square-error*. Table 7.2 shows the results of the preliminary evaluation of this experimented ERCOT demand prediction of 2017. The meaning of the six different algorithms was detailed in the past section 6.2.4.1.

Table 7.2: Mean-Square-Error and Standard deviation of the preliminary ERCOT demand prediction of 2017.

	LR	LASSO	EN	KNN	CART	SVR
MSE	2034.782673	2034.78178	2034.781121	2278.88654	2820.99450	49887.73368
SD	113.288573	113.316929	113.303062	211.223795	150.725277	3819.430337

The observed estimation accuracy of the models shows a tight distribution between the LR, LASSO, EN and nearly CART algorithms. Since data is not yet standardized, probably the SVR algorithm is far from the rest. Next, standardization of the datasets, including the scaling in the dataset are applied as a data transformation to avoid error in the performance. Table 7.3 shows the results of the second evaluation of the experimental ERCOT demand prediction including the standardization and scalation on the demand datasets.

Table 7.3: Mean-Square-Error and Standard deviation of the second ERCOT ML algorithms test. Standardization and scalation on datasets.

	LR	LASSO	EN	KNN	CART	SVR
MSE	2034.782673	2072.213144	3416.104073	1344.06287	2804.72361	4313.07846
SD	113.288573	131.156155	293.632412	96.594996	156.294500	316.340841

The observed estimation accuracy of the models shows KNN as the lowest error, as analysis-1. There is evidence of the good performance of the KNN in the literature, thus, the good behavior is justified here. Different hyperparameters was applied in this analysis to find the optimum. The values used were *n_neighbors*: 1,3,5,7,9,11,13,15,17,19,21. Table 7.4 shows the different parameters for the *n_neighbors*.

Table 7.4: Mean-Square-Error and Standard deviation of the different tuning hyperparameters used in KNN algorithm.

N_neighbors:	MSE	SD
1	1790.17186	112.128953
3	1365.44992	103.016692
5	**1345.89268**	**96.543729**
7	1382.36021	93.641170
9	1425.01895	90.818981
11	1452.79482	90.422161
13	1482.60760	92.650596
15	1511.58812	90.140230
17	1545.26181	88.957689
19	1577.23062	89.980424
21	1597.17240	89.160144

The observed estimation accuracy of the different hyperparameters shows the 5 *N_neighbors* as the best with 1345.89 of MSE. A combination of several machine learning techniques were performed in this analysis. The same algorithms from scenario-1 analysis-2 were used. Table 7.5 shows the results of the evaluation of the experimental machine learning assembled models. The meaning of the four different algorithms was presented in the past section 6.2.4.1.

Table 7.5: Mean-Square-Error and Standard deviation of the different ensemble algorithms.

Ensemble algorithm:	MSE	SD
AB	2502.59706	128.677489
GBM	1710.16247	108.256545
RF	1517.04058	100.424514
ET	**1392.13334**	**74.118990**

The observed estimation accuracy of the different ensembles shows ET as the best with 21392.13 of MSE. Thus, algorithm ET needed to be tuned, as discussed before, to test different hyperparameters. The values used were *n_estimators*: 50, 100, 150, 200, 250, 300, 350, 400. Table 7.6 shows the different parameters for the *n_estimators*.

Table 7.6: Mean-Square-Error and Standard deviation of the different tuning hyperparameters used in ET algorithm.

N_estimators:	MSE	SD
50	1252.695807	46.907031
100	1227.791719	53.546919
150	1232.062253	51.583780
200	1218.557543	56.043937
250	1218.747733	60.690772
300	1217.286517	52.836153
350	**1213.843901**	**51.624784**
400	1219.510958	55.441221

The observed estimation accuracy of the different hyperparameters shows the *350 estimators* with *1213.84* MSE as the best for the ET algorithm.

7.6.2.2 Evaluation (Machine Learning Algorithms)

This final step presents the final configuration of the model as "*the final model*," in this analysis, the ***Extra Tree Regressor (ET)***, which demonstrates having the least MSE. Thus, the final model needs to be trained with the entire training dataset. Prior to this the data needs to be standardized and scaled to better modeling performance as discussed above. The **final MSE result** of the **ET** is estimated as **1208.36**.

7.6.3 EXPERIMENTED MULTI-LAYER PERCEPTRON (MLP) IN DEEP LEARNING NEURAL NETWORK:

7.6.3.1 Interpretation MLP Deep Learning Neural Network Algorithms (DLNN)

From the step-2 in section 7.5.1.2 we formulate a new problem to prove the methodology presented in figure 7.22. The new task will solve a *Binary Classification* problem using peak flood data in a *Multi-Layer Perceptron* framework to experiment *Deep Learning Neural Network*. As discussed in chapter 4, evidence shows that deep learning algorithm improve the performance in data. The libraries explored are shown in Appendix D 3.2. The data transformation was using an *encoded* value of *[0,1]* for *binary classification*. Moreover, the standardization of the data was performed during the evaluation process, within each fold of the cross validation in this experimental analysis. The results are shown in the next step under the interpretation section.

7.6.3.2 First test DLNN: Parameters

Test dataset and train dataset was divided and treated with the same parameters discussed in section 7.5.4. Next, the baseline model was created with the following parameters:

First Dense Layer: 60 neurons.   Activation function: *relu.*

Second Dense Layer: *30 neurons.*   Activation function: *relu.*
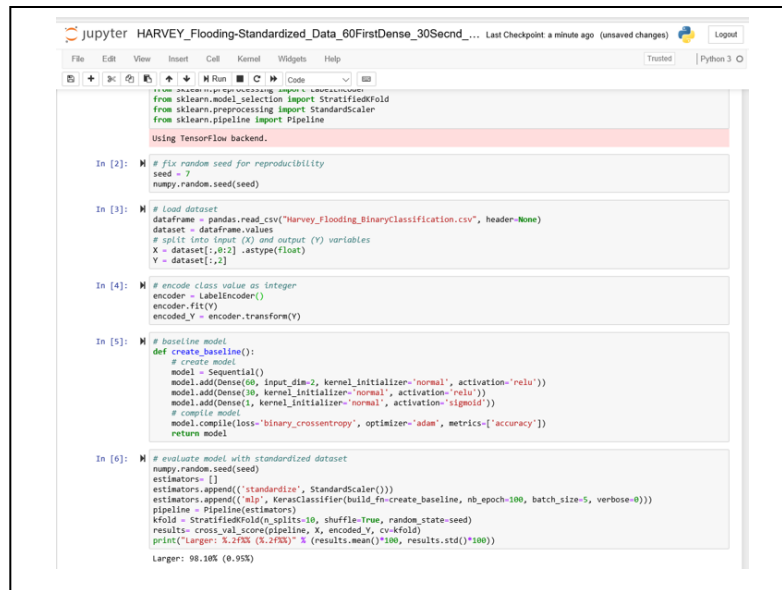
Third Dense Layer: *1 neuron*.   Activation function: *sigmoid.*

Data: *Standardized.*

Number of epochs: *100.*

7.6.3.3 First test DLNN: Interpretation/evaluation

The model used the ***binary cross entropy*** as the logarithm loss function with ***Adam*** optimization algorithm for gradient descent as discussed in chapter 4. The accuracy results show ***98.01% with a standard deviation of 0.95%.*** Figure 7.21 shows a screen capture of the results in python programming of the First test using MLP Deep Learning Neural Network.

Figure 7.21. Results in python programming of the First test using MLP Deep Learning Neural Network.

7.6.3.4 Second test DLNN: Dropout regularization

In this test, *Dropout regularization* was implemented in the hidden layers and an increment in the number of epochs for better performance as discussed in chapter 4. Additionally, the ***optimizer*** was changed to ***Stochastic Gradient Descent (SGD).*** Test dataset and train dataset was divided and treated with the same parameters discussed in section 7.5.4. Next, the baseline model was created with the following parameters:

First Dense Hidden Layer: *60 neurons.*   Activation function: *relu.*

First Dropout: Hidden Layer.

Second Dense Hidden Layer: *30 neurons.*   Activation function: *relu.*

Second Dropout: Hidden Layer.

Third Dense Layer: *1 neuron.*   Activation function: *sigmoid.*

Data: *Standardized.*   Learning rate: *0.1*

Number of epochs: *300.*   Momentum: *0.9*

7.6.3.5 Interpretation/evaluation first DLNN

The loss function was binary cross entropy with *SGD optimizer*. The accuracy results show **99.60% with a standard deviation of 0.53%.** Figure 7.22 shows a screen capture of the results in python programming of the Second test using *MLP Deep Learning Neural Network.*

Figure 7.22. Results in python programming of the Second test using MLP Deep Learning Neural Network.

7.7 System model.

The risk of a blackout in power systems is higher during an extreme-weather event, as discussed in section 1.4.2. In such cases, a blackout is a breakdown in the electric power grid

covering a large area due to the cascading of power failures, in which the loss of an element in the power grid increases the stress on other elements, resulting of a power outage [Ceti18]. Multiple power outages could happen in critical zones like hospitals, fire stations, gas stations, water supplied systems, and shelters.

In an extreme-weather event, the expected *Loss of Load (LOL)* is an index to evaluate the percentage of failures for reliability proposes. Therefore, it is crucial for the operator to take the best decision by evaluating the system in a real-time panorama. The *Optimal Power Flow* (OPF) as discussed in chapter 3 is the most used for a steady state solution on the power network. In most countries, the power grid is operated by the *Transmission System Operators (TSO)*. Thus, *OPF* seeks to minimize the *LOL*. It is also the objective of the *TSO*. The *OPF* is the most difficult task by the *Transmission Systems Operators (TSOs)* which consist of two main operation strategies: a) *Preventive action*, and b) *Corrective action*. The *TSO* needs to take into account a series of constraints for those two actions, as discussed in section 2.5.1. However, the objective function of the *DC OPF* only applies for one particular instant of time. During the path of an extreme-event new scenarios are created in the topology of the power system. Under this dynamic, the *TSO* needs to act in the nearly new topology arrangement. Such a dynamic process should be carried out within a dynamic *OPF* as the stated by the authors [Wang17] and [Sang19]. The transition states probabilities T_{prob} can be expressed by equation 7.1 [Wang17].

$$T_{prob}(S_{path,t}, S_{path',t+1}) = \prod_{k \in \Omega_{C,t+1}} T_{prob}(s_{k,t},\ s_{k,t+1}),\ prob \in \Omega_{S,t} \qquad (7.1)$$

where $\Omega_{S,t}$ represents the set of states at time t, and $S_{path,t}$ and $S_{path',t+1}$ are the time t and $t+1$, respectively. $\Omega_{S,t}$ is a set of Markov states. In the next proposed model several cases were created to represent the system states.

The **System model** refers to the system analysis during the unfolding of the extreme-event. Under this analysis, the trajectory of the hurricane is taken into account to solve the *T-C* from section 7.2.1, to estimate the system failures exposed by the extreme-weather event. This model is shown in Figure 7.23. Each step is described below.
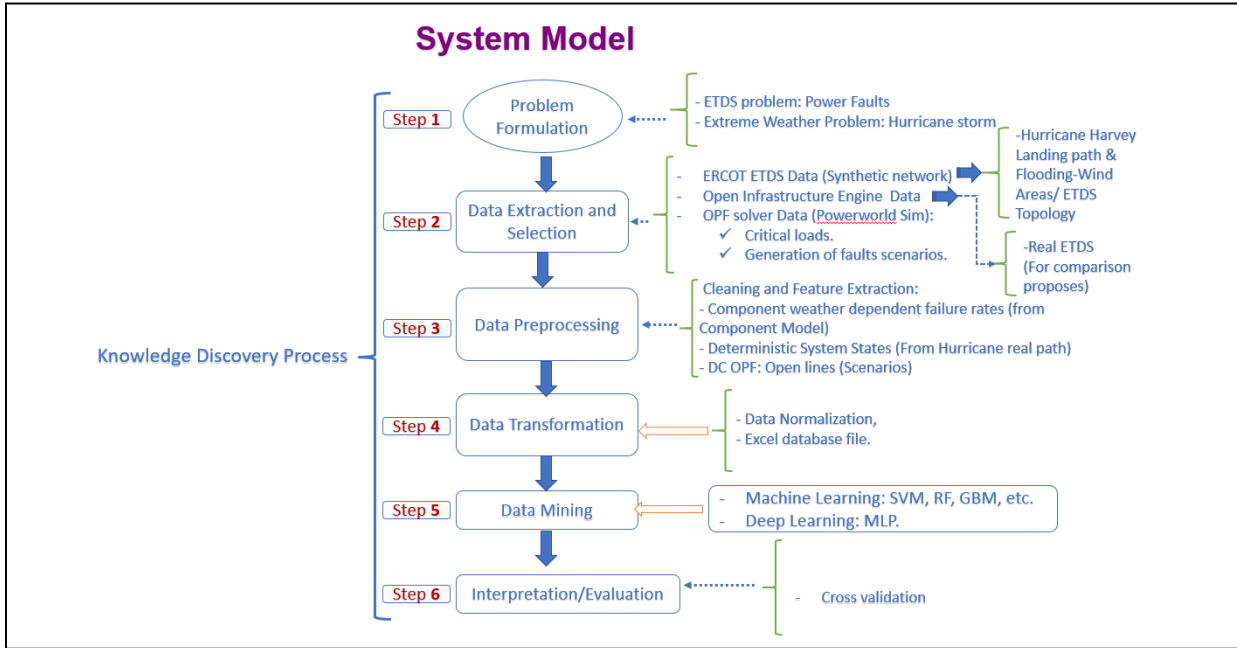
Figure 7.23: System model showing the detailed steps.

7.7.1 STEP 2. DATA EXTRACTION AND SELECTION

In this step, data extraction and selection are explained.

7.7.1.1 Data Extraction

To simulate several system scenarios (*cases*) as discussed above, the data extraction is taken from the *IEEE 9-bus*, which is a *hypothetical topology*, as discussed in section 2.1. Appendix A.2 shows details of the *IEEE 9-bus* used in this analysis. Figure 7.24 shows the experimental power grid topology used in this study.
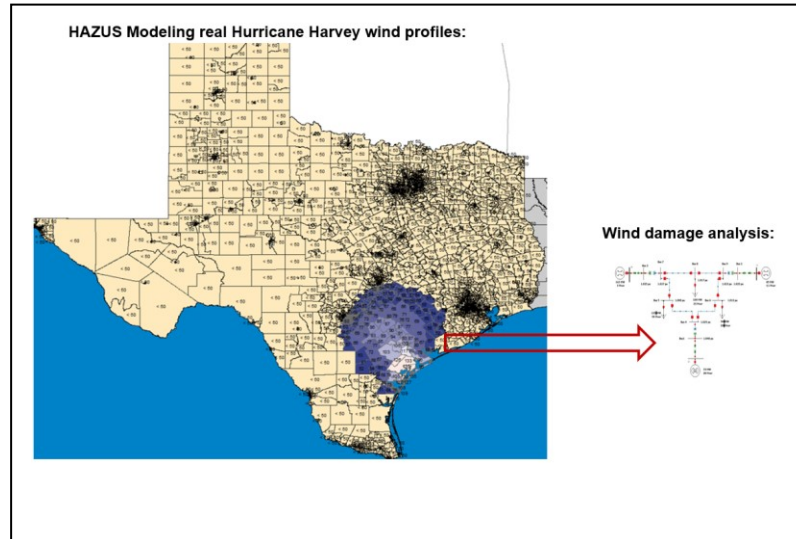
Figure 7.24. Experimental power grid topology for *system model* simulation.

7.7.1.2 Data Selection

The power failures are represented as line disconnections between the electric power grid. *79 cases* were developed to represent *one-line disconnection* and *two-line disconnection*s. Figure 7.25 shows an empirical outage power sequence, which occurred during Hurricane Harvey, and Figure 7.26 shows the hypothetical scenario with the cases created for this analysis.
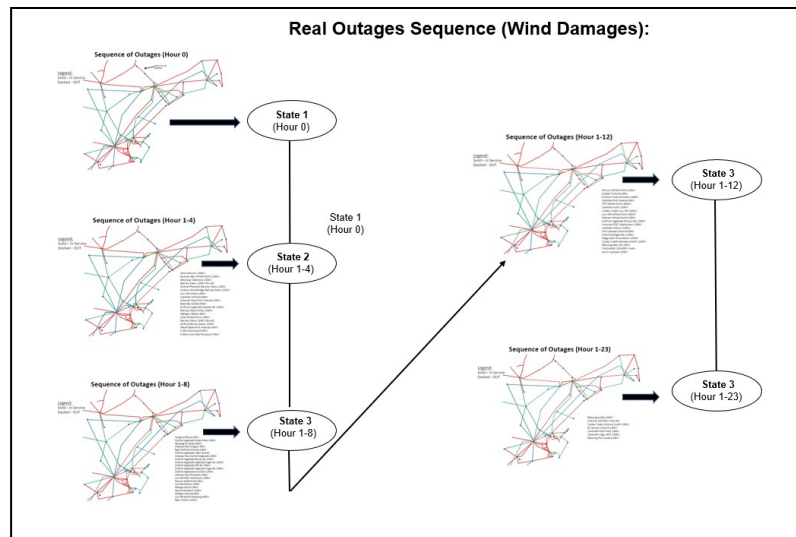
Figure 7.25. Empirical outages sequence during hurricane translation, image lightly modified from [Burf18].
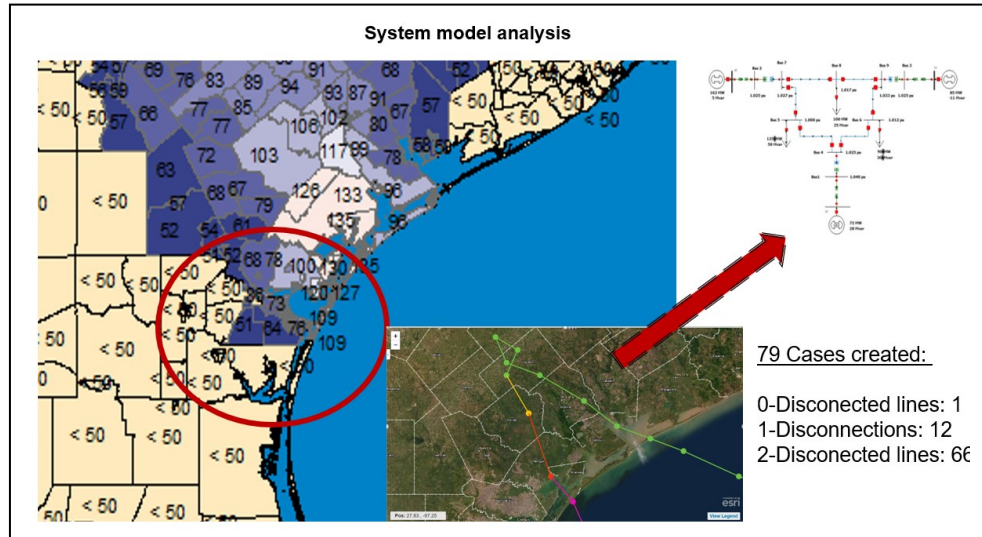
Figure 7.26. Hypothetical scenario with cases created for system model analysis.

7.7.2 STEP 3: DATA PRE-PROCESSING

Missing values and outliers were removed, similar to the past models, in which the dataset was formatted in Excel as table. Appendix C.3 shows the data flow used in this analysis. *Deep Learning Neural Network (DLNN)* experimentation contains a total of 719 data points. Descriptive statistics and visualization were used in this step for the experimentation of machine learning, but was omitted for the experimentation of deep learning neural network algorithms for simplicity purposes. *DLNN*s are discussed in a later section.

The dataset used in *ML* experimentation contains a total of 137 data points. The pairwise correlation applying the Pearson's correlation coefficient between variables is shown in the correlation matrix on Figure 7.27. The description of the variables used are presented below.

0: Injections active power (*MW*).

1: Ybus (DC-OPF solution).

2: 0: *No-line disconnected*. 1: *One-line disconnected*.

113
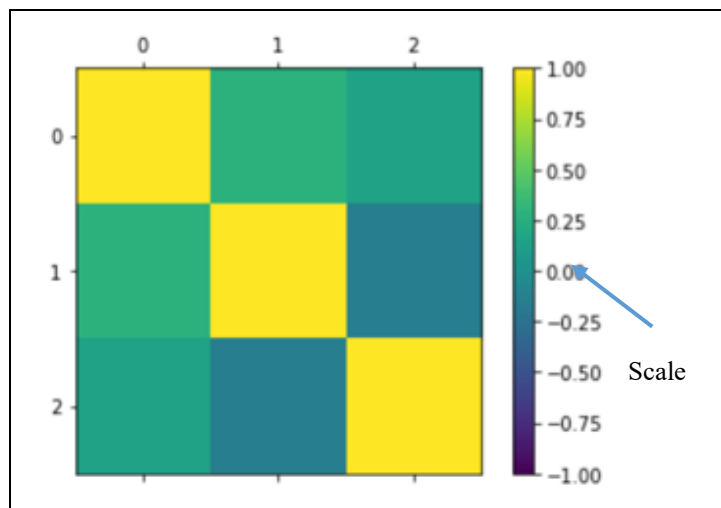
Figure 7.27. Correlation matrix of data attributes in *ML* experimented algorithms.

Figure 7.28 shows the density function of the experimental data.
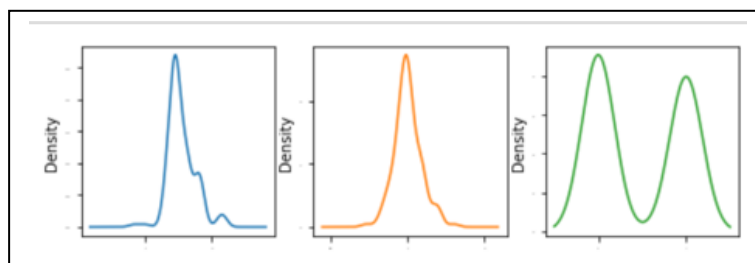
Figure 7.28. Density function of the cases dataset.

Most of the attributes of all regions shows a *Gaussian distribution* and *Double Gaussian* in the binary column classification as expected. This singularity shows a stable system, which is expected for the *DC OPF*.

7.7.3 STEP 4: DATA TRANSFORMATION

In this analysis, data was normalized and standardized during the process of data mining with the same motivation as analysis-1.

7.7.4 STEP 5: DATA MINING:

The whole dataset was divided in two groups: the ***training set*** with the 80% of the data used in this analysis, and the ***validation set*** consisting of the 20% representing the rest of the dataset. The

validation set is discussed in the next step 6. The training set is applied during the process of machine learning modeling implied in this analysis. Additionally, this set was specified as a random seed ensuring that the data was divided in a set of seven to ensure *generalization* and prevent *overfitting* in data.

7.8 Experimented Data mining algorithms

The analysis is split in two, with the discussion of Machine Learning algorithms first and multi-layer preceptor algorithm in a Deep Learning Neural Network second.

7.8.1 EXPERIMENTED MACHINE LEARNING ALGORITHMS:

The problem formulation from step 1 and the observation in the data behavior shows linear data, which can be modeled as a *binary classification model*. Thus, a *machine learning binary classification* is proposed here, to solve the *DC OPF* as *two cases (Disconnected or not disconnected)* binary classification problem during an extreme-event. Several machine learning for binary classification algorithms were analyzed. The libraries explored are shown in Appendix D 3.3. The results are shown in the next step under the interpretation section.

7.8.2 STEP 6: INTERPRETATION / EVALUATION (MACHINE LEARNING ALGORITHMS)

In this step, the interpretation is shown first and then the evaluation of the machine learning algorithms.

7.8.2.1 Interpretation (Machine Learning algorithms)

The *k-fold cross validation* was used in this experimental machine learning. Splitting the data in *K* partitions of equal size, where *i* partition represents the evaluation partition of a remaining *K-1* partition for training. The set was split with the same number of *k-fold* (10 k-parts) as the past models in this dissertation. The results are expressed in terms of *mean-square-error*. Table 7.7 shows the results of the preliminary evaluation of this experimented *binary classification*. The meaning of the six different algorithms are detailed below.

LR: Logistic Regression

LDA: Linear Discriminant Analysis

KNN: K-Neighbors Classifier.

CART: Decision Tree Classifier.

NB: Gaussian Naïve Bayes.

SVC: Support Vector Classifier Machines.

Table 7.7: Mean-Square-Error and Standard deviation of the preliminary binary classification.

	LR	LDA	KNN	CART	NB	SVC
MSE	0.605051	0.550673	0.623906	0.899327	0.734175	0.963131
SD	0.049495	0.023401	0.005724	0.026599	0.025084	0.018687

The observed estimation accuracy of the models shows a tight distribution between the *LR*, *LDA*, *KNN* and nearly *NB* algorithms. Since data is not yet standardized, probably *CART* and *SVC* algorithms are far from the rest, same as the other models. Next, standardization of the datasets, including scaling of the dataset, are applied as a data transformation operation to avoid error in the performance. Table 7.8 shows the results of the second evaluation of the experimental binary classification including the standardization and scalation on datasets.

Table 7.8: Mean-Square-Error and Standard deviation of the second ML Binary classification algorithms tested.

	LR	LDA	KNN	CART	NB	SVC
MSE	0.614478	0.550673	0.660606	0.908586	0.734175	0.789394
SD	0.021886	0.023401	0.006061	0.035859	0.025084	0.043939

The observed estimation accuracy of the models shows CART with the lowest error. Different hyperparameters were applied in this analysis to find the optimum. The values used were *depth*: *4, 5, 6, 7, 8, 9, 10*. Table 7.9 shows the different parameters for the *n_neighbors*.

Table 7.9: Mean-Square-Error and Standard deviation of the different tuning hyperparameters used in KNN algorithm.

N_neighbors:	MSE	SD
4	0.825688	0.025925
5	0.770642	0.029628
6	0.871560	0.035520
7	0.834862	0.038214
8	0.899083	0.026598
9	**0.908257**	**0.035857**
10	0.908257	0.035857

The observed estimation accuracy of the different hyperparameters shows the 9 *depth* as the best with *0.908257 of MSE*. A combination of several machine learning algorithms was performed in this analysis. Table 7.10 shows the results of the evaluation of the experimental machine learning assembled models. The meaning of the four different algorithms was presented in the past section 6.2.4.1.

Table 7.10: Mean-Square-Error and Standard deviation of the different ensemble algorithms.

Ensemble algorithm:	MSE	SD
AB	0.908586	0.035859
GBM	0.899327	0.026599
RF	0.899495	0.044949
ET	0.890067	0.017340

The observed estimation accuracy of the different ensembles shows *ET* as the best with *0.890067 of MSE*.

7.8.2.2 Evaluation (Machine Learning Algorithms)

This final step presents the final configuration of the model as "*the final model*," in this analysis the ***Decision Tree Classifier (CART)*** demonstrates having the least MSE. Thus, the final model needs to be trained with the entire training dataset with a *max_depth* of 9. Prior to this, the data needs to be standardized and scaled to better modeling performance as discussed above. The **final MSE result** of the **CART** is estimated as **0.8571**. Figure 7.29 show the confusion matrix results of this ML binary classification with only 4 misclassifications.

```
0.857142857143
[[16  0]
 [ 4  8]]
            precision    recall  f1-score   support

       0.0       0.80      1.00      0.89        16
       1.0       1.00      0.67      0.80        12

avg / total       0.89      0.86      0.85        28
```

Figure 7.29. confusion matrix results of this ML binary classification.

7.8.3 EXPERIMENTED MULTI-LAYER PERCEPTRON (MLP) IN DEEP LEARNING NEURAL NETWORK:

7.8.3.1 Interpretation MLP Deep Learning Neural Network Algorithms (DLNN)

From the step-2 in section 7.1 a new problem is formulated to prove the methodology presented in figure 7.23. The new task will solve a *Binary Classification* problem using two-disconnected lines using a Multi-Layer Perceptron framework to experiment Deep Learning Neural Network. The libraries explored are shown in Appendix D 3.3. The data transformation was using an encoded value of [0,1] for binary classification. Moreover, the standardization of the data was performed during the evaluation process, within each fold of the cross validation in this experimental analysis. The results are shown in the next step under the interpretation section.

7.8.3.2 First test DLNN: Parameters

Test dataset and train dataset was divided and treated with the same parameters discussed in section 7.5.4. Next, the baseline model was created with the following parameters:

First Dense Layer: *60 neurons.* Activation function: *relu.*

Second Dense Layer: *30 neurons.* Activation function: *relu.*
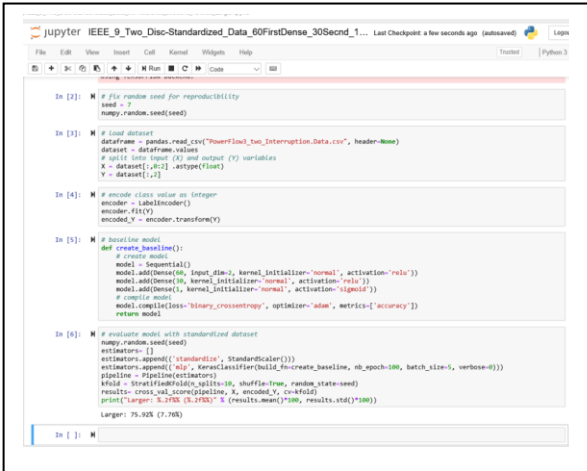
Third Dense Layer: *1 neuron.* Activation function: *sigmoid.*

Data: *Standardized.*

Number of epochs: *100.*

7.8.3.3 First test DLNN: Interpretation/evaluation

The model used the ***binary cross entropy*** as the logarithm loss function with ***Adam*** optimization algorithm for gradient descent as discussed in chapter 4. The accuracy results show ***75.92% with a standard deviation of 7.76%.*** Figure 7.30 shows a screen capture of the results in python programming of the First test using MLP Deep Learning Neural Network.

Figure 7.30. Results in python programming of the First test using MLP Deep Learning Neural Network.

7.8.3.4 Second test DLNN: Dropout regularization

In this test, Dropout regularization was implemented in the hidden layers and an increment in the number of epochs for better performance as discussed in chapter 4. Additionally, the optimizer was changed to Stochastic Gradient Descent (SGD). Test dataset and train dataset was divided and treated with the same parameters discussed in section 7.5.4. Next, the baseline model was created with the following parameters:

First Dense Hidden Layer: *60 neurons*.   Activation function: *relu*.

First Dropout: Hidden Layer.

Second Dense Hidden Layer: *30 neurons*.   Activation function: *relu*.

Second Dropout: Hidden Layer.

Third Dense Layer: *1 neuron*.   Activation function: *sigmoid*.

Data: *Standardized*.   Learning rate: *0.1*

Number of epochs: *300*.   Momentum: *0.9*

7.8.3.5 Interpretation/evaluation first DLNN

The loss function was ***binary cross entropy*** with ***SGD optimizer***. The accuracy results show ***90.26% with a standard deviation of 2.79%.*** Figure 7.22 shows a screen capture of the results in python programming of the Second test using MLP Deep Learning Neural Network.

Figure 7.22. Results in python programming of the Second test using MLP Deep Learning Neural Network.

7.9 Conclusion

A novel method to solve several tasks in the Electricity and Transmission Systems during an extreme-event is presented, and experimental results are summarized. Data Mining Algorithms were implemented by the use of Python software to solve regression predictive ML, binary classification ML, and binary classification DL algorithms. Moreover, Anaconda environment shows a good platform in the management of machine and deep learning model design. The Anaconda environment was used in all the experimental machine and deep learning algorithms. All the experimental data in this chapter could be used as a basis for researching other tasks in the ETDS or other natural disasters. GIS applications are promising tool as aggregation data to better visualize ETDS as a spatiotemporal solution during natural disasters.

CHAPTER 8: SUMMARY, CONCLUSIONS, AND FUTURE WORK

This chapter covers the conclusions of this study and suggests direction of future work on a higher level than contributed chapters, in the explorations of new applications identified during this research work.

8.1 Summary

Analysis of two Scenarios presented in this study are summarized below.

- ***Scenario-1 Analysis-1*** shows that during price forecasting, the uncertainty and non-linearity in the data contribute to inaccuracy in the results. The new method presented using de-noised wavelet as a pre-processing contributes to better results outcomes.

- ***Scenario-1 Analysis 2***, through the observed accuracy of the Machine Learning Regression Prediction algorithm, reveals that the best algorithm for the problem of Demand prediction during a winter storm was the *Gradient Boosting Regressor* with a Mean Square Error of 194.1136.

- ***Scenario-2 Extreme-Weather Model***, through the observed accuracy of the experimental *Supervised Machine Learning* algorithm, reveals that the best algorithm is the *Maximum Likelihood Classification* for the problem of "image classification" during a <u>Tropical Storm</u> (*Hurricane Harvey stage-2)*. It was verified by the *Barker Reservoir* digital image with a Kappa coefficient of 0.9399 and overall accuracy of 97.1968%

- ***Scenario-2, Component Model***, through the observed accuracy of the *Machine Learning Regression Prediction* algorithm, reveals that the best algorithm for the problem of demand prediction (ERCOT) during Hurricane Harvey was the *Extra Tree Regressor* algorithm, with *350 estimators* and MSE of 1208.36. A *binary classification* was presented to solve peak flood classification (above 60 inches) with a Deep network using *Stochastic Gradient Descent optimizer* and *Dropout in hidden layers*. The accuracy of the DLNN algorithm was 99.60 %.

- **Scenario-2 System Model**, through the observed accuracy of the *Machine Learning Binary Classification* algorithm, reveals that the best algorithms for the problem of *one-line disconnected* in hypothetical cases using the IEEE 9-Bus electric topology during an extreme-event is the *Decision Tree Classifier* algorithm, with *a max-depth of 9* and MSE of 0.8571. A *binary classification* was presented to solve *two-lines* disconnected in hypothetical cases in an IEEE 9-bus, using a *Deep Learning Neural Network*. The DLNN was set with a *Stochastic Gradient Descent optimizer,* including *binary cross entropy*, and *Dropout in hidden layers*. The accuracy of the DLNN algorithm was 90.26 %.

8.2 Conclusions

This study serves as proof of benefit of the consolidation of the techniques, methodologies, and modeling explained in chapters 3, and 4. A previous data analysis was conducted to understand the data sets, and its relation to each task in chapter 6 and 7. Moreover, in those chapters a "big picture" was framed in two practical scenarios to demonstrate the achievement of the specific task in data mining through a knowledge discovery framework, as the methodology proposed in this study. The results reveal a powerful tool to solve problem tasks in the area of critical operation of the Electricity Transmission and Distribution systems during an extreme-weather events.

8.3 Future work

From the two analyses conducted on Scenario-1, the exploration of more machine learning and deep learning algorithms is needed to solve the critical operation including the components in the Transmission and Distribution Systems (ETDS) under winter storms. From the multi-model analyses conducted on Scenario-2, the observed future work was to add a *Multi-objective optimization analysis (T-D)* in a post-storm scenario by including shelters as critical loads, i.e. hospital, gas station, etc. Also, it would be beneficial to include neighboring regional interconnectivity, and the possibility of using micro-grids and local mini-generators to enhance the rapid recovery of electricity during this last stage of an extreme-weather event.

Another research area is to conduct several experiments by running different scenarios in the ETDS, for example, tornados, earthquakes, or other natural disasters that are currently escalating, including man-made attacks by the exploration of cyber security areas.

Additionally, it is necessary to explore more experiments with larger datasets using the methods explained in chapter 4. By scaling the datasets, one could use new technology innovation in particular to upgrade the computing power achieved by GPUs, cloud platforms, and Hadoop systems with Spark platform. Larger data could improve deep learning algorithms for more efficient models for data prediction and classification. Furthermore, the use of real-time database management to update records to provide real-time problems.

In general, the use of Machine Learning, and especially Deep Learning, looks promising for the solution of several issues in the ETDS. It can be concluded that research in the electrical engineering area must be continued in order to better consolidate the ideas presented and the algorithms explained in this study. The work done in this dissertation document is the starting point for future applications in finding potential solutions in the Electricity Transmission and Distribution Systems under critical stress to lessen power failures.

REFERENCES

[Abu13] AbuOmar, O., Nouranian, S., King, R., Bouvard, J.L., Toghiani, H., Lacy, T.E., et al, "Data Mining and Knowledge Discovery in Materials Science and Engineering: A Polymer Nanocomposites Case Study", *Advanced Engineering Informatics*, Vol. 27, pp. 615-624 (2013)

[Adde16] Adderly, Sh., "Reviewing Power Outages Trends, Electric Reliability Indices and Smart Grid Founding", *Graduate College Dissertations and Theses*. Paper 531, pp. 1-96 (2016)

[Aept18] American Electric Power, "Hurricane Harvey update: 9-28-2017", *American Electric Power report update*, pp. 1-16 (2017)

[Agar16] Agarwal, P.K., De, D., and Kumar, H.R., "Modernization of Multi Location Live SCADA System-A Case Study", *2016 National Power Systems Conference* (NSPSC), pp. 1-6 (2016)

[Ahme08] Ahmed, M.M., and Soo, W.L., "Supervisory Control and Data Acquisition System (SCADA) Based Customized Remote Terminal Unit (RTU) for Distribution Automation System", *2nd International Power and Energy Conference*, pp. 1655-1660 (2008)

[Alve08] Alvehag, K., and Soder, L., "A Stochastic Weather Dependent Reliability Model for Distribution Systems", *Proceedings of the 10th International Conference on Probabilistic Methods Applied to Power Systems*, pp. 1-8 (2008)

[Anac19] Anaconda distribution. "The Most Popular Python Data Science Platform", Anaconda Inc., 2019, Available: http://www.anaconda.com/what-is-anaconda/

[Anba16] Anbalagan, B. and Valliyammai, C., "#ChennaiFloods: Leveraging Human and Machine Learning for Crisis Mapping during Disasters Using Social Media:. *IEEE 23rd International Conference on High Performance Computing Workshops* (HiPCW), pp. 50-59 (2016)

[Ange17] Angelov, P., Gu, X. and Principe, J., Fast Feedforward Non-Parametric Deep Learning Network with Automatic Feature Extraction, *2017 International Joint Conference on Neural Networks (IJCNN),* pp. 534-541 (2017)

[Arab16] Arabali, A., Majidi, M, Fadali, M.S., and Etezadi-Amoli, M., "Line Outage Identification-Based State Estimation in a Power System with Multiple Line Outages", *Electric Power Systems Research*, vol. 133, pp. 79-86 (2016)

[Azad15] Azadzadeh, A., Kötter, T., and Zebardast, E., "An Augmented Approach for Measurement of Disaster Resilience Using Connective Factor Analysis and Analytic Network Process (F'ANP) model", *International Journal of Disaster Risk Reduction*, No.14, pp. 504-518 (2015)

[Azad17] Azadzadeh, A., Kötter, T., Salehi, P., and Birkmann, J., "Operationalizing a Concept: The Systematic Review of Composite Indicator Building for Measuring Community Disaster Resilience", *International Journal of Disaster Risk Reduction*, No. 25, pp. 147-162 (2017)

[Babr96] Babri, H.A. and Tong, Y., Deep Feedforward Networks: Application to Pattern Recognition, *IEEE International Conference on Neural Networks*, Vol.3, pp. 1422-1426 (1996)

[Bake17] Baker, M., "The Decade's Worst Hurricanes". *Coastal Living*, 2017. Available: http://coastalliving.com

[Band17a] Bandaru, S., Ng, A.H.C and Deb, Kalyanmoy, D., Data Mining Methods for Knowledge Discovery in Multi-Objective Optimization: Part A- Survey, *Expert Systems with Applications*, Vol. 70, pp. 139-159 (2017)

[Band17b] Bandaru, S., Ng, A.H.C and Deb, Kalyanmoy, D., Data Mining Methods for Knowledge Discovery in Multi-Objective Optimization: Part B- New Developments and Applications, *Expert Systems with Applications*, Vol. 70, pp. 119-138 (2017)

[Beng11] Bengio, Y. and Delalleau, O., On the Expressive Power of Deep Architectures, *Proceedings of the 22md International Conference on Algorithmic Learning Theory,* pp. 18-36 (2011)

[Berg08] Berg, R., "Tropical Cyclone Report: Hurricane Ike", *National Hurricane Center* (AL092008), pp. 1-14 (2008)

[Bhui94] Bhuiyan, M.R., and Allan, R.N., "Inclusion of Weather Effects in Composite System Reliability Evaluation Using Sequential Simulation", *IEEE Proceedings – Generation, Transmission and Distribution"*, Vol. 141, No.6, pp. 575-584 (1994)

[Bill02] Billinton, R., and Singh, G.D., "Reliability Assessment of Transmission and Distribution Systems Considering Repair in Adverse Weather", *IEEE CCECE2002 Canadian Conference on Electrical and Computer Engineering. Conference Proceedings (Cat. No. 02CH37373)*, Vol. 1, pp. 88-93 (2002)

[Billi68] Billinton, R., and Bollinger, K.E., "Transmission System Reliability Evaluation Using Markov Process", *IEEE Transactions Power Apparatus and Systems*, Vol. PAS-87, No.2, pp. 538-547 (1968)

[Birc17] Birchfield, A.B., Xu, T., Gegner, K.M.G., Shetye, K.S., and Overbye, T.J., Grid structural characteristics as validation criteria for synthetic networks, *IEEE Transactions on Power Systems*, vol.32, no. 4, pp. 3258-3265 (2017)

[Blak18] Blake, E.S., and Zelinsky, A.D., National hurricane center tropical cyclone report: Hurricane Harvey (AL092017). *U.S. Department of commerce. National Oceanic and Atmospheric Administration and The National Weather Services*, pp. 1-77 (2018)

[Blak18] Blake, E.S., and Zelinsky, D.A., "Tropical Cyclone Report: Hurricane Harvey", *National Hurricane Center* (AL092017), pp. 1-76 (2018)

[Brow18] Brown, P., Daigneault, A.J., Tjernstorm, E., and Zou, W., "Natural Disaters, Social Protection, and Risk Perceptions", *World Development*, vol. 104, pp. 310-325 (2018)

[Brow97] Brown, R.E., Gupta, S., Christie, R.D., Venkata, S.S., and Fletcher, R., "Distribution System Reliability Assessment: Momentary Interruption and Storms", *IEEE Transactions on Power Delivery*, vol.12, No.4, pp. 1569-1575 (1997)

[Budu17] Buduma, N., Fundamentals of Deep Learning, O'Relly Media. Inc., pp. 1-298 (2017)

[Burf18] Burford, B., "2017 Hurricane Harvey", *American Electric Power, Texas A & M Relay Conference*, pp.1-35 (2018)

[Cand17] Candel, A., Parmar, V., LeDell, E., and Arora, A., Deep Learning with H20, *H20.ai, Inc.,* pp. 9-10 (2017)

[Carl12] Carlson, L., Bassett, G., Collins, M., Folga, S., Haffenden, B. et al, "Resilience: Theory and Applications", *Argonne National Laboratory*, pp. 1-64 (2012)

[Chen09] Cheng, D., Liang, Y., Zhu D., and Broadwater R.P., "Real-Time Power Electric System Modeling, Assessment and Reliability Prediction", *2009 IEEE/PES Power Systems Conference and Exposition*, pp. 1-6 (2009)

[Chen18] Cheng, Y., Chen, K., Sun, H., Zhang, Y. and Tao, F., "Data and Knowledge Mining with Big Data Towards Smart Production", *Journal of Industrial Information Integration,* Vol. 9, pp. 1-13 (2018)

[Chol18] Chollet, F., Deep Learning with Python, *Manning Publications, Co.,* pp. 1-360 (2018)

[Cios07] Cios, K.J., Pedrycz, W., Swiniarski, R.W., and Kurgan, L.A. Data Mining: A Knowledge Discovery Approach*, Springer Science + Business Media, LLC*, pp. 3-7, (2005)

[Cois05] Cios, K.J., Pedrycz, W., Swiniarski, R.W., and Kurgan, L.A., "Data Mining: A Knowledge Discovery Approach", *Springer Science + Business Media*, LLC, pp. 3-7, (2005)

[Cui16] Cui, X. and Goel, V., Maximum Likelihood Nonlinear Transformation Based on Deep Neural Networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 11, pp. 2023-2031 (2016)

[Dahl12] Dahl, G.E., Yu, D., Deng, L. and Acero, A., Context_Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 1, pp. 30-42 (2012)

[Davi08] Davis, G.B., Olson, J. and Carley, K.M., "Unsupervised Plan Detection with Factor Graphs, Knowledge Discovery from Sensor Data: Second International Workshop", *Sensor-KDD*, pp. 59-75 (2008)

[Dema17] Demazy, A., Alpcan, T., Mareels, I., and Saha, S., "Assessment of Voltage Stability Risks Under Stochastic Net Loads Using Scalable SVM Classification", *Australasian Universities Power Engineering Conference* (AUPEC), pp. 1-6 (2017)

[Deme17] Demetriou, P., Asprou, M., Quiros-Tortos, J., and Kyriakides, E., Dynamic IEEE test systems for transient Analysis, *IEEE Systems Journal*, vol. 11, no. 4, pp. 2108-2117 (2017)

[Dyna19] Dynamic IEEE test systems web site, "IEEE 9-bus modified test system", *Center for Intelligent Systems & Networks with the University of Cyprus*, 2019, Available: https://www.kios.ucy.ac.cy/testsystems/index.php/about-us

[Egbu16] Egbue, O., Naidu, D., and Peterson, P., "The Role of Microgrids in Enhancing Macrogrid Resilience", *International Conference on Smart Grid and Clean Energy Technologies* (ICSGCE), pp. 125-129 (2016)

[Elec18] Electric Power Distribution web page, "Electric Power Distribution Manual", *Apprentice-Level Library for the Skilled Trades*, 2018. Available: http://free-ed.net/free-ed/SkilledTrades/Electrical%20Power/07ElecDistrib/07ElecDistribFra.asp

[Ener16] Energetic Incorporated report, "Innovation Pathway Study: Smart Grid Thecnologies", *Energetic Incorporated*, pp. 1-36 (2016)

[Erco19] ERCOT web page, *Hourly load data archives: 2017 ERCOT hourly load data*. (2019) Available: https:// www.ercot.com/gridinfo/load/load_hist

[Erjo08] Erjongmanee, S., Ji, Ch., Stokely, J. and Hightower, N., "Large-Scale Inference of Network-Service Disruption Upon Natural Disasters, Knowledge Discovery from Sensor Data: Second International Workshop", *Sensor-KDD*, pp. 134-153 (2008)

[Espi07] Espiritu, J.F., Coit, D.W., and Prakash, U., "Component Criticality Importance Measures for the Power Industry", *Electric Power System Research*, vol. 77, pp. 407-420 (2007)

[Fayy96] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, Vol. 17, pp. 37 (1996)

[Feiz13] Feizifar, B., Haghifam, M.R., Soleymani, S., and Jamilazari, A., "Fault Location in Combined Overhead line and Underground Cable Distribution Networks using Fault Transient Based Mother Wavelets", *12th International Conference on Environment and Electrical Engineering,* pp. 41-45 (2013)

[Fema18] FEMA web page, Hurricane Maria, *U.S. Department of Homeland Security FEMA*, 2018, Available: https://www.fema.gov/hurricane-maria

[Fema19] FEMA web page, *HAZUS: Earthquake, wind, flood, tsunami*. Ver. 4.2, (2019) Available: https:\\www.fema.gov/hazus.

[Flac12] Flach, P., Machine Learning: The Art and Science of Algorithms that Make Sense of Data, *Cambridge University Press*, pp. 13-48, 231-261 (2012)

[Fu15] Fu, Y. G., Zhu, J.M. and Zhang, N., "Analysis on information value of big data in internet finance. International Conference on Logistics", *Information and Service Sciences* (LISS), pp. 1-6 (2015)

[Gale18] Galea, Ch. And Farrugia, R.A., Matching Software-Generated Sketches to Face Photographs with a Very Deep CNN, Morphed Faces, and Transfer Learning, *IEEE Transactions on Information Forensics and Security*, Vol. 13, no. 6, pp. 1421-1431 (2018)

[Gave64] Gaver, D.P., Montmeat, F.E., and Patton, A.D., "Power Systems Reliability I-Measures of Reliability and Methods of Calculation", *IEEE Transactions Power Apparatus and Systems*, vol.83, No.7, pp. 727-737 (1964)

[Glov12] Glover, J. D., Sarma, M. S., and Overbye, T.J., Power system analysis and design, *Cengage Learning*, pp. 1-828 (2012)

[Gonz13] González-Fernández, R.A., Leite da Silva, A.M., Resende, L.C., and Schilling, M.T., "Composite Systems Reliability Evaluation Based on Monte Carlo Simulation and Cross-Entropy Methods", *IEEE Transaction on Power Systems*, Vol. 28, No. 4, pp. 4598-4606 (2013)

[Good16] Goodfellow, I., Bengio, Y., and Courville, A., Deep Learning – Introduction, *MIT Press*, pp. 1-26, 113, 493-510, 654-665 (2016)

[Goog19] Google Earth Pro software applications ver. 7.3.2.5776 (64 bits), *Hurricane Harvey 2017,* Data SIO, NOAA, U.S. Navy, NGA, GEBSO, U.S. Dept. of State Geographer (2019)

[Gull15] Gullo, F., "From Patterns in Data to Knowledge Discovery: What Data Mining Can Do", *Physics Procedia*, Vol. 62, pp. 18-22 (2015)

[Guo18] Guo, Z., Zhou, K., Zhang, X. and Yang, Sh., A Deep Learning Model for Short-Term Power Load and Probability Density Forecasting, *Energy*, No. 160, pp. 1186-1200 (2018)

[Hilb07] Hilber, P., and Bertling, L., "Component Reliability Importance Indices for Electrical Networks", *2007 International Power Engineering Conference (IPEC 2007),* pp. 257-263 (2007)

[Hint12] Hinton, G. et al., Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups*, IEEE Signal Processing Magazine*, vol. 29, no.6, pp. 82-97 (2012)

[Hutc08] Hutchins, J., Ihler, A. and Smyth, P., "Probabilistic Analysis of Large-Scale Urban Traffic Sensor Data Set, Knowledge Discovery from Sensor Data: Second International Workshop", *Sensor-KDD*, pp. 94-114 (2008)

[Ieee12] IEEE Standard Association, "IEEE Guide for Electric Power Distribution Reliability Indices", *IEEE Power and Energy Society Brochure*, pp. 1-31 (2012)

[Ieee99] Reliability Test System Task Force report, "IEEE Reliability Test System – 1996", IEEE Transactions on Power Systems, Vol. 4, No. 3, pp. 1010-1020 (1999)

[Izen08] Izenman, A.J., Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, *Springer Test in Statistics*, pp. 1-731 (2008)

[Kahn17] Kahnamouei, A. S., Bolandi, T.G., and Haghifarm, M-R., "The Conceptual Framework of Resilience and its Measurement Approaches in Electrical Power Systems", *IET International Conference on Resilience of Transmission and Distribution Networks* (RTDN 2017), pp. 1-11 (2017)

[Kaus14] Kausky, C., Informing Climate Adaptation: A Review of the Economic Costs of Natural Disasters", *Energy Economics*, vol. 46, pp. 576-592 (2014)

[Kell15] Kelleher, J.D., Namee, B.M. and D'Arcy, A., Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies- Machine Learning for Predictive Data Analytics, *The Massachusetts Institute of Technology (MIT) press*, pp. 1-20, 396-462, 513 (2015)

[Kenw14] Kenward, A., and Raja, U., Blackout: Extreme Weather, Climate Change and Power Outages, *Climate Central Technical Report,* 2014, Available: https://www.ourenergypolicy.org/wp-content/uploads/2014/04/climate-central.pdf

[Kepl17] Kepler, J. Deep Learning for Event Detection, Sequence Labelling and Similarity Estimation in Music Signals, *Doctoral Thesis Linz University,* pp. 14-17 (2017)

[Kim17] Kim, P., MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence, *Apress*, pp. 2-19 (2017)

[Knab06] Knabb, R.D., Brown, D.P., and Rhome, J.R., "Tropical Cyclone Report: Hurricane Rita", *National Hurricane Center*, pp. 1-33 (2006)

[Kuma14] Kumar, D., and Samantaray, S.R., "Reliability Optimization for Optimal Placement of Multiple Distributed Generators in Primary Distribution Network Using an Evolutionary Approach", *2014 Students Conference on Engineering and Systems*, pp. 1-6 (2014)

[Kuma17] Kumar, K., and Kumbhar, G.B., "A Review on Impact of Distributed Generation and Electrical Vehicles on Aging of Distribution Transformer", *3rd. International Conference on Condition Assessment Techniques in Electrical Systems* (CATCON), pp. 283-288 (2017)

[Li14] Li, G. et al, "Risk analysis for distribution systems in the northeast U.S. under wind storms", *IEEE Transactions on Power Systems,* vol.29, no. 2, pp. 889-898 (2014)

[Liao13] Liao, H., McDermott, E. and Senior, A., Large Scale Deep Neural Network Acoustic Modeling with Semi-Supervised Training Data for YouTube Video Transcription, *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 368-373 (2013)

[Lin16b] Lin, Y., and Bie, Z., "Study on the Resilience of the Integrated Energy System", *Energy Procedia*, vol. 103, pp. 171-176 (2016)

[Lisa15] Lisa Lab University of Montreal, Deep Learning Tutorial: Release 0.1, *Theano Development Team*, pp.1-167 (2015)

[Mast16] Masters, J., "Top Ten Global Weather/Climate Events of 2017: A Year of Landfalls and Firestorms", *WU Weather Underground News- The Weather Company,LLC*, 2016. Available: http://wunderground.com

[Mate15] Mateju, L., Cerva, P. and Zdansky, J., Investigation into the Use of Deep Neural Networks for LVCSR and Czech, *IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM),* pp. 1-4 (2015)

[Math18b] Mathworks web site, Machine Learning with MATLAB – How Do You Decide Which Algorithm to Use? *MathWorks* (2018) Available: https://www.mathworks.com/ (2018)

[Mccu43] McCulloch, W.S. and Pitts, W., A Logical Calculus of the Ideas Immanent in Nervous Activity, *Bulletin of Mathematical Biophysics*, Vol.5, pp. 115-133 (1943)

[Merc16] Mercado, R.M., "People's Risk Perceptions and Responses to Climate Change and Natural Disasters in BASECO Compound, Manila, Philippines", *Procedia Environmental Sciences*, vol. 34, pp. 490-505 (2016)

[Micr17] Microsoft, Data Mining SSAS web page. Data Mining, *Microsoft*, 2017, Available: https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/data-mining-ssas

[Mitc97]Mitchel, T.M., Machine Learning- Introduction, *McGraw-Hill Science/Engineering/Math*, pp. 1-20 (1997)

[Momo12] Momoh, J., "Functions of Smart Grid Components", *Smart Grid: Fundamentals of Design and Analysis,* pp. 12-14 (2012)

[Mont16] Monti, A., Muscas, C., and Ponci, F., "Phasor Measurement Units and Wide Area Monitoring Systems, from Sensor to the System", *Elsevier Inc.*, pp. 1-276 (2016)

[Mose14] Moser, S. C., Davidson, M.A., Kirshen, P., Mulvaney, P., Murley, J. F. et al., "Ch.25: Coastal Zone Development and Ecosystems. Climate Change Impacts in the United States:

The Third National Climate Assessment", U.S. *Global Change Research Program*, pp. 579-618 (2014)

[Mull17] Müller, A. C., Guido, S., Introduction to Machine Learning with Python, *O'Reilly Media Inc.*, pp. 1-378 (2017)

[Muna13] Munang, R., Thiaw, I., Alverson, K., Liu, J., and Han, Z., "The Role of Ecosystem Services in Climate Change Adaptation and Disaster Risk Reduction", *Current Opinion in Environmental Sustainability*, vol. 5, no. 1, pp. 47-52 (2013)

[Murp12] Murphy, K.P., Machine Learning: A Probabilistic Perspective- Introduction, *The Massachusetts Institute of Technology (MIT) press*, pp. 1-25, 46, 210, 999-1011 (2012)

[Naka17] Nakafuji, D., Rogers, L., Bestebreur, J., Rourke, M., and Zweigle, G., "Integrating Synchrophasors and Oscillography for Wide-Area Power System Analysis", *70th Annual Conference for Protective Relay Engineers* (CPRE), pp. 1-10 (2017)

[Naka17] Nakafuji, D., Rogers, L., Bestebreur, J., Rourke, M., and Zweigle, G., "Integrating Synchrophasors and Oscillography for Wide-Area Power System Analysis", *70th Annual Conference for Protective Relay Engineers (CPRE)*, pp. 1-10 (2017)

[Nasa19] Zoom earth application web page, *Storm Harvey 2017*, (2019) Available: https://zoom.earth/storms/harvey-2017

[Naza16] Nazaripouya, H., Wang, B., Wang, Y., Chu, P., Pota, H. R. and Gadh, R., Univariate time series prediction of solar power using a hybrid wavelet-ARMA-NARX prediction method, *2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*, pp. 1-5, (2016)

[Nell18] Nelli, F., Python Data Analytics: With Pandas, NumPy, and Matplotlib, *Apress*, pp.1-569 (2018)

[Newa19] NEWA. NEWA hourly weather data page. *NEWA web page*, (2019) Available: http://newa.cornell.edu

[Noaa17] NOAA National Centers for Environmental Information Web site, "U.S. 2017 Billion-Dollar Weather and Climate Disasters", *NOAA National Centers for Environmental Information (NCEI)*, 2017. Available: https://www.climate.gov/news-features/blogs/beyond-data/2017-us-billion-dollar-weather-and-climate-disasters-historic-year

[Noaa18] NOAA National Centers for Environmental Information Web site, "U.S. 2018 Billion-Dollar Weather and Climate Disasters", *NOAA National Centers for Environmental Information (NCEI)*, 2018. Available: https://www.ncdc.noaa.gov/billions/

[Noaa18b] NOAA National Centers for Environmental Information (NCEI), U.S. Billion-Dollar Weather and Climate Disasters, *NOAA Technical report*, 2018, Available: https://www.ncdc.noaa.gov/billions/

[Noaa19] NOAA and CPHC web page, *Saffir-Simpson hurricane wind scale*, (2019) Available: https://www.nhc.noaa.gov/aboutsshws.php

[Noaa19] NOAA web page, *AL092017 Harvey Report*. (2019) Available: https://www.nhc,noaa.gov/data/to/AL092017_Harvey.pdf

[Nyis19] New York Independent System Operator web page. Open access same-time information system. *NY-ISO web page*, (2019). Available: http://mis.nyiso.com

[Pani17] Panigrahi, B.K., Sahu, S.K., Nandi, R., and Nayak, S., "Probabilistic Load Flow of a Distributed Generation Connected Power System by Two Point Estimate Method", *International Conference on Circuit, Power and Computing Technologies* (ICCPCT), pp. 1-5 (2017)

[Pant17] Panteli, M., Mancarella, P.,Trakas, D.N., Kyriakides, E., and Hatziargyriou, N.D., "Metrics and Quantification of Operational and Infrastructure Resilience in Power Systems", *IEEE Transactions on Power Systems*, Vol. 32, No. 6, pp. 4732-4742 (2017)

[Pasc11] Pasch, R.J., Brown, D.P., and Blake, E.S., "Tropical Cyclone Report: Hurricane Charley", *National Hurricane Center*, pp. 1-23 (2011)

[Patt17] Patterson, J. and Gibson, A., Deep Learning: A Practitioner's Approach, *O'Relly Menia, Inc.*, pp. 1-532 (2017)

[Peng15] Pengcheng, X., Li, Y., Chen, Q., Liu, W., and Liu, Z., "A New Approach for Fast Reliability Evaluation of Composite Power System Considering Wind Farm", *2015 5th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT)*, pp. 2736-2740 (2015)

[Pere08] Pereira, P. and Gama, J., "Pixel Visualization for Mobile Sensor Data Mining, Knowledge Discovery from Sensor Data: Second International Workshop", *Sensor-KDD*, pp. 175-189 (2008)

[Phad08] Phadke, A.G., and Thorp, J.S., "Synchronized Measurement and Their Applications", *Springer*, pp. 1-248 (2008)

[Powe18] Power World Simulation web page. *Power world viewer*, (2018) Available: http://www.powerworld.com

[Quor17] Quora web page, what is the Difference Between Deep Learning and Usual Machine Learning? *QUORA, Inc*, (2017) Available: http:// https://www.quora.com/What-is-the-difference-between-deep-learning-and-usual-machine-learning

[Raus04] Rausand, M., and Hoyland, A., "Chapter 5: Component Importance", *System Reliability Theory: Models and Statistical Methods*, pp. 183-206 (2004)

[Salm15] Salman, A.G., Kanigoro, B. and Heryadi, Y., Weather Forecasting Using Deep Learning Techniques, *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS),* pp. 281-285 (2015)

[Sang19] Sang, Y., Xue, J., Sahraei-Ardakani, M., and Ou, G., An Integrated preventive operation framework for power systems during hurricanes, *IEEE Systems Journal*, pp. 1-8 (2019)

[Shal14] Shalev-Shwartz, S. and Ben-David, S., Understanding Machine Learning: From Theory to Algorithms, *Cambridge University Press*, pp. 184-201 (2014)

[Shap87] Shapiro, S.C., and Eckroth, D., Encyclopedia of Artificial Intelligence- Artificial Intelligence, *Wiley- Interscience Publication*, Vol.1, pp. 9-16 (1987)

[Shen18] Shen, X., Tian, X., Liu, T., Xu, F. and Tao, D., Continuous Dropout, *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-12 (2018)

[Shen98] Shen, B., Koval, D,, Wilsun, X., Salmon, J., and Shen, S., "An Analysis of Extreme-Weather-Related Transmission Line Outages", *Conference Proceedings IEEE Canadian Conference of Electrical and Computer Engineering* (Cat.No.98TH8341), vol. 2, pp. 697-700 (1998)

[Shor04] Short, T.A., "Reliability Indices", *Electric Power Distribution Handbook*, pp. 455-492, (2004)

[Sieg97] Siegelmann, H.T., B.G. Horne, C.L. Giles, Computational capabilities of recurrent NARX neural networks, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol.27, no. 2, pp. 208-215 (1997)

[Smit18] Smith, A., Lott, N., Houston, T., Shein, K., Crouch, J., and Enloe, J., "U.S. Billion-Dollar Weather & Climate Disasters 1980-2018", *NOAA National Centers for Environmental Information (NCEI)*, 2018. Available: https://www.ncdc.noaa.gov/billions/events.pdf

[Tan17] Tan, S., De, D., Song, W.-Z., Yang, J., and Das, S.K., "Survey of Security Advance in Smart Grid: A Data Driven Approach", *IEEE Communications Survey and Tutorials*, Vol. 19, No.1, pp. 397-422 (2017)

[Tech18] Technical Report, "The Definition and Quantification of Resilience", *IEEE PES Industry Technical Support Task Force (PES-TR65),* pp. 1-3 (2018)

[Thom18] Thomas, M.C., Zhu, W. and Romagnoli, J.A., "Data Mining and Clustering in Chemical Process Databases for Monitoring and Knowledge Discovery", *Journal of Process Control,* Vol. 67, pp. 160-175 (2018)

[Tiem18] Tiampo, K.F., McGinnis, S., Kropivnistskaya, Y., Qin, J. and Bauer, M.A., "Chapter 8 – Big Data Challenges and Hazards and Disasters", *Elsevier*, pp. 193-210 (2018)

[Tlei08] Tleis, N.D., "Introduction to Power Systems Faults", *Power Systems Modelling and Fault Analysis: Theory and Practice*, pp. 1-5 (2008)

[Usei18a] U.S. Energy Information Administration, *Electricity Data Browser: Net Generation for All Sector, Monthly,* U.S. Energy Information Administration, 2018. Available: https://www.eia.gov/beta/electricity/data/browser

[Usei18b] U.S. Energy Information Administration web page, *U.S. Energy Mapping Systems*, U.S. Energy Information Administration, 2018. Available: https://www.eia.gov/state/maps.php?v=electricity

[Usgs19] US geological survey web page, *Hurricane H*arvey's *water footprint*, (2019) Available: https://owi.usgs.gov/vizlab/hurricane-harvey

[Vill17] Villmann, T., Biehl, M., Villman, A. and Saralajew, S., Fusion of Deep Learning Architectures, Multilayer Feedforward Networks and Learning Vector Quantizers for Deep Classification Learning, *12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM)*, pp. 1-8 (2017)

[Vill19] Villegas, R., Nava, P., and Vazquez, A., Short-Term Market Price Forecasting Based on De-Noised Wavelet and NARX Neural Network: Data Analytics Approach, *International Conference of Artificial Intelligence (ICAI'19)*, pp. 239-245 (2019)

[Voro17] Vorobev, P., Huang, P., Hosani, M.A., Kirtley, J.L., and Turistyn, K., "A Framework for Development of Universal Rules for Microgrids Stability and Control", *IEEE 56th Annual Conference on Decision and Control* (CDC), pp. 5125-5130 (2017)

[Wang16b] Wang, X.Z., Zhou, J., Huang, X.L.Bi., Ge, Z.Q.Ge and Li, L., A multilevel Deep Learning Method for Big Data Analysis and Emergency Management of Power System*, IEEE International Conference on Big Data Analysis (ICBDA),* pp. 1-5 (2016)

[Wang17] Wang, Ch., Hou, Y., Qiu, F., Lei, Sh., and Liu, K., Resilience enhancement with sequential proactive operation strategies, *IEEE Transactions on Power Systems*, vol. 32, No. 4, pp. 2847-2857 (2017)

[Weat19] National Weather Service web page. *Average monthly and annual temperatures at Central Park*, (2019) Available: http://www.weather.gov/media/okx/Climate/Central-Park/monthlyannualtemp.pdf

[Weat19] Weather Underground web page. Jhon F. Kennedy International Airport, NY: Temperature history daily observation. *Weather Underground-The Weather Company*, (2019). Available: http://www.wunderground.com

[Weis18] Weisstein, E.W., Hyperbolic Tangent, *MathWorld – A Wolfram Web Resource,* (2018}, Available: http://mathworld.wolfram.com/HyperbolicTangent.html (2018)

[Wiki19] Wikipedia web page, *Saffir-Simpson scale*, (2019) Available: https://en.m.wikipedia.org/wiki/Saffir-Simpson_scale

[Xiao17] Xiao, F., Wang, S. and Fan, C., "Mining Bug Building Operational Data for Building Cooling Load Prediction and Energy Efficiency Improvement", *IEEE International Conference on Smart Computing* (SMARTCOMP), pp. 1-3 (2017)

[Yang17] Yang, C., Su, G. and Chen, J., "Using Big Data to Enhance Crisis Response and Disaster Resilience for a Smart City". *IEEE 2nd International Conference on Big Data Analysis* (ICBDA), pp. 504-507 (2017)

[Yang17b] Yang, D., Xie, H., Yin, J., Liu, Y. and Yan, Ch., Supervised Deep Quantization for Efficient Image Search, *IEEE International Conference on Multimedia & Expo Workshops (ICMEW),* pp. 525-530 (2017)

[Yu10] Yu, D., Wang, S. and Li, D., Sequential Labeling Using Deep-Structured Conditional Random Fields, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 4, No. 6, pp. 965-973 (2010)

[Yu16] Yu, C., "Research on Emergency Logistics Decision Support System Design Under Data Mining & WebGIS Technology", *International Conference on Intelligent Transportation, Big Data & Smart City* (ICITBS), pp. 375-378 (2016)

[Yu17] Yu, F., Booth C.D., and Dyśko, A., "Voltage-based Fault Identification for PMU-based Wide Area Backup Protection Scheme", *IEEE Power & Energy Society General Meeting*, pp. 1-5 (2017)

[Yufb17] Yu, F., Booth, C.D., and Dyśko, A., "Voltage-based Fault Identification for PMU-based Wide Area Backup Protection Scheme", *IEEE Power & Energy Society General Meeting*, pp. 1-5 (2017)

[Zhao07] Zhao, J.H., Dong, Z.Y. and Wong, K.P., "A Framework for Electricity Price Spike Analysis with Advanced Data Mining Methods", *IEEE Transactions on Power Systems,* vol. 22, no.1, pp. 376-385 (2007)

APPENDIX A

Data Collection Scenario-1

The following section contains the data collection from Scenario-1.

A.1 DATA COLLECTION FROM SCENARIO-1 ANALYSIS-1 AND 2.

This section presents only a portion of the data collection from Scenario-1 for demonstration purposes. The data sets collection from Scenario-1 is based in information from the New York Independent Service Provider (NY-ISO) web site [NYIS19]. Raw data portion of the electricity price and demand consumption from New York for each hour is presented in a .zip file format, needed to be uncompressed and transfer in an excel sheet file. Figure A.1.1 shows a flow process for the data preparation showing a portion of zip files of price data from NY-ISO web page for case-1. Similarly, Figure A.1.2 shows flow process for the data preparation for case-1 but showing portion of temperature data from New York area. The data was gathered from Network for Environment and Weather Applications web site [NEWA19], and compared with data from Weather Underground web page [Weat19]. Figures A.1.3 and Figures A.1.4 shows the data processing for case-2 which was used the same data arrangement methodology.
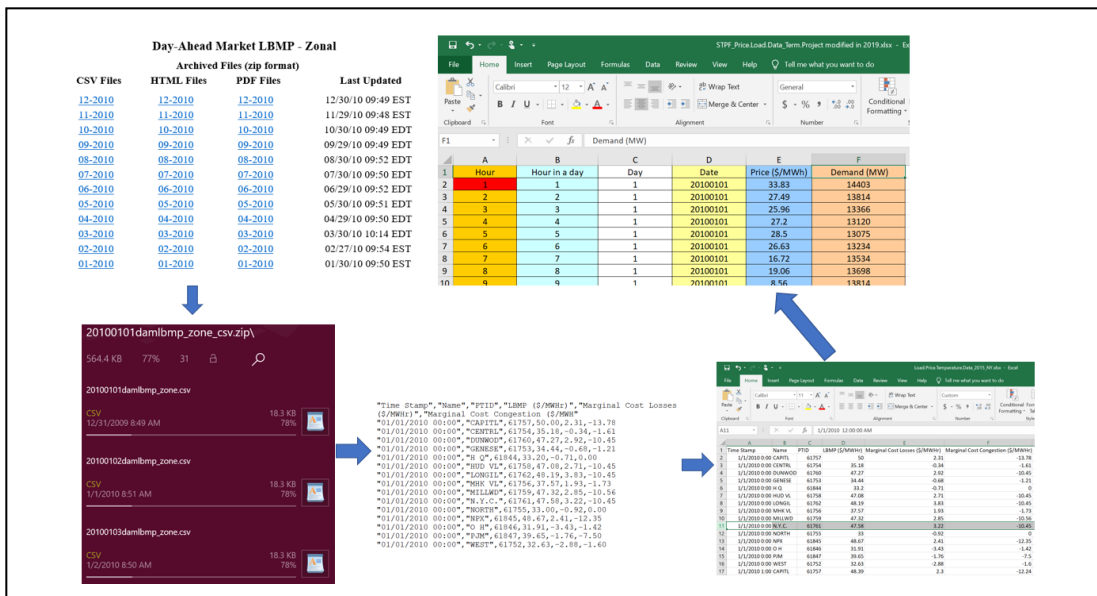
Figure

A.1.1: Flow data preparation showing portion of zip files used for Scenario-1, case-1 for price forecasting.
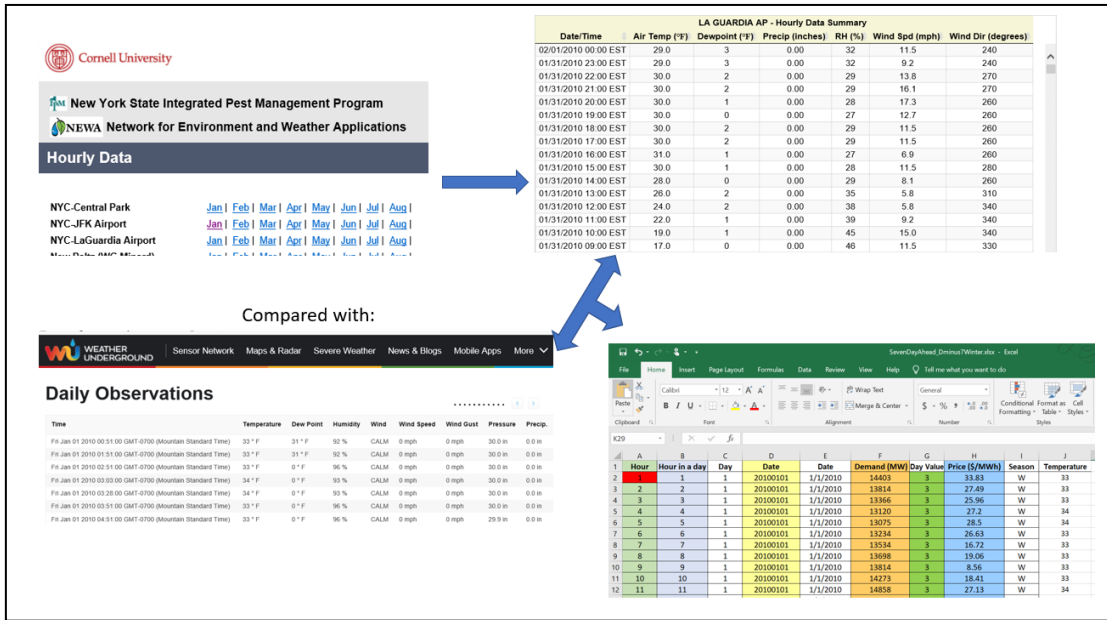
Figure A.1.2: Flow data preparation showing portion of temperatures files used for Scenario-1, case-1 for price forecasting.
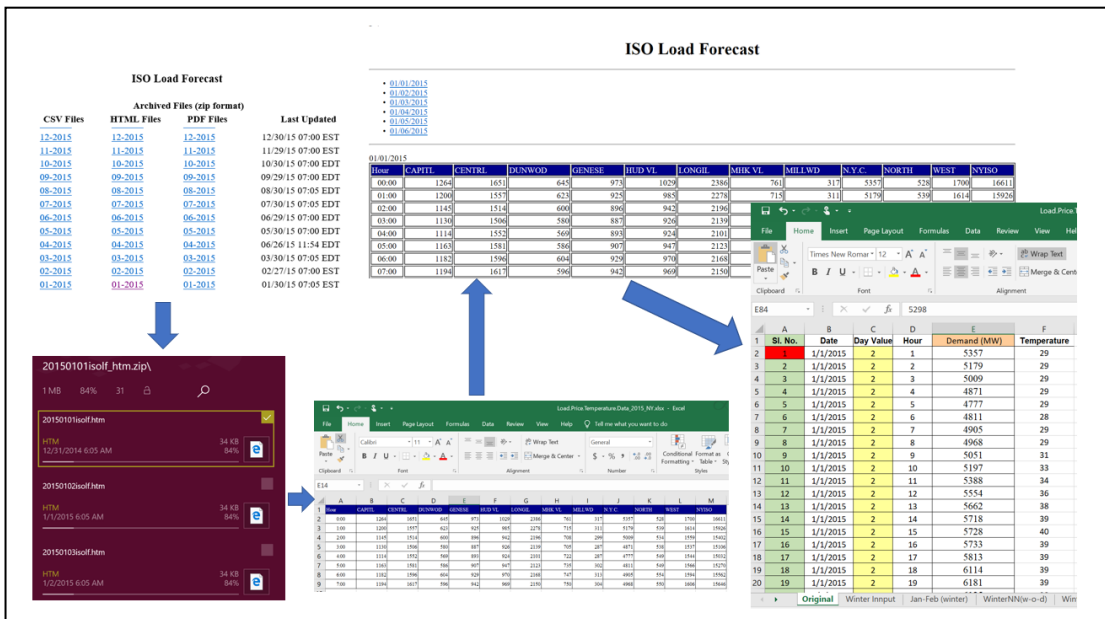
Figure A.1.3. Flow data preparation showing portion of zip files used for Scenario-1, case-2 for demand forecasting.

137
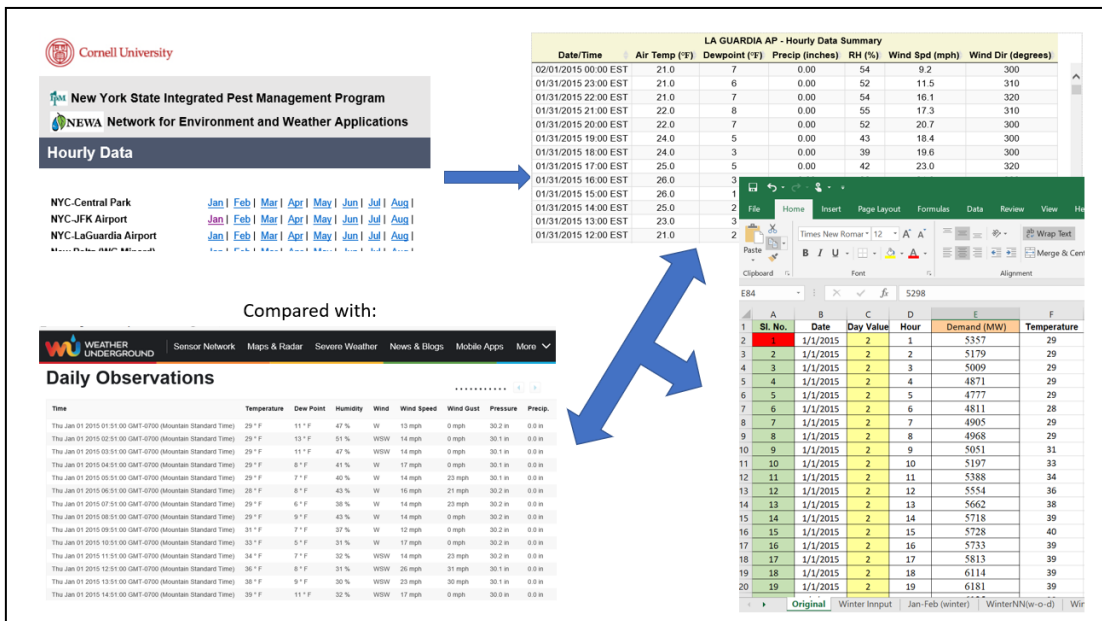
Figure A.1.4. Flow data preparation showing portion of temperatures files used for Scenario-1, case-2 for demand forecasting.

A.1.1 Analysis-2 Data collection

Figure A.1.5 shows part of the data processing used for modeling including part of the modeling showing the libraries used in analysis-2.
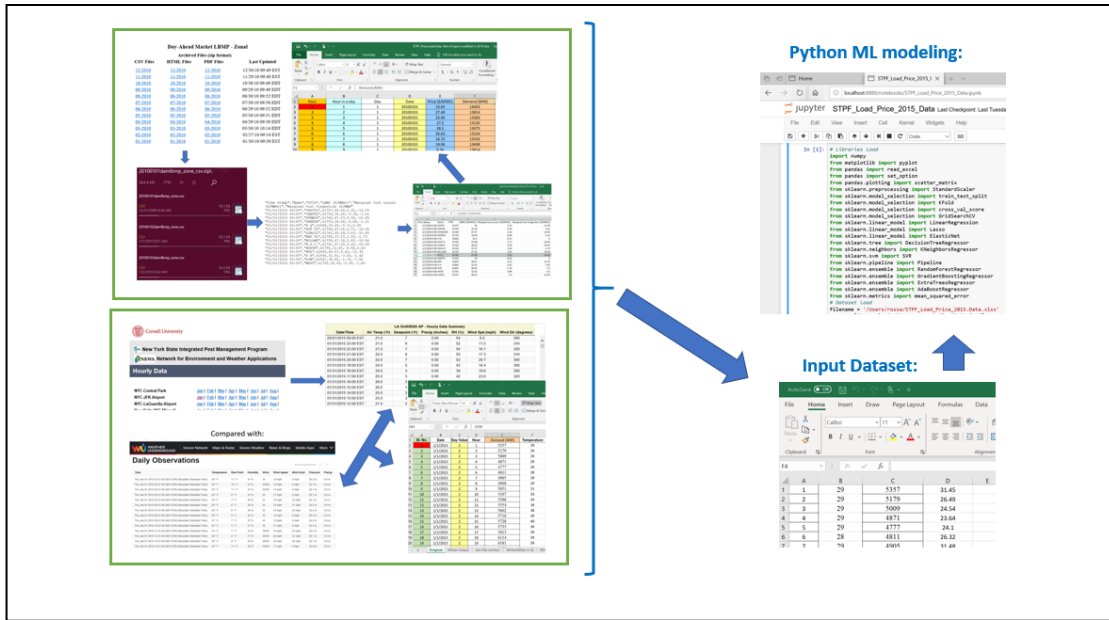
Figure A.1.5. Flow of the data showing portion of files used for modeling Scenario-1
Analysis-2.

A.1.2 Machine Learning modeling process.

Figure A.1.6 shows the steps followed in this dissertation for machine learning modeling.
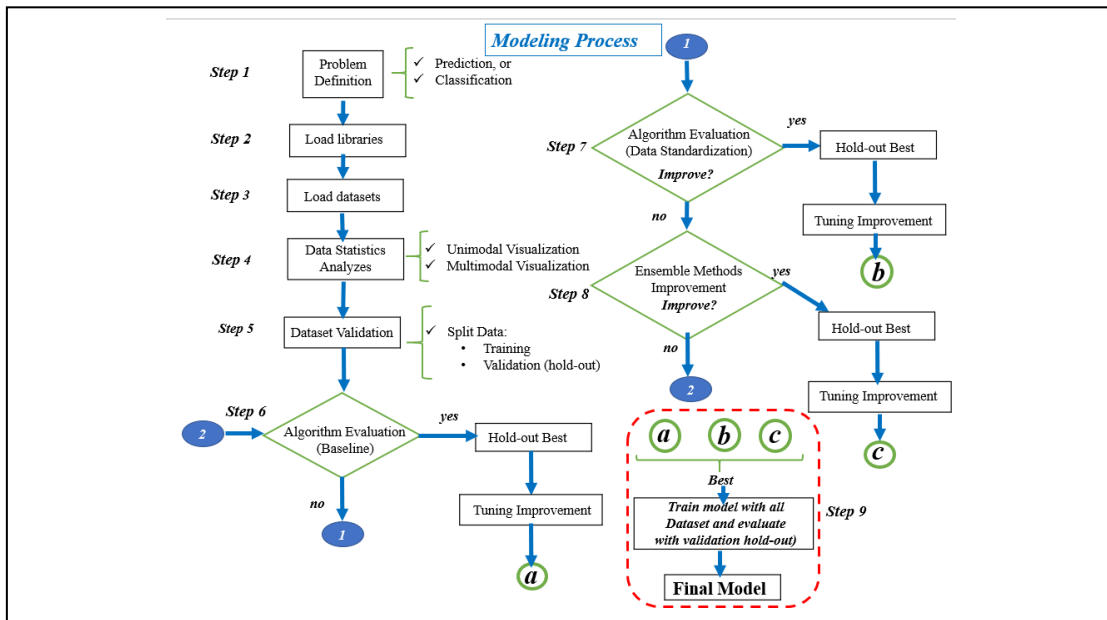
Figure A.1.6. Machine Learning modeling process.

A.2 Hypothetical data collection

The following hypothetical data to represent the ETDS system used in scenario-2 of this dissertation.

The below nomenclature corresponds for both RTS bus systems (IEEE 9-bus and IEEE 14-bus) proposed by [Deme17] showed in Figures A.2.1 (Machine Data), A.2.2 (Exciter Data), A.2.3 (Governor Data) for IEEE 9-Bus, and A.2.4 (Machine Data), A.2.5 (Exciter Data), and A.2.6 (Governor Data) for IEEE 14-Bus, respectively.

Nomenclature:

Rated MVA	Machine-rated MVA; base MVA for impedances
Rated kV	Machine-rated terminal voltage in kV; base kV for impedances
H	Inertia constant in s
r_a	Armature resistance in p.u.
r_d	Unsaturated d axis synchronous reactance in p.u.
r_q	Unsaturated q axis synchronous reactance in p.u.
x'_d	Unsaturated d axis transient reactance in p.u.
x'_q	Unsaturated q axis transient reactance in p.u.
x''_d	Unsaturated d axis subtransient reactance in p.u.
x''_q	Unsaturated q axis subtransient reactance in p.u.
$x_1 \; or \; x_p$	Leakage or Potier reactance in p.u.
T'_{d0}	d axis transient open circuit time constant in s.
T'_{q0}	q axis transient open circuit time constant in s.
T''_{d0}	d axis subtransient open circuit time constant in s.
T''_{q0}	q axis subtransient open circuit time constant in s.
S (1.0)	Machine saturation at 1.0 p.u. voltage in p.u.
S (1.2)	Machine saturation at 1.2 p.u. voltage in p.u.
T_r	Regulator input filter time constant in s.

K_a	Regulator gain (continuous acting regulator) in p.u.
T_a	Regulator time constant in s.
V_{Rmax}	Maximum regulator output, starting at full load field voltage in p.u.
V_{Rmin}	Minimum regulator output, starting at full load field voltage in p.u.
K_e	Exciter self-excitation at full load field voltage in p.u.
T_e	Exciter time constant in s.
K_f	Regulator stabilizing circuit gain in p.u.
T_f	Regulator stabilizing circuit time constant in s.
E_1	Field voltage value, 1 in p.u.
$SE(E_1)$	Saturation factor at E_1.
P_{max}	Maximum turbine output in p.u.
R	Turbine steady-state regulation setting o droop in p.u.
T_1	Control time constant (governor delay) in s.
T_2	Hydro reset time constant in s.
T_3	Servo time constant in s.
T_4	Steam valve bowl time constant in s.
T_5	Steam reheat time constant in s.
F	Shaft output ahead of reheater in p.u.

A.2.1 IEEE 9-bus system data

The numbers shown in the Figure A.5, A.6, and A.7 for the bus numbers corresponds to the default test system and the modified system (in parenthesis), respectively, proposed by the author [Deme17].

IEEE 9-Bus Modified Test System Machine Data			
Type	GENROU	GENROU	GENROU
Operation	Sync. Gen.	Sync. Gen.	Sync. Gen.
Default Unit no. (New Unit no.)	1(12)	2(10)	3(11)
Rated power (MVA)	512	270	125
Rated voltage (kV)	24	18	15.5
Rated pf	0.9	0.85	0.85
H (s)	2.6312	4.1296	4.768
D	2.000	2.000	2
r_a (p.u)	0.004	0.0016	0.004
x_d (p.u)	1.700	1.700	1.220
x_q (p.u)	1.650	1.620	1.160
x'_d (p.u)	0.270	0.256	0.174
x'_q (p.u)	0.470	0.245	0.250
x''_d (p.u)	0.200	0.185	0.134
x''_q (p.u)	0.200	0.185	0.134
x_l or x_p (p.u)	0.160	0.155	0.0078
T'_{d0} (s)	3.800	4.800	8.970
T'_{q0} (s)	0.480	0.500	0.500
T''_{d0} (s)	0.010	0.010	0.033
T''_{q0} (s)	0.0007	0.0007	0.070
$S(1.0)$	0.090	0.125	0.1026
$S(1.2)$	0.400	0.450	0.432

A.2.1: IEEE 9-Bus modified test system machine data, figure taken from [Deme17].

IEEE 9-Bus Modified Test System Exciter Data			
Type	IEEET1	IEEET1	IEEET1
Default Unit no. (New Unit no.)	1(12)	2(10)	3(11)
Rated power (MVA)	512	270	125
Rated voltage (kV)	24	18	15.5
T_r (s)	0.000	0.000	0.060
K_a (p.u)	200	30	25
T_a (s)	0.395	0.400	0.200
V_{Rmax} (p.u)	3.840	4.590	1.000
V_{Rmin} (p.u)	-3.840	-4.590	-1.000
K_e (p.u)	1.000	-0.020	-0.0601
T_e (s)	0.000	0.560	0.6758
K_f (p.u)	0.0635	0.050	0.108
T_f (s)	1.000	1.300	0.350
E_1 (p.u)	2.880	2.5875	2.4975
$SE(E_1)$	0.000	0.7298	0.0949
E_2 (p.u)	3.840	3.450	3.330
$SE(E_2)$	0.000	1.3496	0.37026

Figure A.2.2: IEEE 9-Bus modified test system exciter data, figure taken from [Deme17].

IEEE 9-Bus Modified Test System Governor Data			
Type	BPA_GG	BPA_GG	BPA_GG
Default Unit no. (New Unit no.)	1(12)	2(10)	3(11)
Rated power (MVA)	512	270	125
Rated voltage (kV)	24	18	15.5
P_{max} (p.u)	0.8984	0.8518	1.056
R (p.u)	0.00976	0.01852	0.040
T_1 (s)	0.150	0.100	0.083
T_2 (s)	0.050	0.000	0.000
T_3 (s)	0.300	0.259	0.200
T_4 (s)	0.260	0.100	0.050
T_5 (s)	8.000	10.000	5.000
F	0.270	0.272	0.280

Figure A.2.3 IEEE 9-Bus modified test system governor data, figure taken from [Deme17].

A.2.2 IEEE 14-bus system data

The numbers shown in the Figures A.8, A.9, and A.10 for the bus numbers corresponds to the default test system and the modified system (in parenthesis), respectively [Deme17].

IEEE 14-Bus Modified Test System Machine Data				
Type	GENROU	GENROU	GENROU	GENROU
Operation	Sync. Gen.	Sync. Gen.	Condenser	Condenser
Default Unit no. (New Unit no.)	1(15)	2(16)	3(17)	6(19), 8(18)
Rated power (MVA)	448	100	40	25
Rated voltage (kV)	22	13.8	13.8	13.8
Rated pf	0.85	0.8	0.0	0.0
H (s)	2.656	4.985	1.520	1.200
D	2.000	2.000	0.000	0.000
r_a (p.u)	0.0043	0.0035	0.000	0.0025
x_d (p.u)	1.670	1.180	2.373	1.769
x_q (p.u)	1.600	1.050	1.172	0.855
x'_d (p.u)	0.265	0.220	0.343	0.304
x'_q (p.u)	0.460	0.380	1.172	0.5795
x''_d (p.u)	0.205	0.145	0.231	0.2035
x''_q (p.u)	0.205	0.145	0.231	0.2035
x_l or x_p (p.u)	0.150	0.075	0.132	0.1045
T'_{do} (s)	0.5871	1.100	11.600	8.000
T'_{qo} (s)	0.1351	0.1086	0.159	0.008
T''_{do} (s)	0.0248	0.0277	0.058	0.0525
T''_{qo} (s)	0.0267	0.0351	0.201	0.0151
$S(1.0)$	0.091	0.0933	0.295	0.304
$S(1.2)$	0.400	0.4044	0.776	0.666

Figure A.2.4: IEEE 14-Bus modified test system machine data, figure taken from [Deme17].

IEEE 14-Bus Modified Test System Exciter Data				
Type	IEEET1	IEEET1	IEEET1	IEEET1
Default Unit no. (New Unit no.)	1(15)	2(16)	3(17)	6(19), 8(18)
Rated power (MVA)	448	100	40	25
Rated voltage (kV)	22	13.8	13.8	13.8
T_r (s)	0.000	0.060	0.000	0.000
K_a (p.u)	50	25	400	400
T_a (s)	0.060	0.200	0.050	0.050
V_{Rmax} (p.u)	1.000	1.000	6.630	4.407
V_{Rmin} (p.u)	-1.000	-1.000	-6.630	-4.407
K_e (p.u)	-0.0465	-0.0582	-0.170	-0.170
T_e (s)	0.520	0.6544	0.950	0.950
K_f (p.u)	0.0832	0.105	0.040	0.040
T_f (s)	1.000	0.350	1.000	1.000
E_1 (p.u)	3.240	2.5785	6.375	4.2375
$SE(E_1)$	0.072	0.0889	0.2174	0.2174
E_2 (p.u)	4.320	3.438	8.500	5.650
$SE(E_2)$	0.2821	0.3468	0.9388	0.9386

Figure A.2.5: IEEE 14-Bus modified test system exciter data, figure taken from [Deme17].

IEEE 14-Bus Modified Test System Governor Data		
Type	BPA_GG	BPA_GG
Default Unit no. (New Unit no.)	1(15)	2(16)
Rated power (MVA)	448	100
Rated voltage (kV)	22	13.8
P_{max} (p.u)	0.870	1.050
R (p.u)	0.011	0.050
T_1 (s)	0.100	0.090
T_2 (s)	0.000	0.000
T_3 (s)	0.300	0.200
T_4 (s)	0.050	0.300
T_5 (s)	10.000	0.000
F	0.250	1.000

Figure A.2.6: IEEE 14-Bus modified test system governor data, figure taken from [Deme17].

A.2.3 TEXAS 2000-bus system data

Figures A.2.7, A.2.8, and A.2.9 correspond to the Texas 2000-bus used in this study where the information was taken from Powerworld viewer simulation ver. 20 [Powe18], suggested from the work of [Birc17]. Figure A.11 shows Texas 2000-Bus information, and Figure A.12 and A.13

shows the bus power flows for two different ERCOT weather zones (Central and South), in which, only the affected sampled areas of this study was taken apart (Wind damage area and Flood damage area).
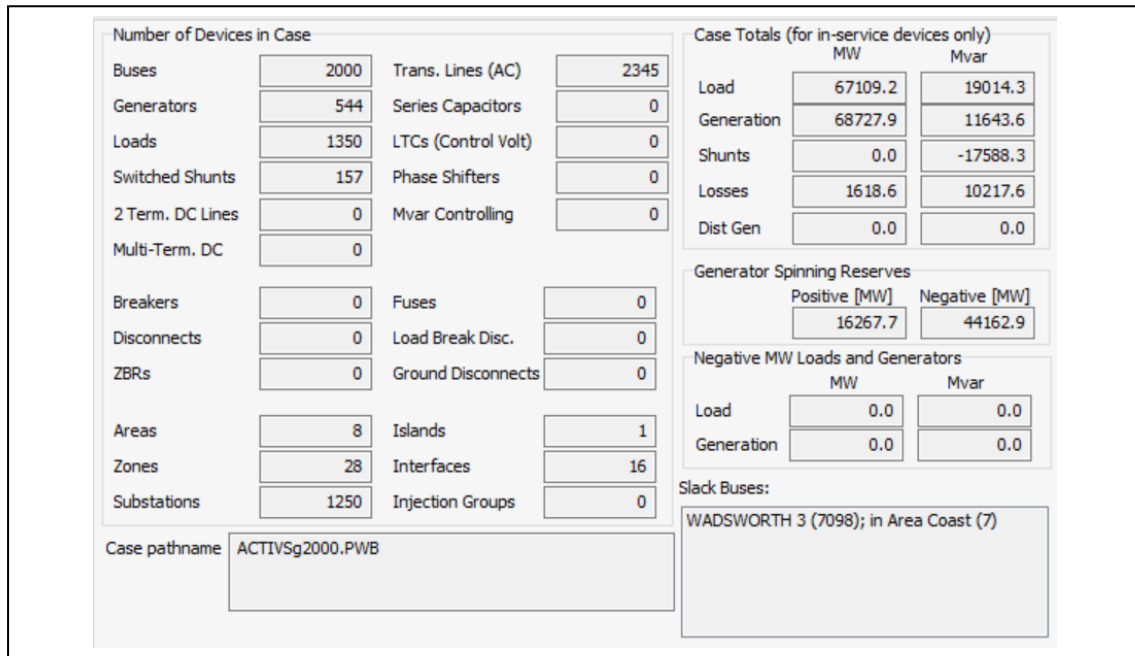

Figure A.2.8: Texas 2000-bus information taken from Powerworld viewer simulation ver. 20 [Powe18], from the work proposed by [Birc17].

Figure A.2.9: Section from bus flow data from Texas 2000-bus showing the study area (Hurricane Harvey Wind damage), information taken from Powerworld viewer simulation ver. 20 [Powe18], from the work proposed by [Birc17].

```
BUS    7038 SUGAR LAND 3 1 230.0   MW    Mvar     MVA   % GIC Amps 1.0074 -32.79   7 Coast
  TO   7037 SUGAR LAND 3 0 1  -277.28 -123.94   303.7  29    0.0N  1.0000TA   0.0
  TO   7039 SUGAR LAND 3 2 1   300.40   35.38   302.5  72    0.0N  1.0000NT   0.0
  TO   7039 SUGAR LAND 3 2 2   300.40   35.38   302.5  72    0.0N  1.0000NT   0.0
  TO   7126 SUGAR LAND 2 1 1  -180.92   -3.23   181.0  24    0.0
  TO   7126 SUGAR LAND 2 1 2  -180.92   -3.23   181.0  24    0.0
  TO   7184 MISSOURI CITY 2 0 1  -309.29    1.66   309.3  48     0.0
  TO   7409 KATY 2 0     1   173.81   28.99   176.2  23    0.0
  TO   7409 KATY 2 0     2   173.81   28.99   176.2  23    0.0

BUS    7039 SUGAR LAND 3 2 115.0   MW    Mvar     MVA   % GIC Amps 0.9987 -37.26   7 Coast
  LOAD 1                  235.50   66.73   244.8 DistGen   0.00    0.00     0.0
  TO   7038 SUGAR LAND 3 1 1  -299.64  -11.78   299.9  71    0.0N  1.0000TA   0.0
  TO   7038 SUGAR LAND 3 1 2  -299.64  -11.78   299.9  71    0.0N  1.0000TA   0.0
  TO   7101 HOUSTON 88 0 1   103.60   -1.22   103.6  47    0.0
  TO   7101 HOUSTON 88 0 2   103.60   -1.22   103.6  47    0.0
  TO   7101 HOUSTON 88 0 3   103.60   -1.22   103.6  47    0.0
  TO   7397 ROSENBERG 0  1    52.96  -39.51    66.1  28    0.0
```

Figure A.2.10: Section from bus flow data from Texas 2000-bus showing the study area (Hurricane Harvey Flood damage), information taken from Powerworld viewer simulation ver. 20 [Powe18], from the work proposed by [Birc17].

APPENDIX B

Hurricane Information

B.1 Saffir-Simpson Hurricane wind scale.

The Saffir-Simpson hurricane scale (SSHS) was created in 1971 by civil engineer Herbert Saffir (civil engineer) and Robert Simpson (director in that time, U.S. National Hurricane Center) [Wiki19]. The SSHS scale is from 1 to 5 ratting based on a hurricane's sustained wind speed [Noaa19], with a related two more classifications [Wiki19]. The SSHS estimates potential damage, Table B.1 sown this scale and related scale which all winds are using the U.S. 1-minute average sustained wind.

Table B.1: Saffir-Simpson hurricane wind scale (SSHWS) [Noaa19], and [Wiki19].

Category	Winds (mph)	Winds (kt)	Winds (kn/hr)	Potential damage
Tropical storm	0-38	0-33	0-34	Initiated a string of tornadoes that damage homes, displaced trees, and overturned vehicles as it moves.
Tropical depression	39-73	34-64	35-63	Heavy rains and strong winds cause minor flooding and property damage
1	74-95	64-82	119-153	Very dangerous wind will produce some damage
2	96-110	83-95	154-177	Extremely dangerous wins will cause extensive damage.
3	111-129	96-112	178-208	Devastating damage will occur.
4	130-156	113-136	209-251	Catastrophic damage will occur
5	157 or higher	137 or higher	252 or higher	Catastrophic damage will occur

B.2 Hurricane Harvey report (AL092017)

In this section a reported from NOAA and the National Weather Service is briefly presented, including figures of the reported rainfall, damages areas, etc.

B.2.1 Reported wind and flood damage.

"… Harvey's maximum winds of 115 kt occurred during a several hour periods concluding with its first Texas landfall. That intensity was based on a blend of peak SFMR measurements of 113 kt near 2122 UTC 25 August and maximum observed 700-mb flight-level winds of 129 kt at 2037 UTC and 2330 UTC 25 August. Both of those 700-mb winds support a surface wind of about 115 kt using a typical flight-level wind to surface wind reduction. Another SFMR measurement of 113 kt at 0419 UTC 26 August is thought to be unreliable due to shoaling. The highest *observed* sustained winds on land were 96 kt near Aransas Pass, with the highest *observed* gust being 126 kt near Rockport, Texas…", "… The highest storm total rainfall report from Harvey was 60.58 inches near Nederland, Texas…", "…36 to 48 inches recorded in the Houston metro area…" [Blak18].

B.2.2 Synoptic history of Hurricane Harvey Texas (17 Aug. 1 Sep. 2017).

"The wave that spawned Harvey moved off the west coast of Africa on 12 August with a large convective mass that had mostly dissipated by late the next day…", "... Harvey moved quickly westward, south of a western Atlantic ridge, reaching an initial peak intensity of 40 kt early on 18 August…", "... The remnants of Harvey moved rapidly to the west and west-northwest for the next couple of days, staying convectively active while they moved over the Yucatan Peninsula on 22 August…", "...Harvey began to rapidly intensify late on 23 August in an environment of light shear, very warm water and high mid-level moisture…", "... Harvey became a hurricane later on 24 August, and by that night a well-defined eye appeared in infrared satellite pictures. The hurricane reached category 3 status by midday on 25 August while it approached the middle Texas coast and intensified into a category 4 hurricane by 0000 UTC 26 August. Harvey's center made landfall on the northern end of San Jose Island about 5 n mi east of Rockport, Texas at 0300 UTC

that day. Sustained winds of 115 kt and a minimum central pressure of 937 mb are estimated for that landfall. The hurricane then made a second landfall on the Texas mainland 3 h later, slightly weaker due to land interaction, with 105 kt winds and an estimated central pressure of 948 mb southeast of Refugio on the northeast coast of Copano Bay west of Holiday Beach. Harvey rapidly weakened over land to a tropical storm within 12 h after landfall and maintained a 35-kt intensity for the next day or so, aided by the sustaining effects of the southeastern portion of its circulation remaining over water…", "… The storm center moved back offshore around 0300 UTC 28 August over Matagorda Bay, its winds slightly re-strengthening with deep convection reforming near and north of the center. However, the vertical wind shear was too strong for much intensification, and Harvey reached a final peak intensity of 45 kt late on 29 August…", "… Extremely heavy rains, however, continued on the north and northwest side of the tropical cyclone, most concentrated then near the Beaumont-Port Arthur area. Harvey made its final landfall in southwestern Louisiana at 0800 UTC 30 August near Cameron with 40-kt sustained winds. Thereafter, the cyclone slowly weakened over land, becoming a tropical depression late on 30 August. Harvey then moved north-eastward over the southern United States while producing heavy rainfall, and it transformed into an extratropical cyclone by 0600 UTC 1 September over the Tennessee Valley. The cyclone dissipated over northern Kentucky late the next day…" [Blak18].
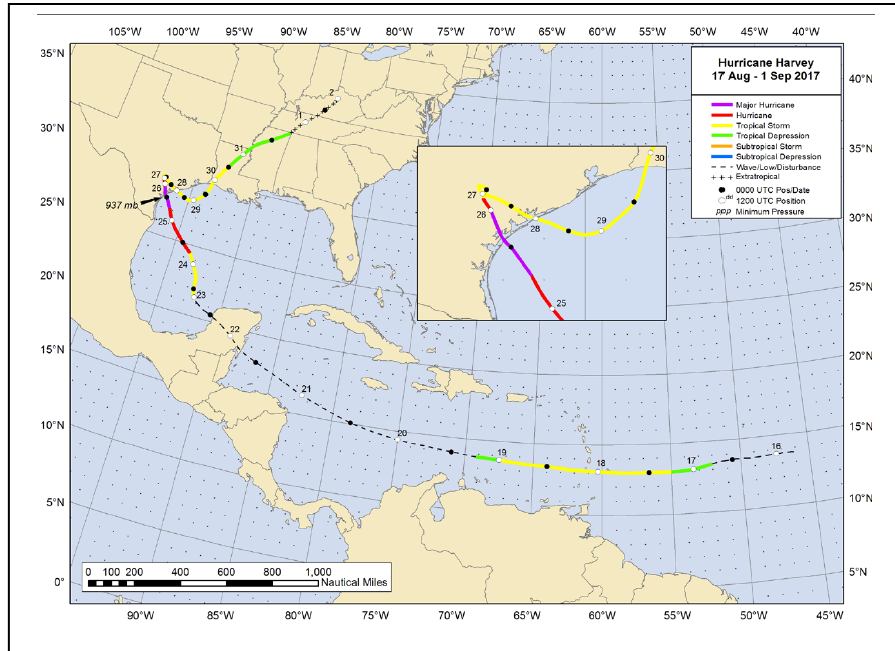
Figure B.2.1: Best track positions for Hurricane Harvey during 17 August – 1 September, image
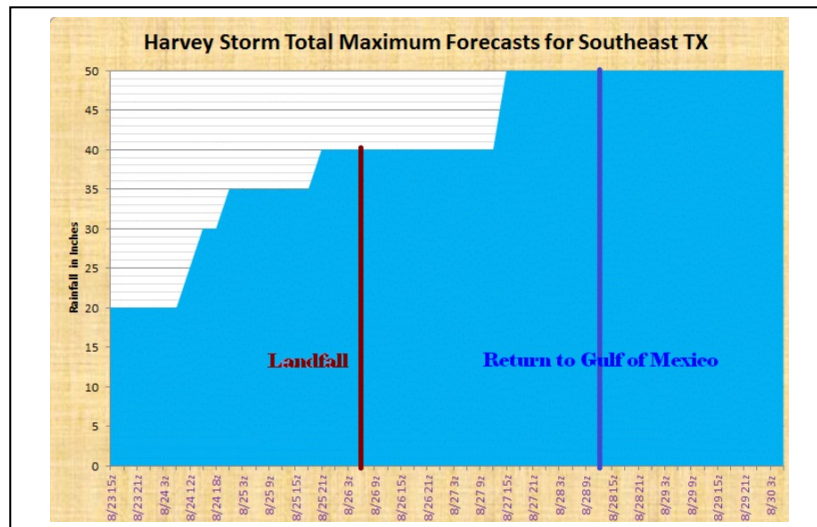taken from [Blac18].

Figure B.2.2: Best track positions for Hurricane Harvey during 17 August – 1 September,
image taken from [Blac18].

Figure B.2.5: Wind damage during Hurricane Harvey right after the landfall at south of

Texas, image taken from [Aept18].

B.2.5: Water rescues that were ongoing during Harvey in Houston on 27 August 2017,

image taken from [Blac18].

APPENDIX C

Data Collection Scenario-2

The following section contains the data collection for Scenario-2. This section gives a glimpse of the data collection from Scenario-2 for demonstration purposes. The dataset collections were based on information discussed in chapter 7. Each dataset was used on several formats, some of them includes images under the formats of .jpg, TIFF, etc., to achieve better behavior on the simulations during this study. Similarly, other data was used as a data arrays for better model simulations i.e., xlsx and .csv as an input data in ML or DL. The following are the main sources of the collection of data from Scenario-2: NOAA, NASA, ERCOT, Powerworld simulator, HAZUS simulator, Open Infrastructure engine, Open Maps engine, Mapillary engine, ArcGIS simulator, ENVI simulator. Many other data were acquired from web sources that was exhaustively searched, and applied all the precaution needed to be added to the data collection of this study.

C.1 DATA COLLECTION SCENARIO-2, EXTREME-WEATHER MODEL

Figure C.1.1 shows the data flow of HAZUS wind damage modeling. A portion of the data was shown for simplification purposes.
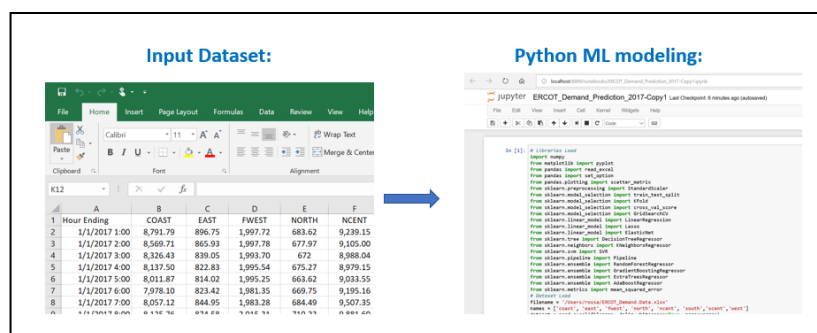

Figure C.1.1: Data flow of the HAZUS wind demand modeling.

C.2 DATA COLLECTION SCENARIO-2, COMPONENT MODEL

Figure C.2,1 shows the data flow to classify the peak flooding using a Deep Learning neural network. A portion of the data was shown for simplification purposes.
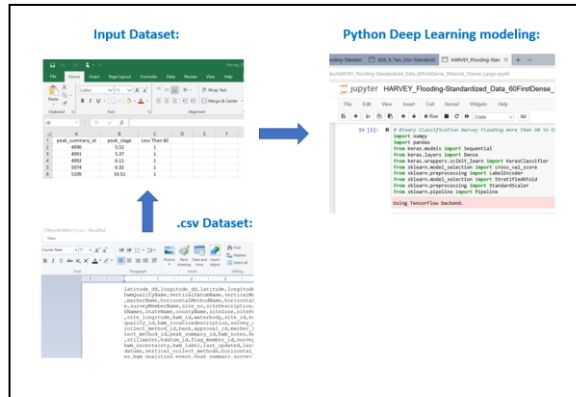
Figure C.2.1: Data flow of Harvey peak flooding deep learning modeling for Scenario-2.

The data flow of the modeling ERCOT demand using Machine Learning algorithms is shown in the Figure C.2.2. A portion of the data was shown for simplification purposes.

Figure C.2.2: Data flow of the ERCOT demand modeling for the component model in Scenario-2.

C.3 DATA COLLECTION SCENARIO-2, SYSTEM MODEL

Figure C.3.1 shows the data flow of the two-disconnected lines for DC OPF modeling, only a portion of the data were shown, for simplification purposes.

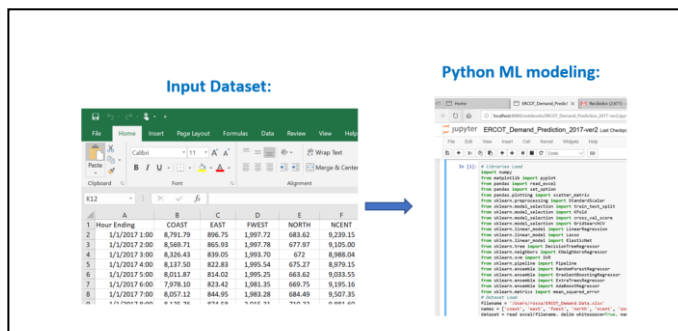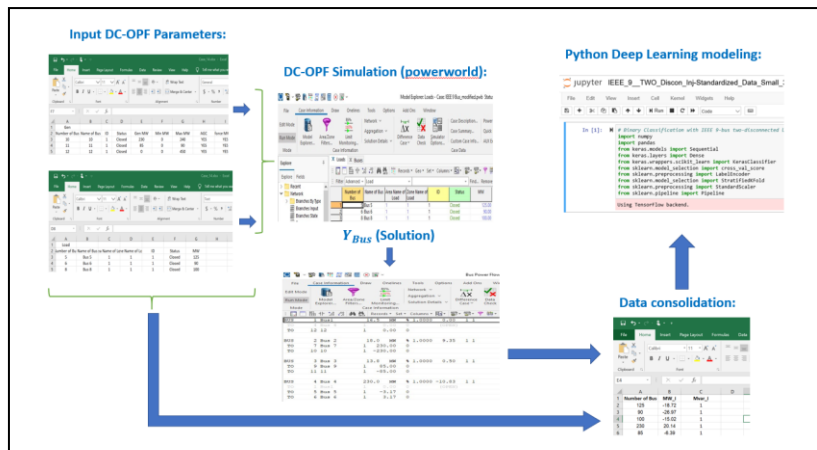Figure C.3.1: Data flow of the two-disconnected lines for DC OPF modeling for the system model in Scenario-2.

APPENDIX D

Simulation setting Scenario-1 and 2.

In order to simulate all the case studies from chapter 6 and 7. This study used a recent efficient developed hardware since the last decade, the GPU (Graphical Processing Units), also, at the time of this study, the latest processor generation for commercial laptop, the Intel Core i7 with 2 GHz as an experimental workstation. Similarly, was used the most novel Machine Learning and Deep Learning software tools like: *Python*, *TensorFlow*. Additionally, *Anaconda ecosystem* was used in all the settings for this study, there are many other options out there but we choose this ecosystem, because has strong attraction between Machine Learning and Deep Learning communities, at the time of this study with 6 million users as reported in Anaconda web site [Anac19]. Anaconda is an open free source which distribute the latest Python packages released, holding with the most of the tools needed to setting up the experimentation frameworks for the purpose of this study. Figure D.1 shows the test bed (computers) used in this study.
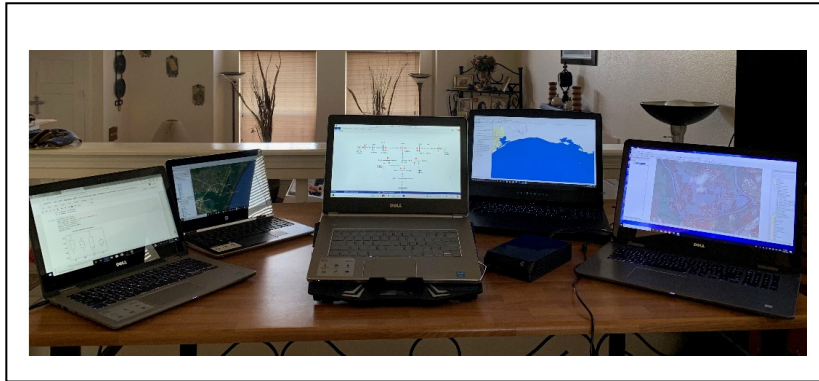

Figure D.1: Test bed (computers) used in this study.

D.1 TERMINOLOGY

The following terms was used in the both data simulations, machine learning (ML) and deep learning:

Terms	Purpose of meaning in the context of data modeling
Instance.	Single row of data in the dataset.
Feature, attribute, field, or variable.	Single column of data being referenced by the learning algorithms, which some could be inputs or outputs of the learning algorithm.
Feature vector or tuple.	List of features.
Dimension.	Subset of attributes that describe data property i.e., Date dimension = 3 attributes (day, month, and year).
Dataset .	Collection of data prepared for ML or DL modeling proposed.

D.2 SETTING UP MACHINE LEARNING FRAMEWORK

Machine Learning framework was setting up on a Dell CPU Intel core i7. Below we briefly detailed the dependencies needed in this study in order to use some algorithms between the packages in machine learning. The Python dependencies can be installed thru *pip* or *conda install*. While the progress of the ML and DL package, and the ML and DL evolving application areas, in a so fast-paced that any current references in the state of the art and current version, is updated at any moment, thus, only as a main reference for this study the language version is presented. Furthermore, the steps to setting up the experiment platforms are briefly detailed below.

Two workstations were employed with different Python versions to evaluate the models in sections 8: 1) version 2.7 and 2) version 3.6:

D.2.1 Dependencies ML

Language Platform: Python 2.7 and 3.6

Python Scikit-learn. Is a complete package for machine learning tasks in a Python platform requiring other dependencies like *Python Numpy* and *Python Scipy*. (http://scikit-learng.org).

Python SciPy. Is a collection of functions for scientific computing in Python. It provides function to handle N- dimensional arrays among other functionality like: advanced linear algebra routines, mathematical function routines, and statistical distributions.

Python NumPy. Is one of the essential packages for scientific computing in Python, containing multidimensional arrays, high-level mathematical functions such as linear algebra operation and other functions.

Python Pandas. Is a package for data wrangling and analysis. Focusing on reading, writing, and manipulating data, this dependency as Scikit-learn need the installation prior to install Pandas.

Python Matplotlib. Is one of the most important plotting packages for Python, used to plot the accuracy of the model by creating 2D charts and plots from data.

D.2.2 ML experimental libraries Scenario-1 Analysis-2

The following machine learning regression libraries was used in Scenario-1 and Analysis-2 of this dissertation.

- *LinerRegression*
- *Lassso*
- *ElasticNet*
- *DescisionTreeRegressor*
- *KneighborsRegressor*
- *SVR*
- Ensemble *RandomForestRegressor*
- Ensemble *GradientBoostingRegressor*
- Ensemble *ExtraTreesRegressor*
- Ensemble *AdaBoostRegressor*

D.2.3 ML experimental libraries Scenario-2, Extreme-weather model.

The following Supervised machine learning regression libraries was used in Scenario-2 in the extreme-weather model analysis.

- *Maximum Likelihood Classification*
- *Sieve classes*
- *Clump classes*

D.2.4 ML experimental libraries Scenario-2, Component model

The following machine learning regression libraries was used in Scenario-2 in the component model analysis.

- *LinerRegression*
- *Lassso*
- *ElasticNet*
- *DescisionTreeRegressor*
- *KneighborsRegressor*
- *SVR*
- Ensemble RandomForestRegressor
- Ensemble GradientBoostingRegressor
- Ensemble ExtraTreesRegressor
- Ensemble AdaBoostRegressor

D.2.4 ML experimental libraries Scenario-2, System model

The following machine learning binary classification libraries was used in Scenario-2 in the system model analysis.

- *LogisticRegression*
- *DecisionTreeClassifier*
- *KNeighborsClassifier*
- *LinearDiscriminantAnalysis*

- *GaussianNB*

- *SVC*

- Ensemble AdaBoostClassifier

- Ensemble GradientBoostingClassifier

- Ensemble RandomForestClassifier

- Ensemble ExtraTreesClassifier

D.3 SETTING UP DEEP LEARNING WORKSTATION

Some deep learning applications like image processing using convolutional neural networks and recurrent neural networks, or scaling to a larger dataset, is fundamental to set the experimental environment to handle such restrictions. In this study we set the experimental for deep learning in two workstations. 1) *Dell (Alienware) GPU Intel® core i7*, and 2) *Intel® Xeon® CPU E5430 @ 2.66 GHz x 8 (CenOS Linux 7)* remote cluster. Below we briefly detailed the dependencies needed in this study in order to use some deep learning algorithms between the specialized deep learning packages, i.e., to use *Keras* the installation of TensorFlow or Theano or both in our case. The Python dependencies can be installed thru *pip* or *conda install*.

While the progress of the ML and DL package, and the ML and DL evolving application areas, in a so fast-paced that any current references in the state of the art and current version, is updated at any moment, thus, only as a main reference for this study the language version is presented. Furthermore, the steps to setting up the experiment platforms are briefly detailed below.

D.3.1 Dependencies

Language Platform: Python 3.6.7

Python Scikit-learn. Is a complete package for machine learning tasks in a Python platform requiring other dependencies like *Python Numpy* and *Python Scipy*. (http://scikit-learng.org).

Python SciPy. Is a collection of functions for scientific computing in Python. It provides function to handle N- dimensional arrays among other functionality like: advanced linear algebra routines, mathematical function routines, and statistical distributions.

Python NumPy. Is one of the essential packages for scientific computing in Python, containing multidimensional arrays, high-level mathematical functions such as linear algebra operation and other functions.

Python Pandas. Is a package for data wrangling and analysis. Focusing on reading, writing, and manipulating data, this dependency as Scikit-learn need the installation prior to install Pandas.

Python Matplotlib. Is one of the most important plotting packages for Python, used to plot the accuracy of the model by creating 2D charts and plots from data.

Python Keras. Is a minimalist Python package for deep learning, which provides enough to achieve an outcome. Keras is a lightweight API that can run on top of Theano or Tensorflow by the process of numerical libraries called *backends*, providing a consistent interface to Theano or **Python Thensorflow.** In this study, be used both backends in Keras: *Theano backends* and *Tensorflow backends*.

D.3.2 DLNN experimental libraries Scenario-2, Component model

The following Deep Learning Neural Network (Multi-Layer Perceptron) libraries was used in Scenario-2 in the component model analysis.

- *Sequential*
- *Dense*
- *Dropout*
- *KerasClassifier*
- *Maxnoem*
- *SGD*
- *Cross_val_score*
- *LabelEncoder*
- *StratifiedKFold*
- *StandardScaler*

- *Pipeline*

D.3.3 DLNN experimental libraries Scenario-2, System model

The following Deep Learning Neural Network (Multi-Layer Perceptron) libraries was used in Scenario-2 in the system model analysis.

- *Sequential*
- *Dense*
- *Dropout*
- *KerasClassifier*
- *Maxnoem*
- *SGD*
- *Cross_val_score*
- *LabelEncoder*
- *StratifiedKFold*
- *StandardScaler*
- *Pipeline*

VITA

Mrs. Rossana Villegas is a PhD. Candidate under the Electrical and Computer Engineering at the University of Texas at El Paso (UTEP), Texas, USA since May 2018. She received a Secondary College Major in Electric Power and Engineering Systems from UTEP in 2018. She received her Master's Degree in Control and Automation from University of Juarez City in 1999. She received her Bachelor of Science Degree in Electrical Engineering from University of Juarez City in 1987.

She worked as Coordinator of Industrial Automation Laboratory. She also worked as coordinator of the computer lab at the Biological Science Institute (BSI). She also worked as Research Assistant in Distributed Computer Lab at UTEP from Spring to Summer 2018.

Her teaching experience cover taught courses in Logic Control System, Robotic, Control I (Analog System), and Control II (Digital System). Her teaching experience also includes as a Lecturer at the University of Texas at El Paso (UTEP) with taught courses in Electromechanical systems, and in University 1301 Seminar. Working also as a Teacher Assistance in Digital and Design II Lab, Electric Circuit I, Electric Circuit II, Introduction of Sensor, Instrumentation, and Measurement, and as a grader assistance in courses in Advanced Functional Materials, Micro-Electro-Mechanical (MEMS) and Bio-MEMS, and as Lab Inventory assistance in Nanomaterials and Devices Lab. She has several publications in journals and conferences, including several tutorials, practices handbook, theses supervised, proposals funded, scholarships, honors, and awards.

She is currently member of the IEEE-ETA KAPPA NU (Zeta Delta Chapter) since 2016, and IEEE Member, including membership in Control System, and Power Energy Society since 2013 to present.

Contact Information: rvillegas7@miners.utep.edu