**Intelligent and Resilient Energy Management in Smart Cities Using Ensemble Methods**

Rajdeep Das

Project Thesis Report

AUGUST 2022

**TABLE OF CONTENTS**

DEDICATION iii

ACKNOWLEDGEMENTS iv

ABSTRACT v

LIST OF TABLES vi

LIST OF FIGURES vii-viii

LIST OF ABBREVIATIONS ix-xi

[CHAPTER 1: INTRODUCTION](#_heading=h.gjdgxs) 1-6

* 1. Background Overview 1-2
  2. Energy System Resilience 2
  3. [Problem Statement](#_heading=h.30j0zll) 2-3
  4. Research Gaps 3
  5. [Proposed Objectives](#_heading=h.2et92p0) 3
  6. Research Questions 4-5
  7. [Scope of Study](#_heading=h.tyjcwt) 5
  8. Significance of Study 5
  9. Organization of Thesis 5-6

[CHAPTER 2: LITERATURE REVIEW](#_heading=h.1fob9te) 7-14

* 1. Resilience framework for energy grid
  2. Resilience and Reliability

2.3 Resilience and Robustness 14

2.4 Weakness in Energy Grids

2.5 Common types of smart grid failure

2.6 Techniques to detect different failures in smart grid

2.7 Artificial Intelligence Techniques in Smart Grid System

2.8 Fault Diagnosis in Smart Grid

2.9 Summary

CHAPTER 3: RESEARCH METHODOLOGY 15-35

* 1. Introduction 15-16
  2. Primary Work Layout
  3. Solving Approach
  4. Energy Management model for Grid outage
     1. Dataset Collection 18-19
     2. Data Pre-processing 19
     3. Data Visualization 20-22
     4. Modelling 22-35
        1. Decision Trees 22-28
        2. Linear Regression 28-32
        3. Exponential Smoothing 32-24
  5. Summary

CHAPTER 4: IMPLEMENTATION AND RESULTS ANALYSIS

4.1 Introduction 36

* 1. Implementation Details 36
     1. Tools 36-38
        1. Pandas Library 36-37
        2. Numpy Library 37
        3. Matplotlib 37

4.2.1.4 Scikit-learn 38

4.3 Evaluation Metrics and Techniques 38

4.4 Results 38-51

4.5 Comparative Analysis of Models 50-52

4.6 Summary 52-53

CHAPTER 5: CONCLUSION AND FUTURE WORK 54-57

5.1 Conclusion 54-56

5.2 Future Recommendations 56-57

REFERENCES 58-61

APPENDIX 62-74

Appendix-1: Dataset Details 62-67

Appendix-2: Code Snippets 68-74

# 

# ACKNOWLEDGEMENTS

# I would like to express my sincere and heartfelt gratitude to the authorities of IIT Kharagpur, especially, I offer my genuine thanks to its chairman, (name), Vice-Chairman (name) and Principal (name) for their constant support and encouragement. I would like to place on record my immense debt and heartfelt gratitude to (name), Head, (department name), (college name) and my colleagues for their motivation during my research work.

# I express my sincere and heartfelt gratitude to my research supervisor, (Guide name here) (occupation of guide), (department name here), (University/college name here) (place name) for her inspiration and constant guidance all through the research work. I extend my sincere gratitude to my well-wishers, colleagues, friends and beloved students for their encouragement and support right through the years of my research. Finally, my greatest debt and profound gratitude to my parents and other family members, who have encouraged and blessed me with their wishes and prayers.

# Rajdeep Das

# ABSTRACT

Energy management is one of the most important requirements in urban cities because of its vital role and its complicated systems. By 2050, about 70% of the whole population will live in cities as predicted and to handle sustainability issues regarding increased energy consumption these cities need to be smart. The concepts of a smart city and smart grid basically depend on the extent of the correlation between information systems and communication technologies.

Recently, Machine learning (ML) methods have been used very widely in the development of prediction models for energy consumption. These models extremely improve the precision of traditional predicting tools and this would provide a lot of advantages both at the environmental and economic levels. In this research, we propose a strategy using an ensemble learning methods to build intelligent and resilient Grid Energy Management in smart cities. These intelligent methods contain integrating various learning models so as to enhance the results gained by every singular model.

# LIST OF TABLES

Table 2.1 Differences between resilience and robustness 14

Table 2.2 Classification of HILF events and examples 15

Table 2.3 Faults that have been mentioned in the literature 16

Table 2.4 Statistics for large blackout causes in U.S. 22

Table 2.5 Techniques to Detect and locate Faults in Smart Grid 23

# LIST OF FIGURES

Figure 1.1. General Architecture of Smart Grid System Utilized in Smart City 3

Figure 1.2. Entities and Service Components in Cloud Integrated Smart Grid Architecture 4

Figure 1.3. Network Architecture in Smart Grid for Communication 6

Figure 1.4. Process of power flow, power system and information flow in Smart Grid 7

# LIST OF ABBREVIATIONS

# Machine Learning……..ML

# Logistic Regression……..LOR

# Decision Tree………….DT

# Artificial Intelligence….AI

# Artificial Neural Network…ANN

# Support Vector Machine…..SVM

# k- Nearest Neighbor……….KNN

# Random Forest….RF

# Convolution Neural Network ….CNN

# Attribute Selection Measure ......ASM

# Cross Validation……CV

# CHAPTER 1

# INTRODUCTION

* 1. **Overview**

Nowadays, the concepts of smart city form a major chance for humankind to meet the increasing risk of environmental challenges, also to build comprehensive cities that focus on the citizens' life quality for a big deal. The concept of a smart city is just similar to the smart grid which deeply relies on the interconnection between all the systems, hence communication, and information technologies play a main role. However, the smart grid and smart cities are new huge topics and complex. Smart cities are aiming to deal with resource optimization and maximum efficiency, rapid urbanization and population growth generate a lot of problems, such as waste management, mobility, and energy supply. Currently, 54% of the world’s population lives in urban areas and it is expected that this percentage will grow to two-thirds of the entire population by 2050. According to statistics, the absolute value showed that by the year 2050 more than six billion people might live in cities [2].

Energy is the most challenge the world is encounter today thus the main goal is energy consumption to be reduced. The prime sector of energy consumption depends on buildings, transport, and the industry which is the main problem that we take into account. The energy needs of cities are intricate and plentiful. As a result, contemporary cities should take use of the synergies between all of these energy solutions to enhance their current systems and install new ones in a coordinated and efficient manner. Consequently, cities are pioneers in the necessary move towards a more efficient economy and low pollution [3]. We can increase communities' resilience by making the energy system more strong, affordable and improving more reliable energy. The necessity to move from the traditional energy models to the aggregated modern energy models has been emphasized by city-level energy policy management which is the main goal of our research.

## **1.2. Energy System Resilience**

The capacity of a system to absorb and adjust to extreme events is known as resilience, and it is a key attribute anticipated of crucial vital systems such as power systems. Modern civilizations rely on easy access to vast amounts of energy; hence the efficient operation of energy systems is essential to their survival. Energy systems and their surroundings are changing quickly as a result of the increasing reliance on energy and energy-based services. In addition to other advancements, the advancement of technology and environmental deterioration are creating new risks that may be challenging or even impossible to anticipate and assess before disruptions occur in practice.

Regardless of how they affect energy systems, certain trends, like climate change, may be undesirable [2], while others, like electrification [3] and digitalization [4], may benefit society more than it would cost to mitigate the risks they entail. Regardless of where they come from, many emerging concerns significantly raise potential risks and harms [5–9]. Traditional risk management techniques may thus not be adequate to handle these [10]. A significant portion of the literature on energy system resilience and overlapping energy security focuses on specific types of threats (such as weather [14-22], technical failures [19,23,24], cyber-attacks [25-29], and geopolitics [29-31]), as well as threats for a particular energy sector (such as electricity [16-20,25-29], oil, and gas [31-34]). By examining and defining relevant terms, ideas, and viewpoints on the challenges, some studies have made an effort to present a more comprehensive understanding of energy security and energy resilience [35,36]. However, energy system resilience is highly tied to the sorts of threats and energy systems in issue, limiting the concept's applicability and necessitating stronger clarification or classification of scenarios considered.

**1.3 Important Concepts**

This research attempts to strengthen the foundation of complicated considerations required for resilience enhancement initiatives. The specific objectives are to, define the term "energy system resilience", sketch out a broad panorama of risks to energy systems from the standpoint of resilience, create a more complete picture of risks and responses for a certain sort of high-risk threat. The definition of resilience in the context of the energy system includes vocabulary and assessment methods. Given the lack of agreement on essential word meanings, introducing the terminology is a required first step. The section on evaluation methods expands on the concepts used to describe energy system resilience in more complex, yet system and threat independent, terms. A basic and grounded framework that incorporates the most important views and observable trends in the literature is used to map out the wide landscape of threats. The landscape includes all energy system levels, industries, and supply networks, as well as various danger characteristics. This is followed by a more extensive explanation of weather-related risks (the source of the majority of energy supply interruptions) and cyber-attacks (a tiny but quickly growing problem). Cyber-attack dangers are examined in terms of rising digitalization of energy systems, increased interest and capability of adversaries, and solutions accessible to energy system operators. Before proceeding further, there are some important related concepts that need to be discussed in-depth for acquiring more knowledge about this thesis work.

**1.3.1 General Cloud Integrated Smart Grid Architecture**

# In industrial sector, the traditional methods of managing the infrastructure have been revolutionized by a technology named as Smart grid. The general framework of smart grid involves three domains i.e. generation, transmission and distribution that is amalgamated with each-other to provide best customer services as shown in Figure 1.1. It basically consists of various traditional power stations whereas the transmission domain provides various services to the distributed domain. Nodes can communicate using multihop network protocols specifically designed for wireless sensor and actor network (WSANs). Typically, one (or more) nodes act as a gateway and provide network connectivity at a given location with a LAN or Internet infrastructure. Cloud computing platforms can also be used to provide storage, analysis, processing, and decision-making services for a smart grid system. In addition, the control center and various users can collect information and issue information and commands to ensure the real-time control of the corresponding system [43].

# figure 3

# Figure 1.1. General Architecture of Smart Grid System Utilized in Smart City [11]

Cloud computing is beneficial in large number of grid applications as it offers high scalability and flexibility in order to build a robust software oriented service delivery infrastructure which provides ‘always-on’ connectivity [12]. There are several advantages of shifting the data from smart grid to cloud computing involving:

* No need to do investment of costly infrastructure. So, it is cost-effective.
* The consumption and price data can be shared in real time which enables the consumers the added informative data to use for further benefits.
* Provides secure and automated services with the help of streamlined sensor.

Irrespective of these benefits, there are critical concerns related to the limitations involved in transferring data from grid to cloud as:

* There are huge chances of data manipulation and various attacks that can be seen while sharing data from one platform to another. Therefore, it is vulnerable to the attacks which can provide threat to security.
* Data is exposed to third parties and various unwanted and malicious entities can destroy the data.

Figure 1.2 in more detailed way (elaboration of Figure 1.1) shows various entities and service

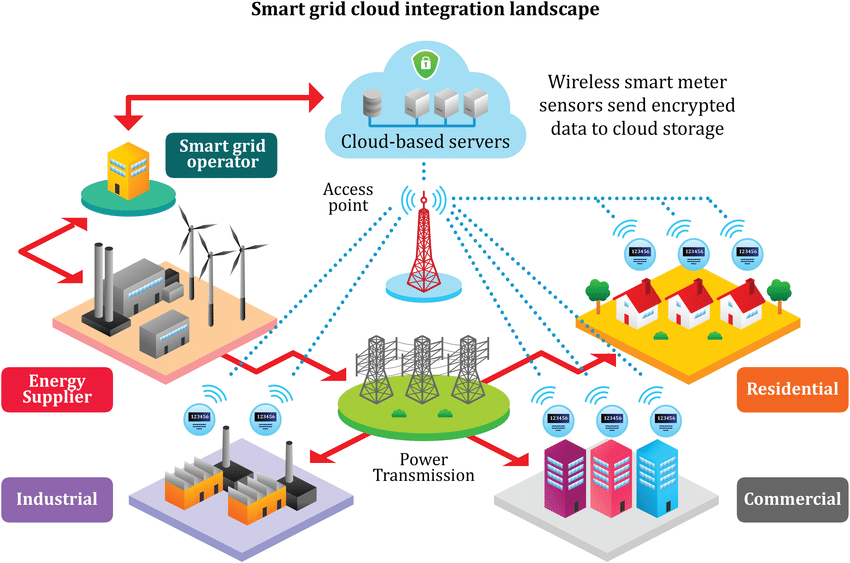


Figure 1.2. Entities and Service Components in Cloud Integrated Smart Grid Architecture [15]

# components in Cloud Integrated Smart Grid Architecture. The figure represents exchange of the data which is encrypted among grid operators and consumers viz cloud oriented services with the help of highly secured communication channels. The encryption method such as Homomorphic techniques data encryption techniques are employed to give the surety about data privacy over the storage phase in the cloud. This advantageous aspect of the applied encryption techniques helps to tackle the overheads involved in communication, complexity of involving large number of encryption and decryption processes, etc.

The transfer of various energy consumption use rates data acquired by sensor meters can also indicate the consumer attributes, behaviors and habits in the energy sector [16]. Further, the use of smart grid entities involving energy suppliers, grid operators, etc. gather the information about energy consumption based on which other data such as lifestyle of the client can be observed. The further subsection discusses about the network architecture of small grid in more detail.

**1.3.2 Network Architecture in Smart Grid for Communication**

# The architecture of Smart Grid system consists of a number of domains including Bulk Generation, Bulk Transmission, Bulk Distribution, Customer, Markets, and Service Providers Operations. The proper assurance of flow of power and information is the responsibility of first two domains. Further, the collection of information and management of power in Smart Grid is the responsibility of last three domains.

# Now, to connect these seven domains, the network should follow a proper hierarchical distribution as shown in Figure 1.3. This figure is the representation of Smart Grid network in which various domains are interconnected involving the backbone network and a bunch of local area networks (LAN). To establish the inter-domain connectivity, backbone network is created. This network comprises of a number of nodes which are interconnected to each-other and involve either gateways for LAN or routers having very high bandwidth to exchange messages in Smart Grid efficiently. In a backbone network, conventional wired communication technologies such as fiber optic technologies can be used to achieve high-speed data and mass delivery of information across domains. For example, a SCADA system is a system for monitoring power operations across the Operations, Transmission and Distribution domains. All power signal quality samples are transferred from local systems in the Transmission and Distribution part of domains through the backbone network to the Operations domain to perform the centralized management. A local network is used for communication within the domain.

Figure 1.3. Entities and Service Components in Cloud Integrated Smart Grid Architecture [17]

#### A local area network (LAN) comprises of ad hoc nodes, which are meters, sensors or intelligent electronic devices (IEDs) installed in the energy infrastructure. They are usually equipped with a small number of bandwidth and have high computing power for certain monitoring or protection purposes. To represent the process of power flow, power system and information flow in Smart Grid as explained in previous page, Figure 1.4 is utilized. Smart Grid is a modern and popular intelligent network adapted to the new global platform. Intelligent systems connect various components such as energy, information and communication technologies and can be used for various purposes like improving performance, quality, production, transmission, distribution and marketing services. This flow of information and power from generation to consumption by the help of the power system through WAN, Neighborhood Area Network (NAN) and Home Area Network (HAN) can result in simulation and deployment of Intelligent Environments with Sensors in the Home, Data Control Architecture, Complex Computer Systems. Therefore, it directly reflects the relationship between Smart Grid and Internet of Things (IOT) which can improve people’s quality of life to a greater extent.

# 

Figure 1.4. Entities and Service Components in Cloud Integrated Smart Grid Architecture []

**1.3.3 Cyber-Physical (CP) Structure in Smart Grid**

Cyber-Physical (CP) attacks is the growing but an important topic that remains largely unclear. The interconnection of huge number of devices (millions in number) in the field constructed a large attack surface, while the associated vulnerabilities were remotely positioned across CP domains; moreover, interfaces among CP structures are also prone to various attacks launched from both the domains. A typical CP structure for a smart grid is shown in Figure 1.5.

Knowledgeable attackers can directly damage vulnerabilities of control systems to exert sudden and significant impacts in the Smart grid. Based on the target systems or policy, the effect of control-based attacks ranges from transient voltage and frequency instability, steady-state line overloading and load shedding, to massive blackouts resulting from cascading failures. Measurement-based attacks pose another critical and crucial threat in the Smart grid. Instead of directly manipulating the control signals, attackers can compromise the measurements to weaken situation awareness by concealing the occurrence of disturbances or mislead control actions by reporting non-existing contingencies. Both types of the attacks are within the scope of CP attacks. Two types of factors are commonly considered in the design of a critical attack scheme: the cost of the attacker and the cost of the defender. For the attacker, the costs typically include the large number of resources and knowledge required to launch the attack; in many cases, the risk of being detected is also included. For defenders or operators, costs typically comprises of equipment damage, power outages, and financial loss. Connected power systems are managed by their respective control zones. Between control zones, redundant power generated in one zone flows along the inter-connect to another zone. Generation in each control area is required to meet dynamic load demand, and grid nodes are required to maintain a great balance between the load and generation.

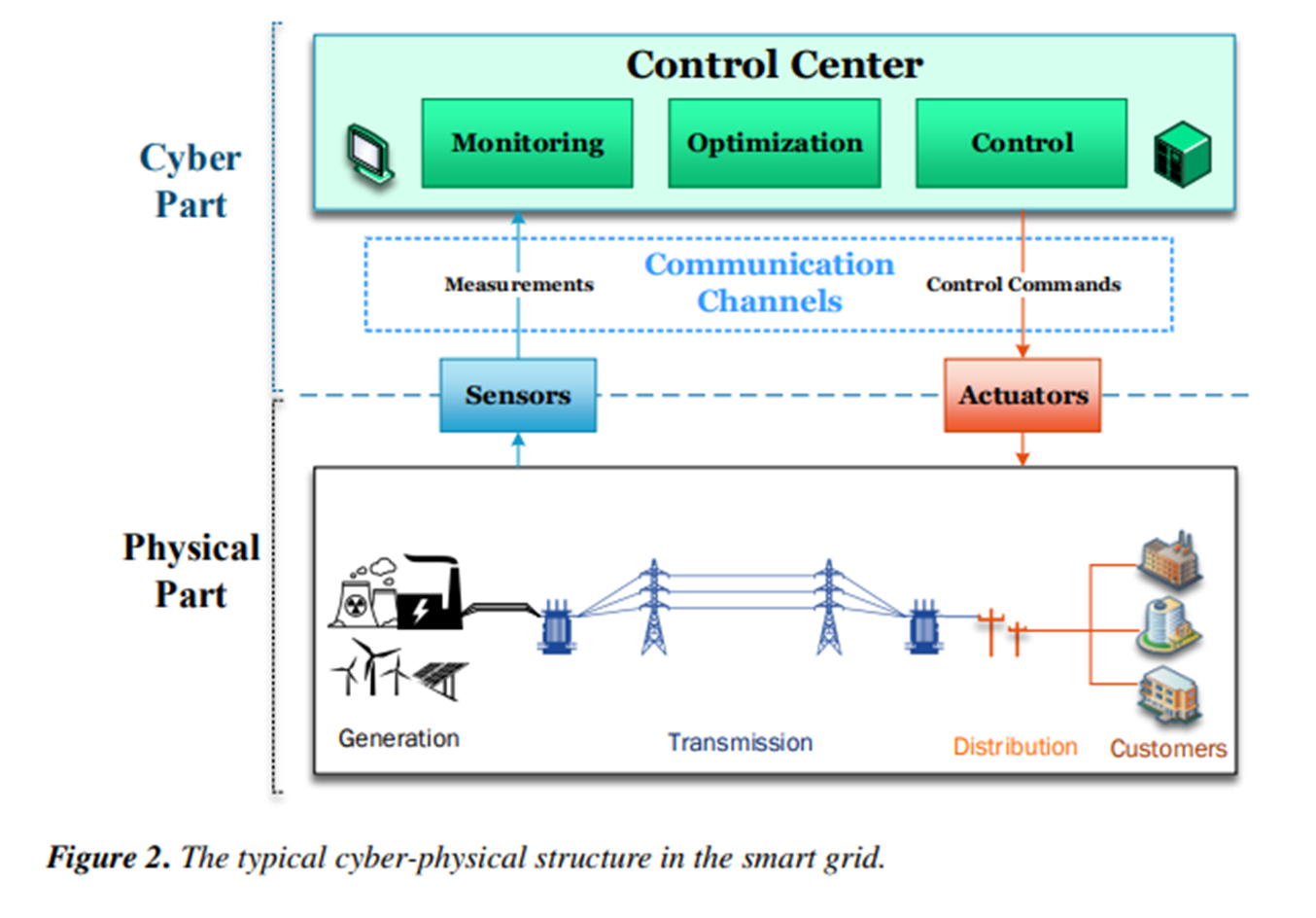


Figure 1.5. Typical CP Structure in Small Grid

**1.3.4 Typical Power System Control Loop**

Power systems are functionally divided into generation, transmission and distribution. The control center receives readings and measurements from the input device known as sensors that

Interact with field devices(power lines, transformers, etc.). The algorithms running in the Control Center process these measurements to make operational decisions. The decisions are then sent to the actors to implement these changes in the field device. Figure 1.6 shows a generic control loop representing this interaction between the control center and the physical system.

Smart grids can provide an efficient way to supply and consume energy by providing two-way energy flow and communication. It can integrate multiple renewable distributed energy resources (DER), which is environmentally friendly, has low greenhouse gas emissions, and effectively reduces transmission power loss. The associated connectivity and advanced information/communication infrastructure make smart grids vulnerable to Cyber- attacks. Energy sector statistics show that there were more than 150 Cyber-attacks in 2013, and he had more than 79 in 2014[73]. Costs as a result of blackouts About $80 billion a year in the US. Energy suppliers typically amortize this by raising energy prices, but

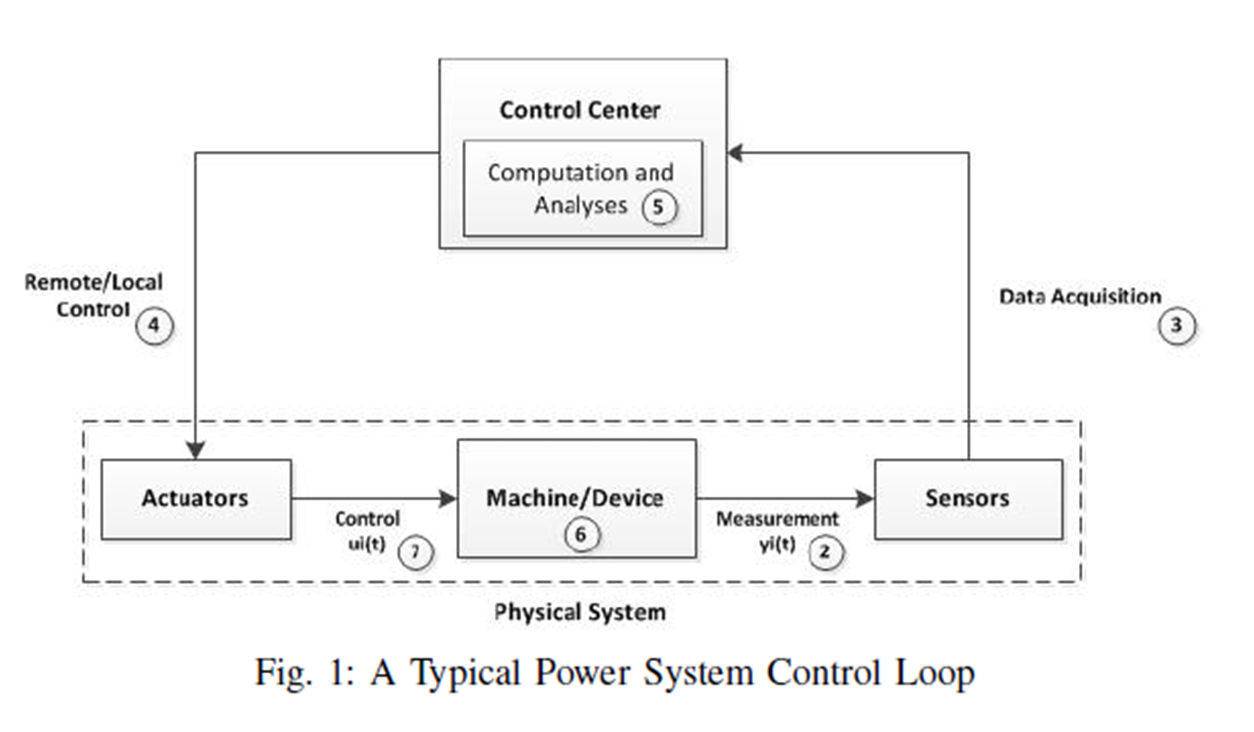
****

Figure 1.5. A typical Power System Control Loop

unfortunately this is passed on to consumers at a cost [373]. Renewable energy microgrids, including DER, could be a potential solution, but their production patterns depend on weather and environmental conditions and need to be properly monitored. One of the features of the smart grid is the ability to integrate multiple microgrids and monitor them using a highly reliable-communication network.

**1.4 Scope of Study**

In this work, we have used ensemble methods to make an energy management for resilient smart grid system, which predicts the power outage in adverse weather and in case of cyber-attack. It also handles the resiliency of the grid to high-low impact probability in the case of attack and provides architecture of the energy management system which states the working of the system on the smart grid in case of attack. The prediction and classification of various factors such as natural disaster outages, energy components state is performed in this work. The results obtained after experimentation can be utilized in further research works for more robust analysis and proposal of energy management systems.

**1.5 Significance of Study**

Following the recent power outages in numerous regions of the world, grid resilience, which has been a topic of industry concern as extreme weather events are occurring more regularly and the threat of cyber-attacks is expanding, has risen to the top of a larger public discussion. The development of resilience measurements and frameworks to direct the targeted deployment of cutting-edge technologies on the grid is a top priority for the electric power sector. The development of such metrics and frameworks is extremely complex because utilities must be able to compare the costs of taking action (the cost of deploying technology under ideal conditions to increase power system resilience) to the costs of inaction (increased societal impacts and recovery costs) in the context of a risk environment that is rapidly changing and becoming more uncertain.

**2.14 Organization of Thesis**

After Chapter 1 i.e. Introduction, the rest of the thesis is organized as follows: Chapter 2 provides the detail of the literature review, problem statement, research gaps and proposed objectives. The proposed research methodology is presented in detail in Chapter 3. Chapter 4 provides a description of the implementation and an analysis of the results obtained. Finally, the conclusion and future work is provided in Chapter 5 which is followed by a list of references.

# CHAPTER 2

# LITERATURE REVIEW

# 

# Before discussing the research techniques used in this thesis, a foundation of academic and industry information will be evaluated and provided in the following review of literature. The analysis of data obtained after performing a literature review guided this work about the aspects that need to be covered and may be proved to be beneficial for the community as a whole.

# Resilience Framework for the Energy Grid

The energy grid is more vulnerable to outages than ever before, due to extreme weather disasters, human attacks or faults, ageing, and other causes will be discussed in more detail later. The extreme events such as (hurricanes and earthquakes) and man-made attacks like (physical and cyber and attacks or errors) has roughly impact on energy systems starting from long outage times to main components' ruin. In order to protect and maintain the energy system operations and components, system resilience must be high thus failures controlling must be fast. Issues foreseeing and controlling is basic to play down synchronization issues and to make strides in the framework financially. In addition, it is critical to identify and detect the fault within the smart grid system quickly sufficient to avoid a total breakdown within the system [8] [9] [10]. In [11] possible solutions, strategies for resilience improvement and several modeling have been provided. Many studies concentrated on fault events caused by natural causes. However, reasons such as attacks or human errors, as well as their faults, are rarely reported. Since even a small event can cause an energy outage, centralized energy plants, substations, transmission lines, energy transformers, as well as distributed generation (DG), all is considered potential weak points.

As previously mentioned, resilience may have a variety of definitions and criteria. In [12] the author attempted to standardize and explain the concept of resilience in the field of energy grids. According to the findings of this study, a resilient system can evaluate threats and take a series of actions over time to ensure its efficiency in the challenging situations, attacks, or faults. According to the English Cabinet Office, infrastructure resilience is achieved by effective system and network architecture, which ensures the necessary resistance, reliability, and the capacity to turn or split the grid into other sections (redundancy) in order to preserve service continuity through an outage [13] [14].

Moreover, to be a resilient system, it must have the capacity and capability to adapt and recover. Jufri in [15], discussed that the amount of disruption caused by an extreme event on the grid or the ability of the grid to continue operating during the affected condition was used to measure power grid stability. Above mentioned energy grid resilience framework has two common structures: evaluation and enhancement. The risks are evaluated in the grid evaluation, and the energy grid system is improved in the grid enhancement to ensure availability and reliability during an outage and reduce the amount of time to bring back to its normal case.

Many other authors have attempted to identify system resilience in temporal line. In [16] Bruneau introduced the resilience triangle and taking into consideration that the system does not have a degraded state. Figure 2.1 illustrates what is known as the resilience triangle. Panteli concluded in 2017 that for a system to be prepared to handle with the conditions that are associated with a fault effectively, it must exhibit the characteristics of the resilience trapezoid, as seen in Figure 2 [17].

# Figure 2.1. Resilience triangle

**2.2 Resilience and Reliability**

Reliability is a term that describes a system's overall capacity to function satisfactorily. Under specific situations, energy systems are designed to be reliable throughout a given time interval. Due to two phenomena, increased frequency and length of extreme natural events, and a start moving in the generation mix toward renewable energy resources that rely on climatic conditions to operate, the idea of resilience has gained traction in recent years. The meaning and definition of resilience, as well as the distinction between resilience and dependability, blurs the lines between the two ideas and confounds the definitions. Different organizations have adopted definitions of resilience that encompass both dependability and resilience in some situations, or focus only on high-level emergency management in others, underlining the need for the resilience concept to be properly understood and assessed. Table 2 below shows the main differences between resilience and reliability [24].

There has been a lot of new development about resilience engineering, a new and rising discipline. This field has quickly aligned itself with several existing safety disciplines, but it has also removed itself from human reliability analysis. To date, the debate has been relatively one-sided, with significant emphasis placed on the novel insights provided by resilience engineering. Grid resilience stems from the need to establish values and actions to safeguard the grid from disasters and stress, as well as build contingency plans. Grid resilience is defined as the ability of the grid to withstand grid interruptions and return to normal operating circumstances without or with minimum external intervention, or to change with the pressure to reduce compromise through graceful decline.

Grid resilience is defined in Policy Directive (PPD) 21 as the capacity to promptly plan for and adapt to extreme situations, as well as resist and recover from disturbances. The ability to survive and recover from natural disasters or direct anthropogenic dangers is referred to as resilience.

The increase in power outages caused by extreme weather as a result of worsened climate change has accelerated research towards improving community resilience. Several researchers and policymakers have contributed to the characterization and parameterization of energy resilience and reliability in particular, which necessitates a collection of coordinated studies to highlight the findings and incorporate them into future grid resilience and reliability enhancement efforts. The concepts of grid resilience and dependability should be defined and separated so that the systems may be easily understood, appraised, and operated to ensure perfect functioning and environmental sustainability. This study [1] satisfies the stated objectives by delving into grid resilience and dependability, their measurement metrics, and augmentation strategies. Using Monte Carlo Simulations and the stated measures, the article also divides the United States into four groups based on grid dependability and grid resilience. This paper proposes two new terms, resilience risk factor and grid infrastructure density, which will be important criteria in determining grid resilience.

**2.3 Resilience and Robustness**

The terms resilience and robustness are used similarly in the disciplines in the social and organizational systems, they are regarded differently in the field of power systems, as shown in Table 3. In actuality, resilience and robustness are two distinct design approaches. The first is concerned with system flexibility, whereas the second is concerned with system strength. Robustness is a fundamental component of a robust system, and so falls under the umbrella of resilience. Robustness is generally built into the system's architecture, whereas resilience is built into its operating components. In extreme events, a robust grid may break like a tree, but a resilient grid may break and survive like a reed. Table 2.1 below shows the main differences between resilience and robustness [25].

Table 2.1 Differences between resilience and robustness

| **Resilience** | **Robustness** |
| --- | --- |
| Capability to resist and recover from external shocks. | Ability to withstand change without losing stability |
| Flexibility, adaptability, and speed are required. | Stronger connection between network components is required. |
| Service quality is prioritized. | The focus is on asset use. |
| An active approach | A passive approach |
| Segments network into several subsystems to continue functionality | Segments network into limited sub-systems to continue functionality |
| Depend on agility and flexibility to survive extreme disturbances | Depend on component redundancy and topological changes to cope with specific threats |
| Considers unexpected catastrophic failures caused by High Impact Low Frequency (HILF) events | Considers only expected failures |

Based on WG C2.25 survey results [27], table 2.2 provides a comprehensive categorization and includes instances of HILF events. HILF has been classified into two types of events: physical, which includes man-made or natural hazards, and cyber threats.

Natural, man-made, and man-made disasters, as well as cyber threats, have the potential to have major consequences for the energy system and grid operability due to the loss of grid infrastructure assets and the operation of vital grid services. As a result, adequate preparations and countermeasures must be devised and executed to handle such occurrences, taking the probability and degree of their potential impact into account.

Table 2.2 Classification of HILF events and examples

| **Classification of threats** | **Examples** |
| --- | --- |
| Physical-man- made | Physical Security violations, Terrorist Threats, Vandalism pandemic |
| Physical natural | Seismic Events, Floods, Hurricanes / Superstorms, landslides / avalanches, Snow / ice Storm, Tsunamis, Wildfires / Cyclones, Drought, Earthquake |
| cyber | Denial of service, Malware, Man-in-the-middle |

**2.4 Weakness in Energy Grid**

In order to determine which failures occur more frequently and what causes them, we need to highlight the weaknesses of the energy grid and summarize the literature review. The objective is to know the faults and the related causes of them. The scale of the cause would have an impact on the consequences of the emerged faults. If a small cause occurs, it will result in a minor fault that will only affect a few residential houses and will be simple to fix, perhaps in a matter of hours. In most situations, particularly in developing nations, the national electric grid's fragility is due to overcrowding, aged infrastructure, high temperatures, and severe weather. Power outages in such systems can last from several hours to many days. To avoid power outages, users might utilize diesel generators, which are both noisy and dirty. Alternative, renewable energy sources and storage devices, on the other hand, can be employed for load shifting. This has the potential to increase grid resiliency and supply continuity. On the other hand, if the cause of failure is large-scale, such as a terrorist attack or a hurricane, it will result in a large-scale fault, such as an outage or a cascading failure, impacting a large geographic region and probably taking days or even weeks to recovery from. According to research, three major cause clusters have been identified:

* **Errors**: creates due to the human error or equipment failure.
* **Natural Causes:** Numerous types of Natural disasters that can trigger a failure in the energy system. Hurricanes, winds, floods, earthquakes, storms, and extreme weather are examples of natural disasters.
* **Attacks:** It may be of cyber or physical source. Cyber-attacks like denial of service (DOS) which is the most common attack, or physical attacks such as human-made attacks (terrorism).

According to the literature review, Natural Causes are listed in 84 % of publications. Hurricanes and storms are the most often mentioned natural causes, with 22% of articles citing them, followed by other natural phenomena such as thunderstorms, which were discussed in 20% of the articles in this survey. Windstorms and tornadoes are mentioned in 14% of the articles, and earthquakes are mentioned in 11% of the articles [26] [27] [28]. Human and equipment failures were analyzed simultaneously, representing 17% of the papers reviewed. Finally, cyber-attacks represent 17% of reading papers, which is a large percentage as compared to physical attacks, which represent just 6% of the literature reviewed for this survey. Table 2.3 lists the main references where these causes/faults are discussed [29] [30].

Table 2.3 Faults that have been mentioned in the literature

| Causes | Faults | Refs |
| --- | --- | --- |
| Natural | Blackout  Cascading fault  Collapse of transmission towers  Damage and faults on substations  Fault currents  Fault of transformers  Faults and damage to overhead transmission  Stability limits exceeded  Power loss  Thermal overloads  Underground cable loads affected | [7][31][32][33][34][15][35] |
| Errors | Blackout  Cascading outages  Fault currents  Fault of transformers  Frequency deviation  Voltage and frequency instabilities  Line faults  Line overloads | [7][36][12][37] |
| Attacks | Blackout  Cascading failures  Control infrastructures of smart grids affected  Delay, blockage or corruption  Widespread damage  Power loss  Localized blackouts and momentary interruptions | [1][7][33][12][38] |

**2.5 Common types of Grid Failures**

High-speed wind, falling trees, lightning, storms, human errors and assaults, insufficient insulation, overloads, protection failure, and other factors can all cause energy failure in smart grids. Smart grid failures are classified as two forms of abnormal electric currents in the energy system [39]. External faults and interior faults, the first of which can be short circuits or open circuits, and the second of which can occur on a DC bus or in a storage system. [40]. To summarize, breakdowns in the smart grid system can occur at numerous points such as the producing part, transmission part, or load portion, resulting in unforeseen places in the energy system. In [4] faults in the smart grid system are classified into three types: abrupt, intermittent and incipient faults. An incipient fault is an arc fault that keeps on for a harmless as mentioned in [41]. In [42] authors discussed that power failure in the system causes sudden signals changing which they defined as abrupt faults. Finally, authors in [43] defined a short period temporary that is considered to be an incipient event that leads to a permanent fault in the system as the intermittent fault. Large-scale outages have also significant economic and social consequences for users. When a large-scale failure occurs, the robust system is supposed to recover and restore to its original state. According to research, three major cause clusters have been identified. Authors in [44] argued that advanced machine learning algorithms may use acquired data to predict and also provide early warnings regarding grid faults and instabilities on a system-wide level, such as frequency oscillations between generators, as well as on a local level, such as impeding component and line failure.

The reasons for outages vary in many significant ways. According to [45] the two most important differences are as follow: (1) how warning system operators can take protective action before disruption is coming, and (2) how the energy system stays effective as soon as the interruption has dissipated and how much the physical and cyber control systems can manage it. We can review the causes of outage as follow (Figure 2.2):

1. *Earthquake*

The possible for disruption of main energy system component by earthquake is high. This phenomenon often occurs in the West for example, the central Mississippi valley and the coastal area of South Carolina. This can result in significant damage to substations, transmission towers and distribution poles. For instance, after a magnitude 5.8 earthquake struck Virginia in 2011, the North Anna Nuclear Power Station was shut down for more than ten weeks while the owner and operator performed extensive damage evaluation and the Nuclear Regulatory Commission approved restart approval [46]. Although earthquakes generally occur without warning, the natural frequency of earthquake waves is very slower than the speed of light, enabling for several seconds of advance warning in some cases with proper instrumentation. If that possible, a notification like this might offer enough time to de-energize vital components and reduce damage.

1. *Physical Attack*

Physical attacks on main system components, particularly large transformers and other complicated substation and transmission equipment, may result in serious physical harm. Physical attack such as attacks, bombings, and terrorist activity targeted transmission and distribution than generation systems around the world for example, in Afghanistan, Colombia, Iraq, Peru, and Thailand [47]. This might occur without warning or with just a limited amount of warning.

1. *Cyber Attack*

Preventing intruders into vital systems and detecting and eliminating malware before it becomes active are the strongest defenses against cyber-attacks. The results of a successful cyber-attack could be almost immediate, the results of a successful cyber-attack could be almost immediate, it may require a few seconds to a few minutes to completely manifest, or an attacker could remain hidden for months collecting data, as happened in the 2015 cyber-attack on Ukraine's energy system. In certain situations, cyber-attacks do not result in significant physical damage to the device, while physical damage may result in certain instances, particularly if the hackers are professional. It may be necessary to take certain action to reduce the potential harmful effects between the detection of an intrusion and the occurrence of any consequences. It is unclear how long it would take to recover depending on the severity of the attack.

1. *Operations Error*

A lot of historical outages have been induced by one or more errors, usually when the system is heavily loaded, which might have been handled if a series of consequential system failures had not occurred. Cascading failure, for example, occurred in the southwestern United States in2011, when a failure at a single substation in Arizona quickly developed into a network outage across Southern California [46]. Operator error may occur over time periods ranging from minutes to hours, depending on system conditions and the type of faults, and there could be chances to detect faults and take the appropriate action. The risks of operator error due to cascading failure have been decreased by improved control techniques. Simultaneously, external risks such as cybercrime and pandemics can put operators under stress and possibly increase the chances of errors.

1. *Tsunamis*

It is usually limited to coastal areas. While locating large facilities in areas not vulnerable to tsunamis is the best way to minimize risks to the energy grid, neglecting and transferring existing installations is costly, and there may be other safety measures that can be taken, such as elevating backup generators. This is will be more important consideration in utility planning such as in Hawaii and along the West Coast in U.S.

1. *Ice Storms*

As demonstrated by the 1998 ice storms in Ontario and upstate New York, ice storms or freezing rain can cause significant damage, with full recovery taking weeks. Ice storms disrupt energy by gathering ice on transmission or distribution lines, causing lines to collapse and poles and towers to be damaged. Moreover, when wind blowing against ice-laden transmission lines can lead to high-amplitude (1m) oscillations low-frequency (1Hz) that stress towers and insulators even more. Ice buildup on surrounding trees can cause branches to fall on power lines or bring plants close enough for an arcing current to create a short. Snow storms are a major cause of power outages across the country, but they do not get the same level of worldwide attention as more focused events such as hurricanes.

1. *Floods*

It may harm distribution or transmission lines and their pilings, as well as ground-based facilities. Several energy providers have used historical flood data to select locations for infrastructure facilities, such as substations, that are impossible to be flooded. Nevertheless, as the environment changes, so does the rate of flooding.

1. *Hurricanes and extreme Weather*

Tropical cyclones could impact negatively on energy systems. As a hurricane approaches, modern prediction techniques usually offer advance notification, with more detailed and reliable predictions about the hurricane's strength and landing location. Some of the main causes of power grid disruptions are hurricanes and tropical storms for example, Hurricane Katrina and Sandy has also been a massive and meteorological turmoil which has led to outages in 17 states and the Columbia district and has had comparatively short-lived impacts. Moreover, weather conditions may cause massive disruption to the energy system. Scientific research about the causes of extreme weather events, as well as the detecting changes in risk, varies from country to country. Most rapidly-changing risks, such as the probability of more frequent extreme heavy rainfall events as well as more intensive hurricanes, are fairly well understood in both cases. Extreme weather events and natural disasters are the major cause of power outages, resulting in significant economic, social, and physical disruptions and cause considerable inconvenience for residents living in disaster areas. It costs the U.S. economy between $20 billion and $55 billion every year [48]. Furthermore, the northwestern part of Saudi Arabia suffered a blackout, forcing residents to rely on the headlights of their vehicles for illumination, according to the Al-Medina newspaper. The outage in the Tabuk region lasted three hours and was caused by damage to electrical wires caused by severe weather changes in the Kingdom's northwestern region in 2018 as reported in saudi gazette n.d. A sudden and unexpected outage is pretty harmful, particularly when hotels and hospitals are affected. In [49] a logistic regression-based outage prediction model was suggested to assess the probability of power grid component outages in the event of an impending hurricane.

1. *Volcanic Activity*

In most of the world this is not an issue such as in Arabian countries, but it is a high risk for the explosion, ash fall, lava flow and lahars in some countries such as the Pacific Northwest and Hawaii. Isolate vital infrastructure from sensitive sites is the best method for preventing issues is simply. Even so, the geographical range of falling ash can extend the immediate danger area considerably and can affect insulators and transformers that can be disabled.

1. *Drought*

The effects of drying or drought on energy systems are numerous, for example, decreased hydroelectricity generation, increased power use for treatment and pumping or low cooling water in energy systems. As the IPCC states, “medium trust exists that in certain seasons and areas droughts will rise in the 21st century as the precipitation is decreased or evapotranspiration increased” [50]. The energy system gets really stressed by excessive heat. Consequently, the chance of hardware failure or user error leading to major failure rises when the energy system is heavily stressed. The 2014 U.S. National Climate Assessment state that "the global mean surface temperature is increasing at a higher frequency with longer-lasting heat waves" [51].

1. *Wildfire*

Fire normally may have a significant effect on particular substations and transmission systems but not cause widespread disruption to the energy system. As a result of continuing climate change, climate experts have long expected wildfires to be more widespread and intensive.

Natural disasters caused about 679 power outages in the United States between 2003 and2012, each affecting at least 50 000 consumers [52]. From 1984 through2006, Hines et al. in [53] identify 933 incidents that caused disruptions and Table 2.4 shows the information.

Table 2.4 Statistics for large blackout causes in U.S.[53]

| **Cause** | **% of events** | **Mean size in MW** | **Mean size in customers** |
| --- | --- | --- | --- |
| **Earthquake** | 0.8 | 1,408 | 375,900 |
| **Tornado** | 2.8 | 367 | 115,439 |
| **Hurricane/tropical storm** | 4.2 | 1,309 | 782,695 |
| **Ice storm** | 5.0 | 1,152 | 343,448 |
| **Lightning** | 11.3 | 270 | 70,944 |
| **Wind/rain** | 14.8 | 793 | 185,199 |
| **Other cold weather** | 5.5 | 542 | 150,255 |
| **Fire** | 5.2 | 431 | 111,244 |
| **Intentional attack** | 1.6 | 340 | 24,572 |
| **Supply shortage** | 5.3 | 341 | 138,957 |
| **Other external cause** | 4.8 | 710 | 246,071 |
| **Equipment failure** | 29.7 | 379 | 57,140 |
| **Operator error** | 10.1 | 489 | 105,322 |
| **Voltage reduction** | 7.7 | 153 | 212,900 |
| **Volunteer reduction** | 5.9 | 190 | 134,543 |

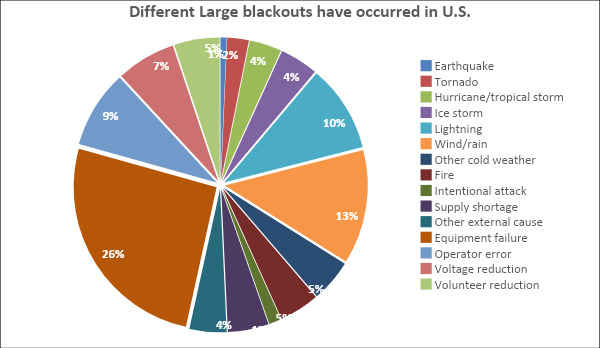


Figure 2.2. Blackouts occurrance in U.S.

**2.6 Techniques to detect different failure in smart grid**

To ensure system stability, fault detection systems must be fast enough to identify a power outage. The Table 2.5 explains the many methodologies that some researchers examined in order to detect numerous types of power failures in the system. At every node in the system, measurement devices may be utilized to read critical data, which helps in the detection of anomalous changes.

After detecting a fault event, it is critical to locate the location of the energy failure in the system in terms of dealing with it and maintain system stability. To detect a failure in the system, a variety of methods are used. The Figure 2.3 displays a chart that helps in the selection of several viable approaches for locating the system's power failure.

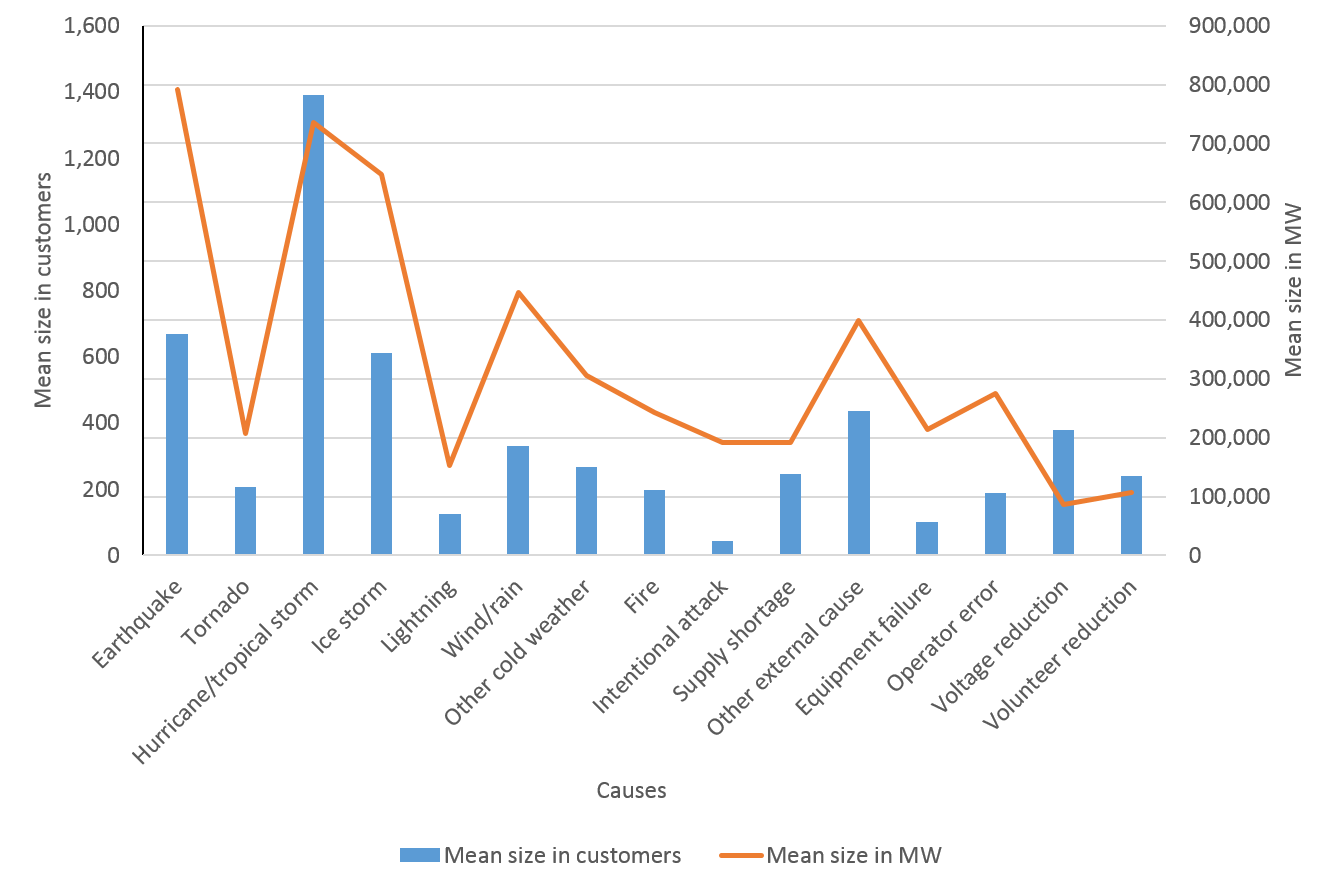
  
Figure 2.3. Selection of several viable approaches for locating the system's power failure

Table 2.5 Techniques to Detect and locate Faults in Smart Grid

| **Ref** | **Methods** | **Weakness** |
| --- | --- | --- |
| [2] | Data-driven computational approach based on machine learning matching pursuit decomposition (MPD) using Gaussian atom dictionary and hidden Markov model (HMM) | The phasor measuring device was unable to give sufficient information regarding the system's power failure. |
| [54] | Wavelet analysis in conjunction with a neural network or a support vector | Wavelet analysis needs a high sample rate, which reduces the method's accuracy. |
| [55] | Analyzing changes in impedance characteristics after a high-frequency current is injected into the system | High-quality smart meters are required. |
| [56] | Gaussian Markov method to locate the fault | Lack of accuracy based on the size of the power network. |
| [57] | Voltage sags and swell measurements to locate the fault | Lack of accuracy |

**2.7 Techniques to detect different failure in smart grid**

Due to the obvious hybrid energy system's rapid revolution, more distributed smart grid components, such as smart metering infrastructure, communication infrastructure, distributed energy resources, and electric vehicles, are tightly integrated into the power system by encompassing a massive power transmission energy grid with the underlying communication system. Those components create massive amounts of data in order to automate and improve smart grid performance by enabling a wide range of applications. Because traditional computing approaches are incapable of processing the massive amounts of data generated by smart grid systems, artificial intelligence (AI) techniques have attracted a lot of attention. Several research efforts have been directed toward investigating these AI techniques to address the problems, since they utilize large-scale data to improve smart grid performance.

AI techniques in smart grids may be broadly classified as follows:

* + - 1. **Expert Systems (ES):** A human specialist in the loop approach, which is utilized to solve specific issues.
      2. **Supervised learning:**An artificial intelligence model in which the mapping of inputs and outputs has been investigated in order to predict the outcomes of future inputs.
      3. **Unsupervised learning:** An ML class wherein unlabeled data are utilized to capture data similarity and difference.
      4. **Reinforcement learning (RL):** it is an intelligent agents technique, which seeks to maximize the concept of cumulative reward, distinguishing it from supervised and unsupervised learning.
      5. **Ensemble methods:** Combine the outcomes of many AI algorithms to overcome the constraints of a single algorithm while improving overall performance.

*Can Artificial Intelligence contribute to addressing energy grid issues?*

The answer is unequivocally "yes." Artificial intelligence (AI) has the potential to do the following:

1. Enhance grid resilience in the face of extreme weather and/or events.
2. Make general deployment of clean energy practicable in order to minimize carbon combustion and emissions.
3. Reduce the cost of producing energy to customers on a daily basis.

**2.8 Fault Diagnosis in Smart Grid**

In the context of the smart grid (SG), fault diagnosis (FD) along electricity transmission and distribution networks must be reliable, rapid, and accurate. Because of the complexity, including the installation of distributed generators (DG) and the changing character of power distribution networks, FD faces greater issues than transmission systems. Power distribution networks often span large regions and are made up of a large number of nodes, thousands of end user loads, a large number of distribution transformers, and small lines with varying resistances and inductances.

This paper [1] proposes the use of smart sensors and advanced communication technology that will be available in future smart grids to carry out automated fault diagnosis tasks using signal processing techniques. Standard deviation aspects of fault transient signal and fault location factors are used in suggested methods. The performance of different scaling levels, features, and components of fault transient current signals extracted using the most recent non-traditional Symlet mother wavelet function is assessed and compared. The attempt is made to pick appropriate fault transient current characteristics and components in order to enhance the performance of the existing restricted types of accessible fault locators.

# 2.9 Problem Statement

The energy management industry in today's world has become a vital focus of concern for nearly every government across the globe. Since the development of smart grid there has been a growing need to predict the attacks on the grid and also maintain its resiliency in every condition and case. Hence, usage of ensemble methods to predict the attacks to the grid is the need of the hour. Also, ensemble methods are efficient in terms of data processing and backup and predicting threats, systems are to be developed so as to make this happen and with the growing AI industry the previous methods are proving to be inefficient in comparison to the growing need for smart grid to provide and manage energy across the globe.

**2.10 Research Gaps**

The research work on former smart grid system identified certain gaps which motivated us to pursue further research on this domain:

* Lack of an efficient use of network resources and available power, to ensure reliability and security in the smart grid .
* Lack of time synchronization, and latency of data delivery with comparatively poor support of multicast in the smart grid system.
* Lack of significant change management in the smart grid and its processing.
* Lack of information and communication technologies, sensing, measurement, control and automation technologies, power electronics and energy storage technologies in terms of smart grid.
* Lack of use of hybrid learning techniques such as ensemble learning platform which can yield better results.

**2.11 Proposed Objectives**

This thesis work supports the development of an outage prediction model for resilient smart grid systems applying Ensemble methodologies and effective Machine Learning algorithms have been rigorously followed in this research project. This research work specifically focused on resolving the major problem of analysis of the energy management and resiliency of smart grids. This study investigates how in case of adverse conditions a smart grid is capable of maintaining its resiliency and backup the data in case of power outage and cyber-attacks. Therefore, based on the gaps obtained, this work proposed some objectives that need to be accomplished as highlighted below:

* To perform a comprehensive review of literature for a better understanding of the considered research area and the amount of work performed.
* To propose an effective power outage prediction model based on various machine and ensemble learning techniques.
* To predict natural disaster outages and energy component state using effective learning approaches.
* To analyze the attacks and extreme conditions using the applied learning models.
* To make a comparison of the employed machine learning approaches to each other for performance evaluation based on various performance evaluation parameters to find the best among all.

**2.12 Research Questions**

# Why do we need this Resilient Smart Grid System?

The quality of the power supply, rising energy demand, and significant aggregate technical and commercial (AT&C) losses are three major issues facing the electrical sector. Due to these difficulties, a smart grid system was created, bringing stability, sustainability, and the highest possible quality in order to fulfill the rising demand for power. The global need for energy is rising quickly. Smart Grid is urgently needed to satisfy the increase in demand and to offer consumers an affordable, dependable, and sustainable supply of power. Both the utility and the consumer gain from the installation of the smart grid. Hence is the rising demand for resilience of the smart grid developed. In order to assist in the speedy restoration of electricity after significant storms, utilities have historically hired and educated field technicians. For these repairs to happen, however:

* Customers who will be affected must first notify their utility suppliers, which take time.
* Then, utilities must send out field teams to identify the precise nature and location of the issue (using trial and error)
* Once the problem is localized, field teams may finally address the root cause.

Resilience in this situation is gauged by how long it takes for these two steps to follow one another. The notice, reaction, and repair times are continually being slashed by utilities and independent power providers. However, the world's greening grid's increasing complexity and reliance on the Internet of Things (IoT) have led to a rise in the need for tools to manage this complexity, especially in the face of new dangers like harsher storms or cyber-attacks.

# What are Ensemble methods and why use it?

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produce more accurate solutions than a single model would. The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator. Ensemble methods are ideal for regression and classification, where they reduce bias and variance to boost the accuracy of models. Sequential ensemble techniques and parallel ensemble techniques are the two main categories into which ensemble methods belong. Base learners are produced via sequential ensemble approaches, such as adaptive boosting (AdaBoost). The dependency between the basic learners is encouraged by their consecutive generation. The model's performance is then enhanced by giving previously misrepresented learners more weight.

Base learners are created in a parallel fashion, such as RF (Random Forest) in parallel ensemble approaches. To promote independence among the preliminary learners, parallel techniques make use of parallel generations of base learners. The mistake resulting from the use of averages is greatly decreased by the independence of base learners. The majority of ensemble approaches only use one algorithm for base learning, which makes all base learners homogeneous. Basis learners that have comparable traits and are of the same kind are referred to as homogenous base learners. In order to create heterogeneous ensembles, other approaches must be used to heterogeneous base learners. Diverse sorts of learners make up heterogeneous base learners.

**How will Ensemble methods help with the Resilient Smart Grid system?**

The traditional electric power grid will change from being an electromechanically managed system to an electronic network under the notion of the smart grid. Massive volumes of data are used in ensemble approaches to build intelligent computers that can do jobs that call for human intellect. AI systems are created using ensemble approaches that employ ML algorithms. Neural networks, robotics, expert systems, fuzzy logic, and natural language processing are some broad approaches to creating AI systems. In general, AI approaches allow for quick and precise decision-making. In smart grid applications, artificial intelligence (AI) is the process through which computers imitate the cognitive processes of grid operators to provide self-healing capabilities. But in other circumstances, AI might not be able to take the role of grid operators. Although using AI to improve smart grid systems can make them more accurate, dependable, and comprehensive, there are still numerous obstacles to overcome. Informatics that can facilitate grid operators' work performance is included in virtual AI systems. Self-aware AI systems that can optimize and manage certain grid activities with or without human involvement are included in physical AI systems. Artificial narrow intelligence (ANI) and artificial general intelligence (AI) are two further categories for AI systems in the smart grid (AGI). ANI refers to artificial intelligence (AI) systems created for particular jobs with relevant criteria and restrictions, such as an AI system that does load forecasting using various datasets.

**2.13 Summary**

This chapter reviews the various studies conducted on the faults that affect the operation and resilience of the energy grid system. It aims to improve the grid's resiliency. It starts with a clarification of the definition of smart grid resilience and presents a state- of- the-art review of the causes of outage, and consequent faults that affect energy smart grid. Natural reasons are mentioned 84 % in the literature reviewed for this study of causes, while human or technological mistakes are mentioned 22 %, and physical or cyber-attacks are mentioned 23 %. Hurricanes, storms, earthquakes, tornadoes, lightning storms, and heat waves are examples of natural disasters that can result in catastrophic disruptions. Errors, on the other hand, might result from either human or technical faults. The ability to resist an outage and maintain working during the damage state, as well as the ability to respond and recover after the outage, define the resilience of an energy system. This chapter covers the different types of energy failures that can happen in the smart grid. The various techniques used to prevent these failures also has been discussed. Further, this chapter provides the designed problem statement, identified research gaps from the literature, various research questions and proposed objectives.

# CHAPTER 3

# RESEARCH METHODOLOGY

* 1. **Introduction**

Research methodology plays an important role in the successful accomplishment of a task. It comprises several important steps starting from data collection to prediction results. The literature shows that there are numbers of prediction and classification methods exist that can be used to achieve high accuracy. This work applied different machine learning techniques either used as a single classifier or in the combination known as Ensemble learning method.

* 1. **Proposed Methodology**

At first the prediction is trained by a historical data set, this process as shown in Figure 3.1. is being carried out by ensemble methods and the accuracy of each data set is taken into account in

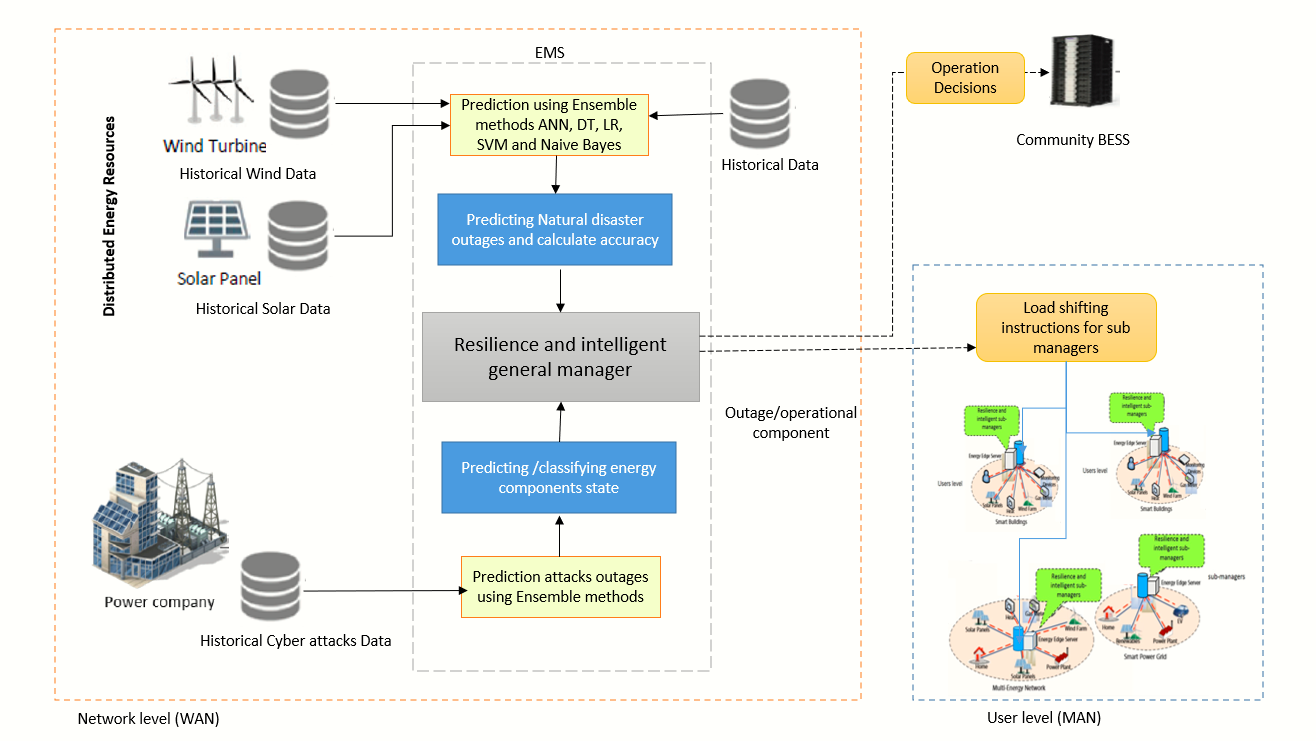


Figure 3.1. Process From Data Collection to Output

order to deliver a nearly correct prediction under all circumstances. During training of the model while the data set that are needed for distributed energy resources are split into two categories, i.e., the wind data and solar data. While in case of attack in the grid the data is categorized into Cyber Attack Data. The working of the model is further more categorized in two components that are the operation decisions and the outage/operational component. The later part mainly consists of load shifting instructions for multi energy networks, resilient power grid sub managers and smart buildings .These are handled at user level.

In general, the proposed framework or methodology followed in this thesis work analysis is comprised of several stages i.e. Data Collection, Data Preprocessing Feature Selection, Classification and Output, as shown in Figure 3.2. **Stage-I** includes the collection of relevant data related energy consumption which will be further used for the experimentation. In **Stage II**, data is preprocessed to make it clean to perform the operations efficiently. **Stage III** involves the implication of feature section techniques to extract the most relevant features among all. In **Stage-IV**, classification is performed is performed by employing machine learning techniques and lastly, **Stage-V** gives the output predicting various goals set in this work. A detailed explanation of all these stages is provided as under:

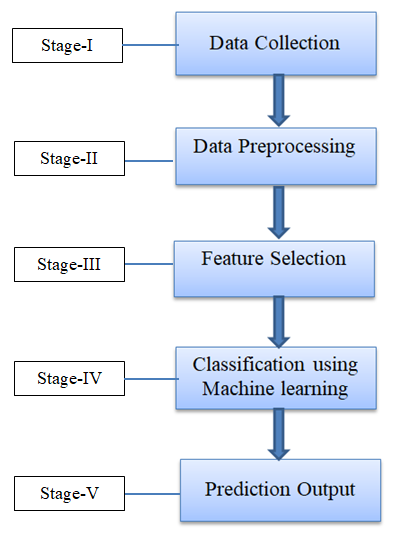


Figure 3.2. Proposed Methodology

* + 1. **Data Collection**

The resilient system will input the primary data set in two formats i.e., weather data and cyber-attack data as presented in Figure3.3. The weather consists of the solar and wind forecast and cyber-attack data set is concerned with attacks in the smart grid. The detection method used in the smart grid is for and to prevent threats, to protect systems against them, to respond to threats and to recover from them, rises in priority.

The architecture is capable of handling the increasing amount of data but it also allows in-depth monitoring. This enables the investigation of the impact of diverse threat scenarios like technical errors, human failure and cyber-attacks. Thus, it is the basis for developing resilient algorithms, improved communications channels and monitoring mechanisms for energy management systems. The community energy management system may be implemented using

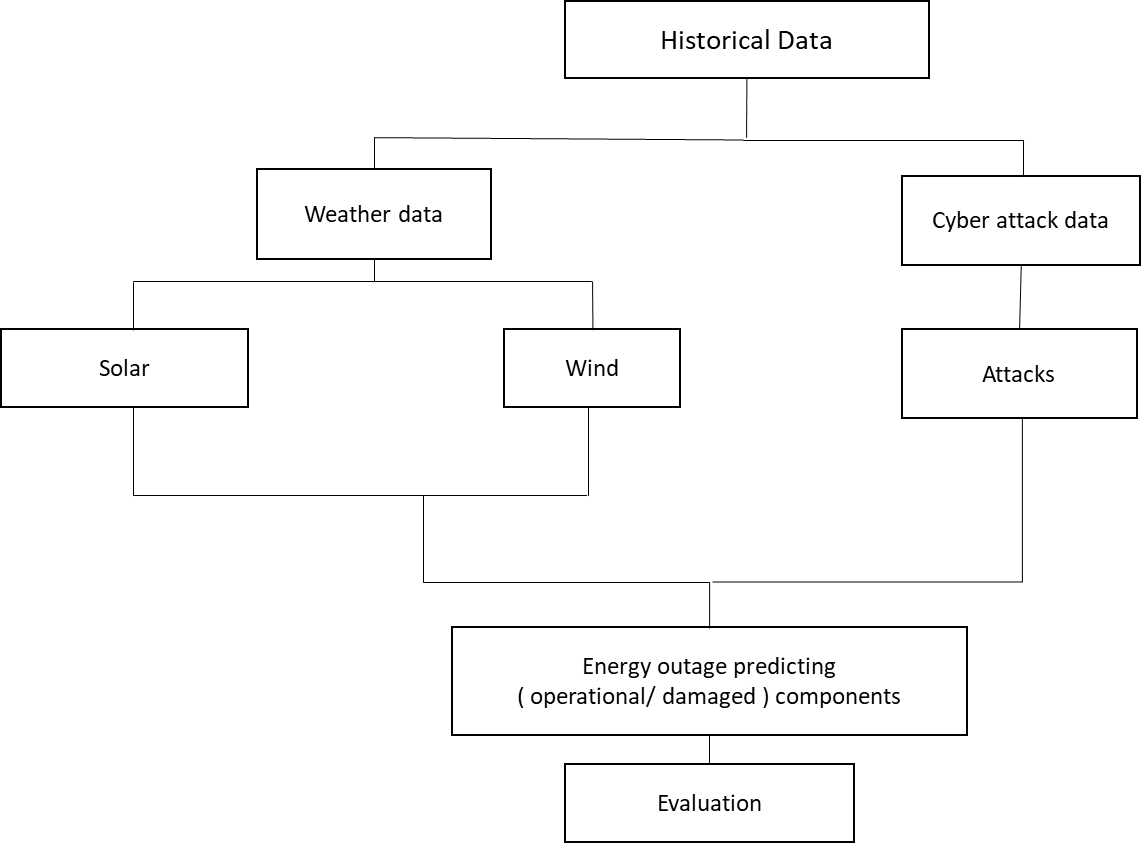


Figure 3.3. Data Collection Process

several domains of communication protocols: Wide Area Network (WAN), Neighborhood Area Network (NAN), and Home Area Network (HAN) (HAN). In the provided schematic, Wide area Network is used to facilitate the interchange of data/information between the EMS and the utility. NAN provides support for the communication between CEMS and HEMSs. Several wireless standards, such as WiMAX, 4G/5G, and IEEE 802.22, can be used in NAN [].

* + 1. **Dataset Preprocessing**

Preprocessing is a crucial stage in which various operations such as noise filtering, resizing, rescaling, cropping, etc. are performed to enhance the quality of the data. The purpose of preprocessing is to transform the raw into processed form to attain more accuracy and increase the performance of the model. Preprocessing is the form of data preparation which is most widely used in the process of data mining and data analysis in various fields. In this work, different data preprocessing techniques were used including data transformation, data cleaning, missing data handling and resampling as depicted in Figure 3.4.

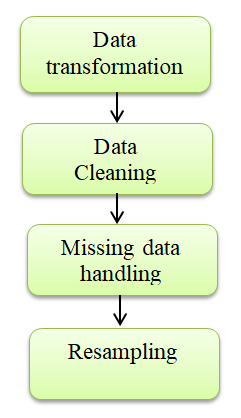


Figure 3.4. Data Preprocessing methods used

Let us overview these methods applied one by one:

1. **Data Transformation**

Data transformation is the process of converting data from one format or structure into another. Transformation processes can also be referred to as data wrangling, or data munging, transforming and mapping data from one "raw" data form into another format for warehousing and analyzing. Open source data has been used in this research. Independent and dependent variables can be found in this equation. The features are listed in the following table: As a user-server approach, the smart grid data interface was developed to make it easier to communicate amongst the various smart grids. The interface accepts crucial data in XML format. On the provider side, this file will also be used as a Hybrid Model parameter. Smart Grid Dataset was used to obtain the raw data. As a result, numerous ways have been used to clean up the data, including deleting duplicates and null entries and so on.

Data transformation or data conversion techniques include methods that help transform data before before applying machine learning algorithms. The aim is to remove unnecessary data and highlight the most important aspects of the data to be processed. Data conversion can be divided into the following steps: Each of these steps is applied based on the complexity of the transformation

• **Data Discovery:** This is a more exploratory step that involves profiling the data using data profiling tools or possibly manual scripts. The goal of this step is to understand the structure and properties of your data.

• **Data Mapping:** This is the process of defining individual field mappings, modifications, joins, filtering, aggregation, etc. Produce the final desired output.

• **Data Transformation Code**: This is the process of generating code (SQL, Python, R, etc.)

That transforms data based on data mapping rules.

• **Code implementation**: This is the process of running the generated code against the data

to produce the desired output.

• **Data Validation:** This process is designed to ensure that the output data meets conversion requirements. This step is most often performed by the company or the end user.

1. **Data Cleaning**

Most people concur that the quality of your insights and analysis while utilizing data depends on the data you are using. In essence, bad data equals bad analysis. If you want to develop a culture inside your business centered around sound data decision-making, one of the most crucial first stages is data cleaning, also known as data cleansing and data scrubbing. It involves correcting or eliminating inaccurate, damaged, improperly formatted, duplicate, or insufficient data from a dataset. There are several ways for data to be duplicated or incorrectly categorized when merging different data sources. Even though results and algorithms appear to be right, they are unreliable if the data is inaccurate. Because the procedures will differ from dataset to dataset, there is no one definitive way to specify the precise phases in the data cleaning process. But it is essential to create a template for your data cleaning procedure so you can be sure you are carrying it out correctly each time.

Therefore, data cleansing is the process of preparing data for analysis by removing or modifying

inaccurate, incomplete, irrelevant, redundant, or improperly formatted data. But, as mentioned above, it's not easy to clean up rows or remove information to make room for new data. Cleaning data is a muscle job. There's a reason data cleansing is the most important step when creating a data culture, not to mention valid predictions. It involves:

* Correcting Spelling and Syntax Errors
* Standardizing Data Records
* Correcting Errors
* Correcting mistakes such as null or empty fields
* Determining data points with duplicity

1. **Missing Data Handling**

Missing data are those values that are not there but would have significance if they were. Any number of things, including missing files, incomplete features, missing files, incomplete information, and data input errors, might result in missing data. In the actual world, the majority of databases have missing data. It's not always a problem when your data has missing values. Nevertheless, it presents a chance to carry out the proper feature engineering to direct the model to interpret the missing data in the proper manner. Missing data may be automatically found and handled using machine learning techniques and software. It is still advised to manually modify the missing data using analysis and coding methods.

Real data often have many missing values. Missing values ​​can be due to data corruption or data recording errors. Many machine learning algorithms do not support missing values, so it is very important to handle missing data during dataset preprocessing. Missing data can be due to human factors (such as deliberately not answering survey questions), electrical sensor issues, or other factors. Various crucial and important information can be lost when this happens.  
There is no perfect way to handle missing values ​​that gives accurate results for missing values. However, there are some techniques you can use to get decent performance. There are different ways to handle missing values in the dataset:

1. Discarding Rows having missing values
2. Input missing values for continuous variable
3. Impute missing values for categorical variable
4. Other Imputation Methods
5. Using Methods that support missing values
6. Prediction of missing values
7. Imputation using Deep Learning Library
8. **Data Resampling/Balancing**

Resampling methods are essential tools in modern statistics. You should iteratively draw samples from the training set and refit the model of interest to each sample to obtain additional information about the fitted model. This gives you more information than you could get by just tuning the model once. Typically, the goal of any data science project is to use training data to build a model to make predictions on new data. Therefore, the resampling method allows us to see how the model performs on untrained data without collecting new data. Resampling and regularization are two crucial steps that can significantly improve both the model’s performance as well as model’s reliability. There are different such methods that can be utilized including various type of cross-validations (CV).

Use cross-validation to estimate the test error associated with your model, assess its performance, or choose the appropriate level of flexibility. Evaluation of model performance is usually defined as model evaluation and model selection is used to select the degree of flexibility. The term is widely used in the field of data science. Leave-One-Out Cross-Validation (LOOCV) is a better option than the validation set approach. Instead of splitting the dataset into two subsets, only one observation is used for validation and the rest for model fitting.

Precision in predictive modelling is hindered by imbalanced classifications. In machine learning algorithms, the number of instances utilized for each class is the same. When it comes to the experiences of people of color, the models are wrong. This arrangement is dangerous because the minority group has more sway and is more prone to misclassifications than the majority. As a consequence, we could get rid of the outliers in the sample and recalculate the data. As a result of this work, new resampling techniques have been developed. We can, for instance, sample the majority of class data using sampling and remove records from each cluster to preserve information. Synthetic samples don't have to completely mimic minority class data; we may make small tweaks throughout the sampling process to obtain more varied samples. For data mining research, a balanced and uniform dataset is required [41].

* + 1. **Feature Selection**

In this thesis work, Features are selected by first selecting the percentile and finding the correlation between the attributes. Once the values are obtained, the attributes which showed high correlation to each-other were selected for the further experimentation as presented in Figure 3.5.

The dimensionality of data is growing rapidly, and poses challenges to most of the existing mining and learning techniques, such as: Feature selection has seemed to be an most effective and efficient approach for preparing high-dimensional data for the purpose of data mining and machine learning. The recent emergence of new technologies and new types of data and features not only advances existing research on feature selection to a larger extent, but also makes feature selection continuously more advanced and applicable to a wider range of applications.

In general, features are characterized as follows: (i) related: functions that affect the output and no other function can assume that role; (ii) irrelevant: functions that do not affect the output; (iii) redundant: one function plays the role of another function. can take over. The goal of feature selection is to find the best subset of m features with selected out of a total of n feature. A significant problem with many features selection methods

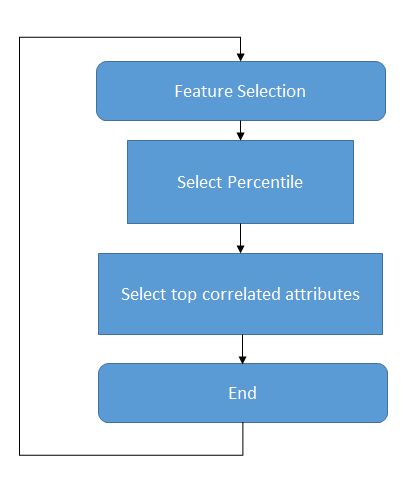


Figure 3.5. Process of Feature Selection

is the need to apply exhaustive search strategies to find the best subset among all possible feature subsets. This usually complicates the computation considerably. Alternative sub-optimal

feature selection methods offer more pragmatic solutions in terms of computational complexity, but we cannot promise that the final selected feature subset will be globally optimal. Trait association can be assessed individually (univariate approach), or multivariate. Univariate approaches are attractive because they are simple and fast. However, possible correlations and dependencies between characteristics are not considered. Multivariate search techniques are therefore useful. Some restrictions limit the use of multivariate search.

The dimension of a dataset increases with the size of the power grid. For example, the number of features in the measurement data for state estimation is 27 for the IEEE 9-bus system and is 1122 for the IEEE 300-bus system. With increasing feature dimensions the complexity and elapsed time for training the models increase very steeply, causing the so-called “curse of dimensionality” (Verleysen and François, 2005). To minimize this problem, feature selection is often used to eliminate the least discriminating features from the dataset, thereby reducing the dimensionality without sacrificing much of the information. Selecting the best feature and best number of features could lead the method to achieve its optimal performance while minimizing its running time. The feature selection phase is one of the most crucial phases of model classification, which can be done by various inbuilt mechanisms or by using domain knowledge. The ensemble framework currently supports random forest classifier (RFC) as a feature selection algorithm. Data from one domain is utilized to develop functions for learning machines using Feature Engineering. Extracting the most significant properties from raw data, it turns it into machine learning formats.

* + 1. **Classification**

An overview of the machine learning and ensemble-based attack detection method is given in this section. The approach puts individual and ensemble methods—supervised and unsupervised classifiers—into pipelines. The detailed schematic diagram depicting the different methods used in this thesis work is given in Figure 3.6. According to the diagram the various listed algorithms

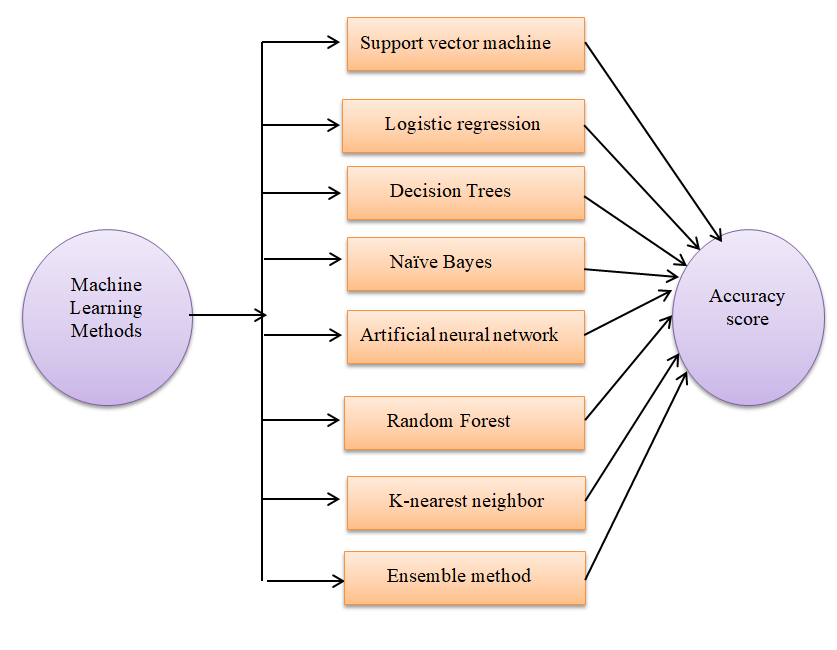


Figure 3.6. Machine Learning Techniques used

are run for each data set and the accuracy score is measured, until the highest score. While the testing is done online in real-time deployment mode, the training is done offline using historical

data.

* **Support Vector Machines (SVM)**

We seek a hyperplane that linearly separates attacked and secure measurements into two half spaces using hyperplanes in a dimensional feature space, which is constructed by a non-linear mapping A hyperplane is represented by a weight vector and a bias variable ,which results in

(3.1)

where is the feature vector of the sample that lies on the hyperplane in as shown in Fig (). The hyperplane is chosen such that it is at the largest distance from closest positive and negative samples. This constraint can be formulated as

(3.2)

* **Logistic Regression (LOR)**

Logistic regression works somewhat similar to linear regression. Logistic regression predicts the outcome on the basis of the individual characteristics of each feature. The logistic regression model is very easy to regularize. It calibrates output based on predicted possibilities. Suppose *Y* is a predicted output feature that depends on the predictor variable *X* and *X* can be given as ,*y* is regarded as the output variable given as follows:

(3.3)

(3.4)

(3.5)

where are the regression coefficients. It uses logarithmic or logistic function for cost evaluation.

* **Decision Trees (DT)**

Decision trees are one of the important methods in machine learning that works on linear as well as non-linear data. These algorithms work according to these algorithms works according to the rules made on data. The accuracy of the decision trees heavily depends on the decision to split the tree i.e. correctly deciding the number of splits. The basic motive of a decision tree is to predict the target variable’s value based on the simple decision rules extracted from the related features set. A decision tree employs a tree-like model to represent options and their potential results, including several variables and chance event outcomes. In a decision tree, the internal node is the depiction of the applied test and the tree branch represents the output of the test performed. A decision tress comprises a tree flowchart-like structure that can able to handle both categorical as well as numerical data. Therefore, it consists of leaf and decision nodes as depicted with the help of Figure 3.7.

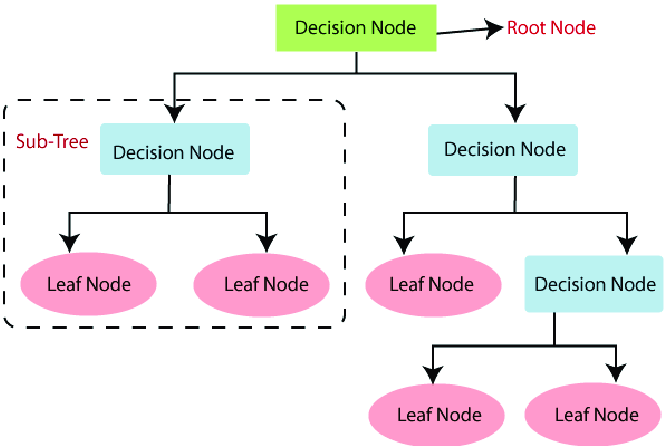


Figure 3.7. Decision tree flowchart structure (*Structure of the decision tree (DT) algorithm. | Download Scientific Diagram*, no date)

Decision Trees (*Decision Tree Algorithm, Explained - KDnuggets*, no date) are one of the important methods in machine learning that works on linear as well as non-linear data. These algorithms work according to these algorithms works according to the rules made on data. The accuracy of the decision trees heavily depends on the decision to split the tree i.e. correctly deciding the number of splits. This strategy is different for regression as well as classification trees. Therefore, multiple algorithms are utilized to divide a node into two or more sub-nodes.

The nodes are partitioned by the decision trees on all the variables which are available and the division that yields identical sub-nodes is chosen and selected.

Initially, the algorithm starts from the decision’s tree root node to predict the class in a given dataset. Then the values of the root are compared with the record attribute present in the real dataset. Depending on this comparison, the jump is made to the other branch to start with the next node.

The value of the attribute is re-compared with the value of other sub-nodes for the next node to move in a further direction. This procedure will continue till the algorithm reaches the destination node of the tree i.e. leaf node. The entire process can be easily understood with the help of the following algorithm.

1. Start with the root node of the tree ‘R’. This node includes the entire dataset.
2. Determine the attribute which is best among all in the dataset with the help of some attribute selection measure (ASM).
3. Split the root node ‘R’ into subsets ‘S’ which contain all the possible values for the attributes that are best.
4. The node i.e. decision tree node which has the best attribute is generated.
5. By utilizing the subsets of the entire dataset as constructed in step no.3, generate new decision trees recursively.
6. Continue with this procedure till where the further split of a node is not possible and this node will be the final node i.e. leaf node.

To select the attribute in step 2, there exist several methods but the most widely used methods include information gain and the Gini index. These measures tell the number of nodes in which the split should be done.

* **Artificial Neural Network (ANN)**

ANN is an important part of machine learning that has become most popular in large scale in research and developments today. ANNs are inspired by the biological human brain, which consists of a set of up to 60 trillion interconnected neurons to execute network patterns of decision making. Based on this basic idea, the artificial neural network process starts with very simple and easily understandable interconnected neurons acting as a single processor. His concept of a perceptron was introduced based on the neuron model by McCulloch and Pits. As a result of the very basic concept of the information processing cycle, ANNs perform complex mathematical formulations to arrive at optimal results for any given dataset or problem segment.

The human brain is made up of billions of cells called neurons. Information travels in electrical signals inside neurons. Any information that needs to be communicated to any other part of the brain is collected by the dendrites in the neuron, which is then processed in the neuron body and is passed on to other neurons via the axon. Another neuron can accept or reject a traveling electrical signal based on its strength. An ANN is a set of artificial neurons that try to mimic the functioning of a biological neuron inside the brain. Each connection can transmit a signal from one node to another, where it can be further processed and transmitted to the next connected artificial neuron. ANNs can learn and distinguish between very complex patterns that are difficult to manually extract and feed into a machine. ANN consists of three types of layers, each layer is a stack of artificial neurons. The number of artificial neurons in each layer can vary depending on your personal choice.

The basic structure of ANN is shown in Figure 3.8. The input layer is composed of artificial

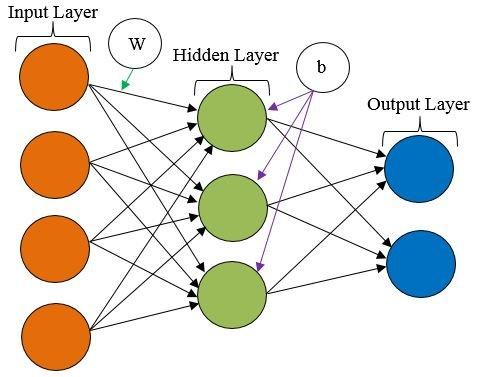


Figure 3.8. ANN Structure

neurons that take input data into the system for further processing by subsequent artificial neurons. The input layer is present at the beginning of the ANN. The weight and bias is associated with each input layer.. The hidden layer is between the input layer and the output layer. It takes a set of weighted inputs and produces an output via an activation function. This layer is called hidden because it is not an input or output layer. This is the layer where all the processing takes place. There may be one or more hidden layers depending on the complexity of the problem. The output layer is the last layer of the ANN architecture that produces the output for a particular problem. This is the last layer where any processing takes place before predicting the outcome of the problem. The summation and activation functions are related to ANN. The role of summation function is to sum input and weight while activation function produces output.

There are different types of activation functions that can be used in ANN but the most simplest one is the rectified linear activation function, or ReLU. For the output, the ReLU function typically selects the highest value from the linear combination of inputs from the preceding nodes [48]. ReLU was chosen since it produces either all zeros or all ones as its output. In addition, the grid includes all numerical functions within a specific range and is either stable (represented by "1") or unstable (represented by "0") with regard to our dataset. Because the dataset only includes two prediction classes, the "sigmoid" function is utilized as an activation function for the output layer, indicating that the dataset will be categorized logistically.

(3.7)

(3.8)

The adaptive optimization method, often known as "Adam," is an optimization approach which is most widely used to forecast grid stability in order to improve the performance of ANN. The Adam optimizer function can aids in ANN network weight optimization. The Adam optimizer also aids in improving the ANN model's learning rate.

* **Naive Bayes (NB)**

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks. In this article, we will understand the Naïve Bayes algorithm and all essential concepts so that there is no room for doubts in understanding.

The Naive Bayes classifier works on the principle of conditional probability, as given by the Bayes theorem.

The Bayes theorem gives us the conditional probability of event A, given that event B has occurred. In this case, the first coin toss will be B and the second coin toss A. This could be confusing because we've reversed the order of them and go from B to A instead of A to B. Naive Bayes classification is mainly used for weather prediction, medical diagnosis, and news classification. Some benefits of the Naive Bayes classifier are it is simple and easy to implement, doesn’t require as much training data, handles both continuous and discrete data, highly scalable with the number of predictors and data points, it is fast and can be used to make real-time predictions and is not sensitive to irrelevant features .

* **Random Forest (RF)**

A random forest is a commonly used machine learning algorithm trademarked by Leo Breiman and Adele Cutler that combines the output of multiple decision trees to produce a single result. Its ease of use and flexibility have fueled its adoption as it handles both classification and regression problems. The random forest method is shown in Figure 3.9.

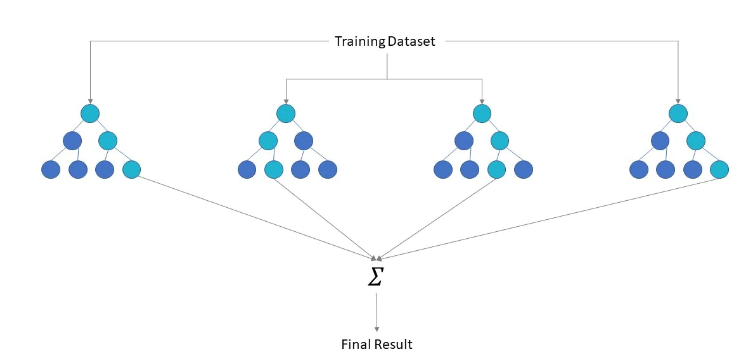


Figure 3.9. Random Forest method

The random forest algorithm is an extension of the bagging method because it uses both bagging and feature randomness to create an uncorrelated forest of decision trees. Phenomenal randomness, also known as element storage or random subspace method. It creates a random subset of features that ensures low correlation between decision trees. This is the key difference between decision trees and random forests. While decision trees consider all possible distributions of elements, random forests only select a subset of those elements.

Random forest algorithms have three main hyperparameters that need to be set before training. These include node size, number of trees, and number of elements sampled. From there, the random forest classifier can be used to solve regression or classification problems. A random forest algorithm consists of a collection of decision trees, and each tree in the set consists of a sample of data taken from a training set with replacement, called a bootstrap sample. From this training sample, one-third is set aside as test data, known as the out-of-bag (oob) sample, which we will return to later. An additional case of randomness is then introduced via feature storage, adding more variety to the dataset and reducing the correlation between decision trees. Depending on the type of problem, the determination of the prediction will vary. For the regression task, the individual decision trees will be averaged, and for the classification task, voting will be by majority - i.e. the most common categorical variable - gives the predicted class. Finally, the oob sample is used for cross-validation, completing this prediction.

* **K-nearest neighbor (KNN)**

A k-NN classifier is designed to classify unlabeled observations by assigning them to the class of most similar labeled examples. Observational features are collected for both training and testing data. For example, fruits, vegetables, and grains can be distinguished by their texture and sweetness (Figure 3.10). Only two features are used to represent them in a two-

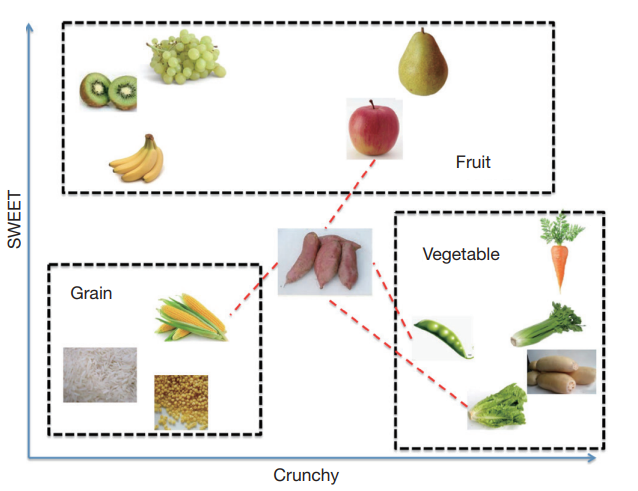


Figure 3.10. k-NN Algorithm Example

dimensional diagram. In practice, there can be any number of predictors, and this example can be extended to include any features. In general, fruits are sweeter than vegetables. The grains are neither crunchy nor sweet. His task for us is to determine which category sweet potato belongs to. For this example, select the four closest types of foods: apples, green beans, lettuce, and corn. Vegetables received the most votes in, so sweet potatoes are assigned to the vegetable class. It turns out that the key concepts of k-NN are easy to understand.

There are two important concepts in the example above. One is how to calculate the distance between sweet potatoes and other types of food. By default, the knn() function uses Euclidean distance. This can be calculated using below equation

 (3.8)

where *p* and *q* are objects to be compared with the features of *n*. Another concept is the parameter k, which determines how many neighbors the k-NN algorithm chooses. A good choice of A large k k has important implications for the diagnostic performance of the kNN algorithm. A large k reduces the effects of variance caused by random errors, but runs the risk of ignoring small but important patterns. The key to choosing an appropriate k value is finding a balance between overfitting and underfitting. Some authors suggest setting k equal to the square root of the number of observations in the training set.

https://www.plagiarismremover.net/assets/download_icon.svg

* **Ensemble Learning Method**

Ensemble learning is a machine learning paradigm where multiple models (often called weak learners) are trained to solve the same problem and combined to achieve better results. The main hypothesis is that when weak models are properly combined, we can obtain more accurate and/or robust models.

In ensemble learning theory, we call weak learner models (or base models) that can be used as building blocks for designing more complex models by combining several of them. Mostly, these basic models don't perform that well on their own, either because they have high bias (for example, low-degree-of-freedom models) or because they have too much variability to be robust (high-degree-of-freedom models, for example). Then the idea of ​​ensemble methods is to try to reduce the bias and/or variance of such weak learners by combining several of them together to create a strong learner (or ensemble model) that performs better.

In order to set up the ensemble learning method, we first need to select our base models to be aggregated. Most of the time (including the well-known storage and reinforcement methods) one basic learning algorithm is used, so we have homogeneous weak learners trained in different ways. The file model we get is then called homogeneous. However, there are also some methods that use different types of underlying learning algorithms: some heterogeneous weak learners are then combined into a heterogeneous ensemble model.

One important point is that our selection of weak learners should be consistent with the way we aggregate these models. If we choose base models with low variance but high variance, it should be with an aggregation method that tends to reduce variance, while if we choose base models with low variance but high variance, it should be with an aggregation method that tends to reduce the deviation.

In this thesis work, we have used the ensemble of five machine learning models and applied bagging technique for the voting and to predict the final output. Bagging is a homogeneous model of weak students that learns independently in parallel and combines their output to determine the average of the model. The concept of ensemble learning using bagging is presented with the help of Figure 3.11.

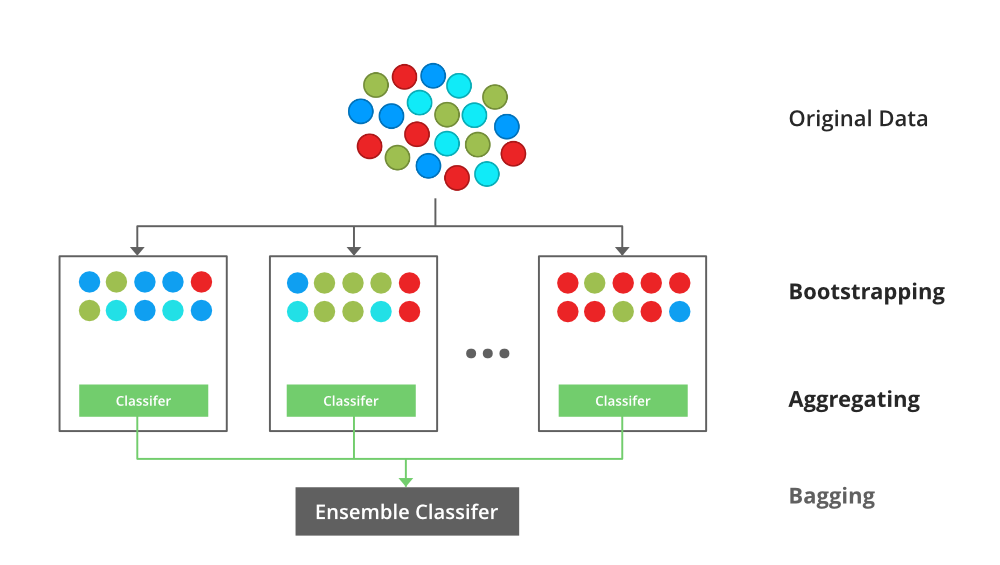


Figure 3.11. Ensemble learning technique with bagging

**3.3.5 Output**

After performing all the operations, the final output is obtained using different classifiers predicting the components state. Then, the performance of all th emthods is evaluated using various perfromance evaluation techniques to find the best one.

**3.3 Summary**

This chapter discusses the detailed research plan which is followed to achieve the purpose of

predicting the state of smart grid components. The methodology designed or proposed consists of five main stages i.e. Dataset collection, Data Preprocessing, Feature selection, Classification, and Output. Each stage is explained in a detailed manner to get an insight into the strategy and techniques which are used in this thesis work. Firstly, data is collected and then data is preprocessed to make it more meaningful for performing further operations. Afterward, feature selection techniques are applied to select the best and optimal features among all the features. Various machine learning techniques and ensemble learning approach are applied namely Decision trees, Logistic regression, Support vector machine (SVM), Naïve bayes (NB), K-nearest neighbor (KNN). Artificial neural network (ANN), Random Forest (RF) and ensemble of five classifiers. The output produced by each method is then comparatively analyzed to determine the best among them using evaluation metrics. The next chapter explains the implementation and the analysis results obtained after performing experimentation.

# CHAPTER 4

# IMPLEMENTATION AND RESULTS

# 4.1 Introduction

As earlier discussed, in this thesis work, we have implemented different machine learning models. The purpose was to suggest an Intelligent and resilient energy management in smart cities using machine learning and Ensemble Methods. The classification was divided into three parts: prediction of natural disaster outages, prediction of energy component state and lastly, to classify different types of attacks based on various effective techniques. This chapter discusses the evaluation metrics and methods used to evaluate the efficacy of the models in terms of accuracy and other parameters. The details regarding the tools, platform, libraries, etc. used for the implementation of the learning approaches are briefly defined in this chapter. The training and testing data are discussed and the results obtained are presented with the help of graphs. The efficiency of the applied models is estimated by comparatively analyzing them.

# Implementation Details

# 4.2.1 Tools

For the implementation of different techniques,thePython platform is used (*Top 10 Python Libraries - InterviewBit*, no date).Python is a general-purpose, high-level, interpreted programming language. Code readability is prioritized in its design philosophy, which employs heavy indentation. One of the important benefits of using this language is that it is simple to use and very easy to implement as it contains various in-built libraries and functions. Due to its application in various areas, this language has become the most popular language among academicians and researchers. Therefore, it comprises various libraries having different functions. The collection of modules is present in python’s library which includes huge snips of code which can be utilized again and again in various programs. To implement a method or something, there is no need the write the identical code several times. These libraries have a significant role in fields like machine learning, data science, data mining, data visualization, etc. Some of the important python libraries used in this work are discussed under (*Libraries in Python - GeeksforGeeks*, no date):

# Pandas

Pandas (Tutorials Point, 2015) are a Python-based open source data analysis and manipulation tool that is quick, strong, adaptable, and simple to use. It is one of the most popular libraries among data scientists. This library provides high-level data structure shaving greater flexibility and various crucial tools. The use of this library helps in easy analysis of data, manipulation, and making data cleaner. Various operations can be performed using the Panda library such as iteration, sorting, indexing, data conversions, etc. The following data can be used using this library as

* Dataset’s data
* Time series data
* Labeled data of rows and columns matrix
* Unlabeled data
* Statistical data

# NumPy

The meaning of the name NumPy refers to the Numerical python (Shell, 2012). It is a python library that supports faster processing of larger multi-dimensional arrays and matrices, along with a huge collection of mathematical functions especially from the section of linear algebra to apply to any arrays or vectors. With machine learning applications, this library is mostly preferred. It comprises built-in functions of mathematics which make it easier to use for performing computations. The other libraries of python like TensorFlow (Tutorials Point, 2015) also make use of this library to perform various tasks on tensors internally. Some of the features of NumPy are as:

* User-friendly and very interactive library.
* Allows easy implementation of complex mathematical functions and equations.
* Easy to understand and code.
* Array interface facility.

# Matplotlib

Matplotlib (Data, 2022) is a graph plotting library of Python and its extension NumPy. For integrating charts into programs utilizing all-purpose Graphical user interface (GUI) toolkits like Tkinter, wxPython, Qt, or GTK, it offers an object-oriented API. These libraries plot numerical data, therefore, mostly used for data analysis tasks. Matplotlib is open source and capable of producing high-definition figures such as graphs, histograms, etc.

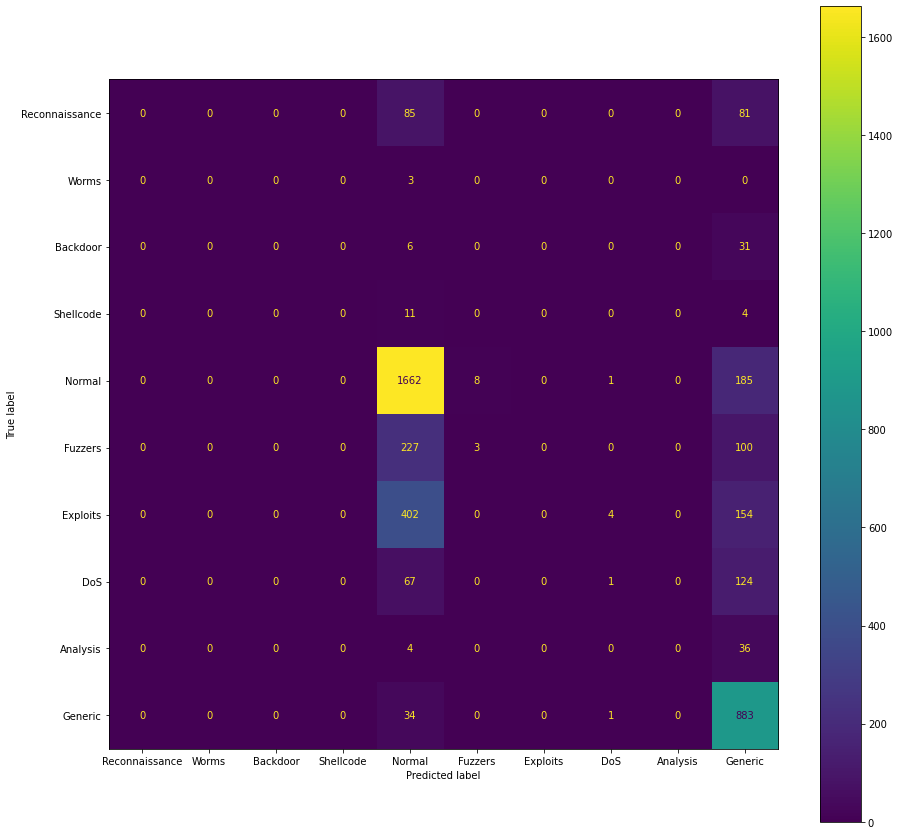
* + - 1. **Scikit-Learn**

For Python, Scikit (Wang and Lu, 2018) is a free and open-source machine learning package. Among the clustering, regression, and classification methods it provides support to various algorithms such as support vector machines, random forests, gradient boosting, k-means, etc. It is also designed to operate with Python's scientific and numerical libraries, NumPy and SciPy. This library is most popular for providing music-related good suggestions on Spotify.

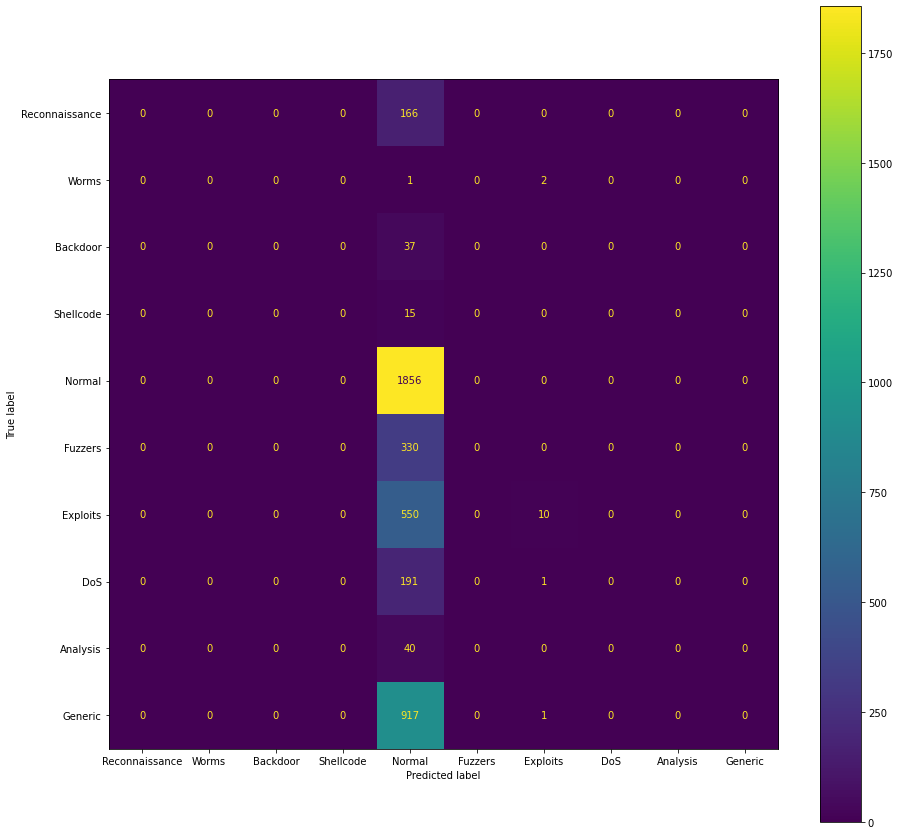
# Evaluation Metrics

Performance evaluation metrics play a vital role to evaluate the efficacy of a model in correctly completing a classification or regression task. These measures simply tell how well the model has performed in terms of various metrics.

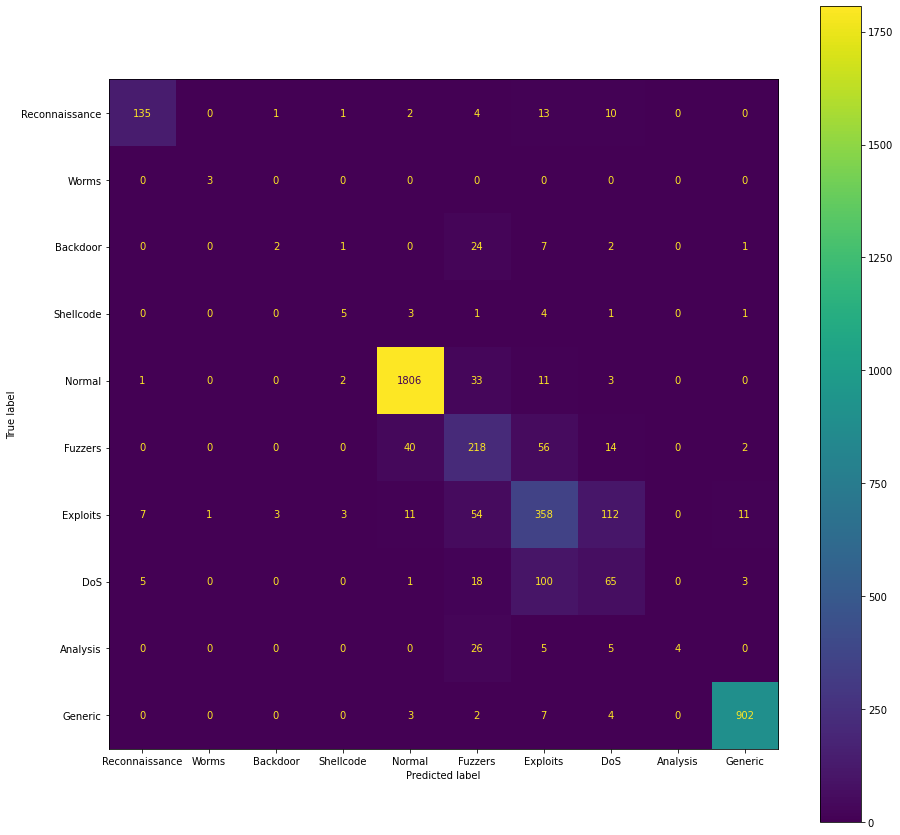
* In logistic regression the accuracy/precision score obtained after execution of the code was 0.6191401. In this model the data set was fit in the axes and the number of iterations which was used to fit the data was scaled according to the data size . The results obtained from this model were terminated at the end as the data was non-convergent.



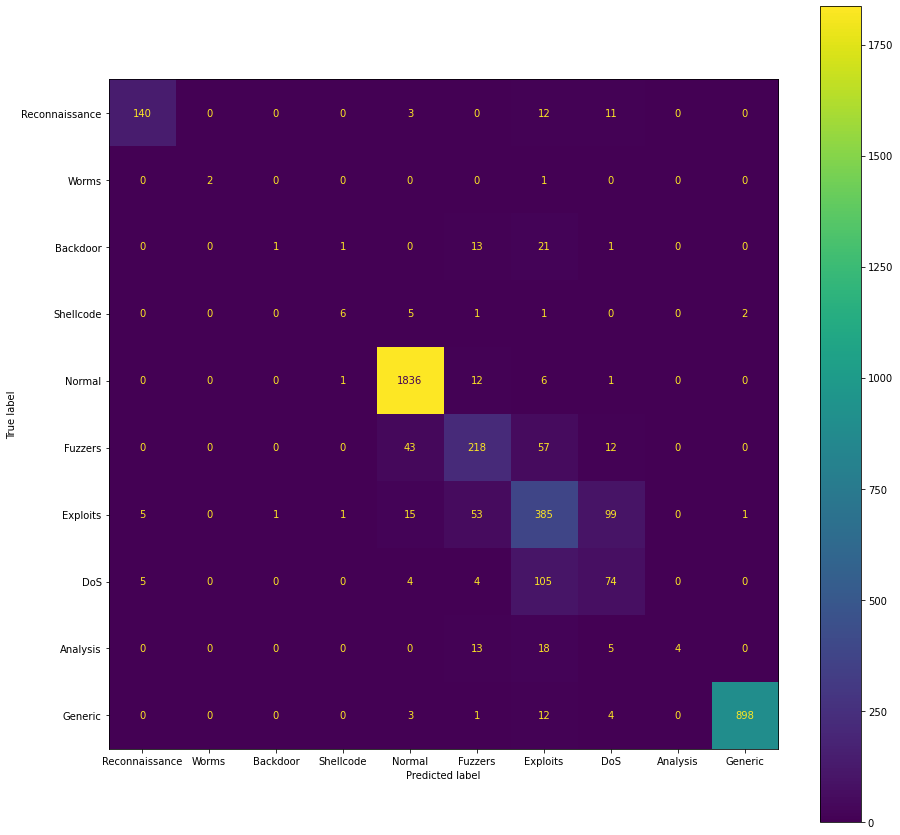
In SVM the obtained accuracy/ precision score of execution of the code was 0.45324265241.

MKL can be considered as a kernel selection method. In the experiments, SVM algorithms are implemented with different kernels and these kernels are combined under MKL.Feature level fusion methods combine the feature spaces instead of the decisions of the classifiers. One of the feature level fusion methods is MKL in which different feature mappings are represented by kernels that are combined to produce a new kernel which represents the samples better than the other kernels. Therefore, MKL provides an approach to solve the feature mapping selection problem of SVM.

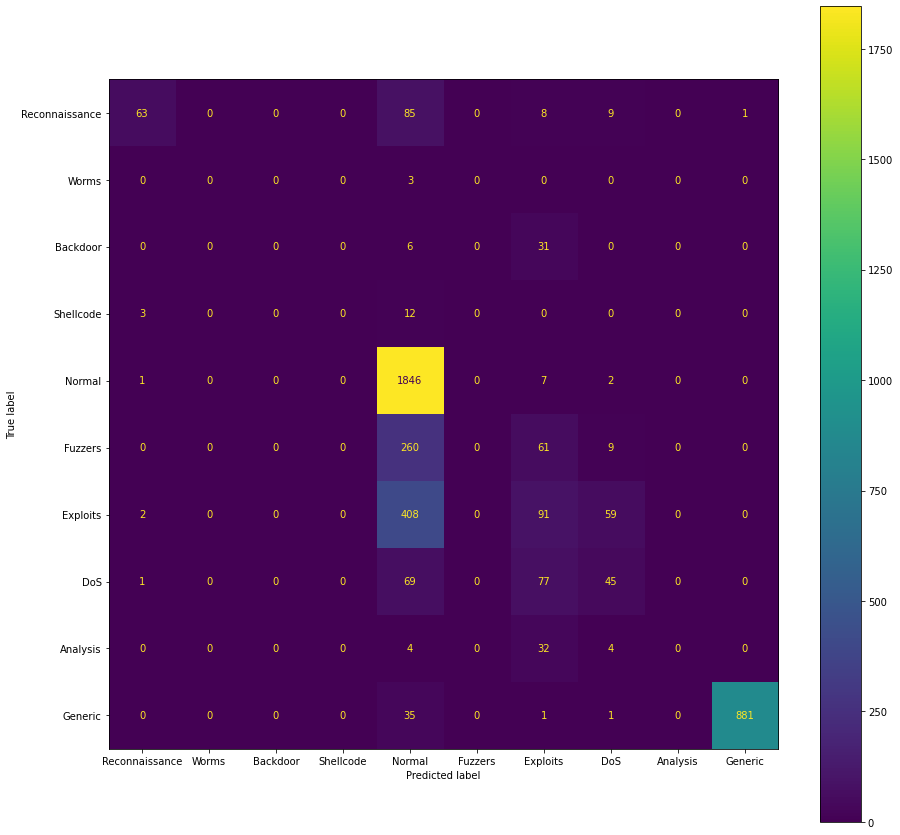
In the Decision Tree Classifier the obtained accuracy/precision score after execution of the code was 0.849647801.

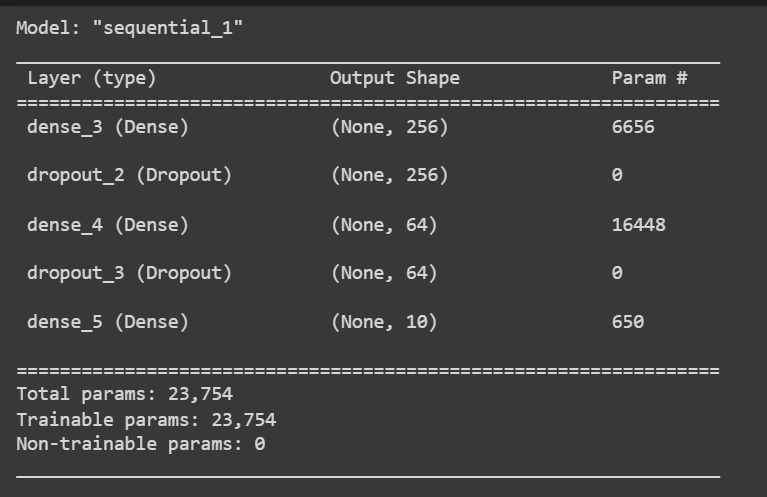


In Random Forest Classifier the obtained accuracy/precision score after execution of the code was 0.865678893.

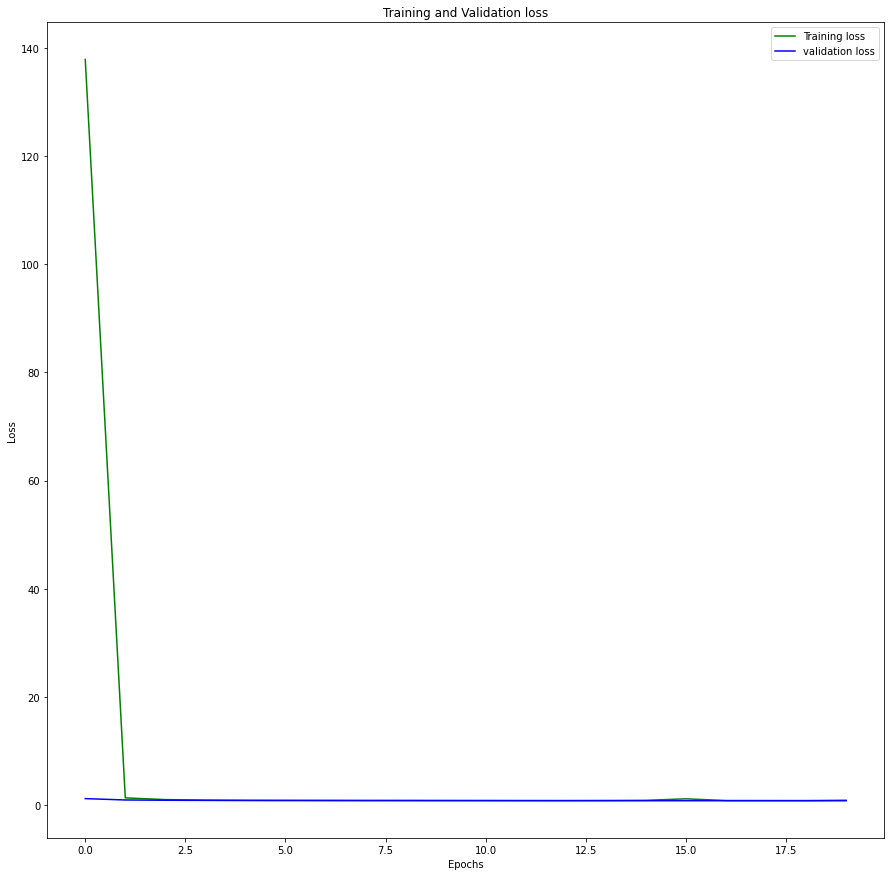


In the Artificial Neural Network(ANN) the obtained accuracy/precision score after execution of the code was 0.710711683.

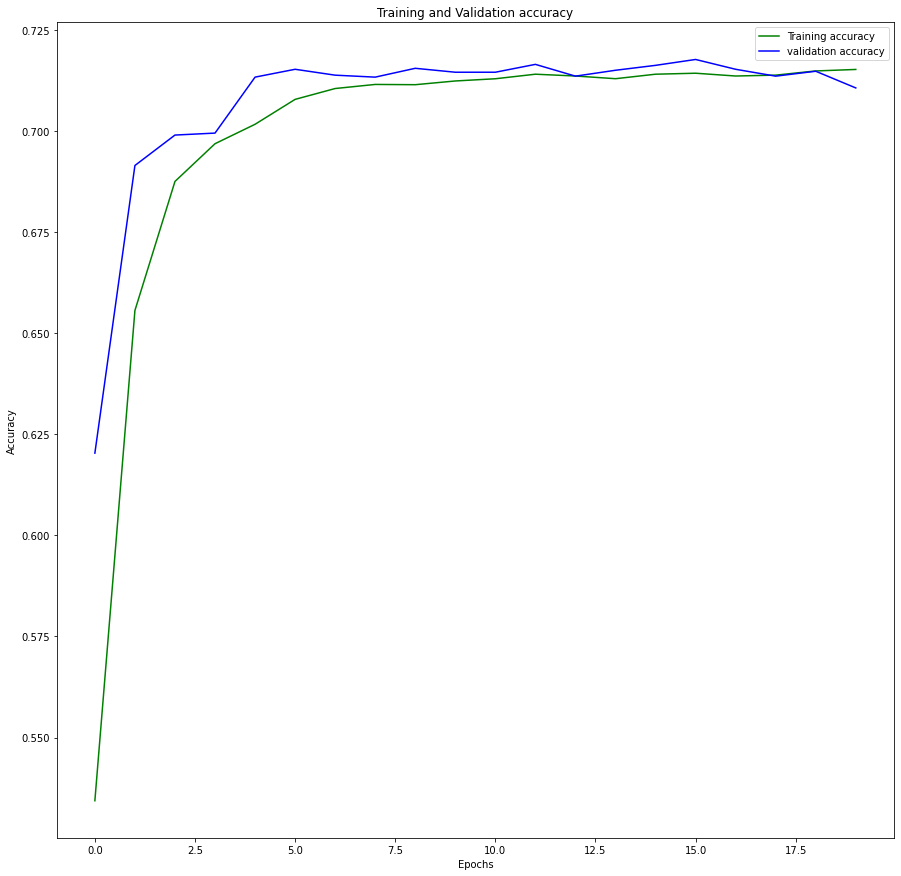
A total of 5 layers were used in the neural network out of which three were dense and the other two were dropouts.



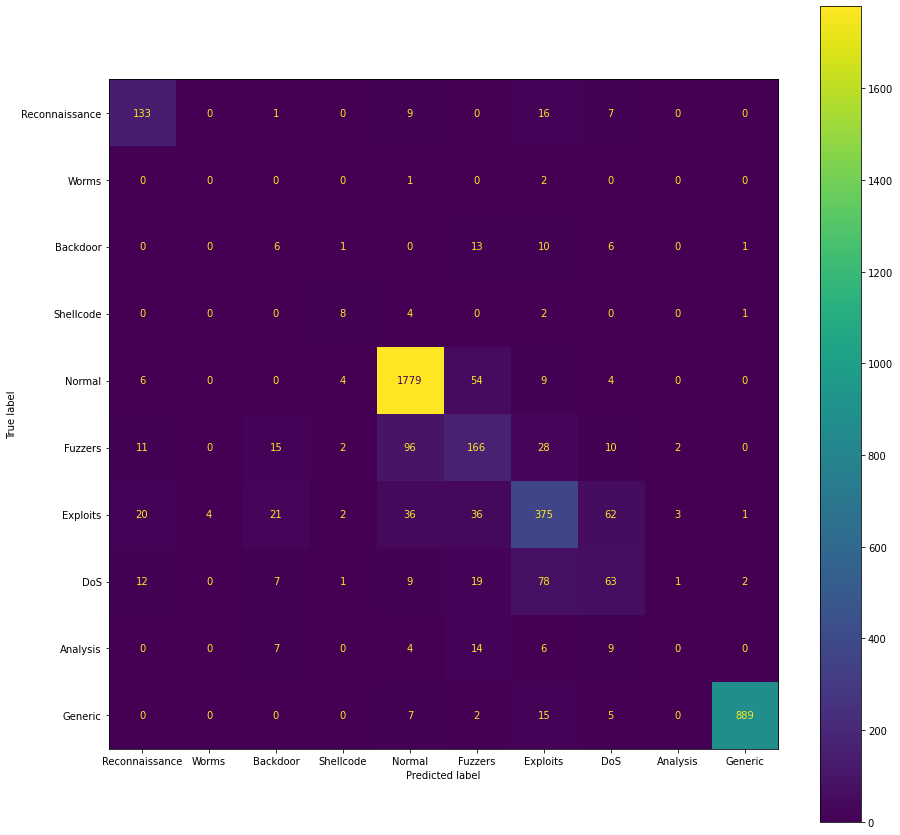
The training and validation loss curves are plotted below



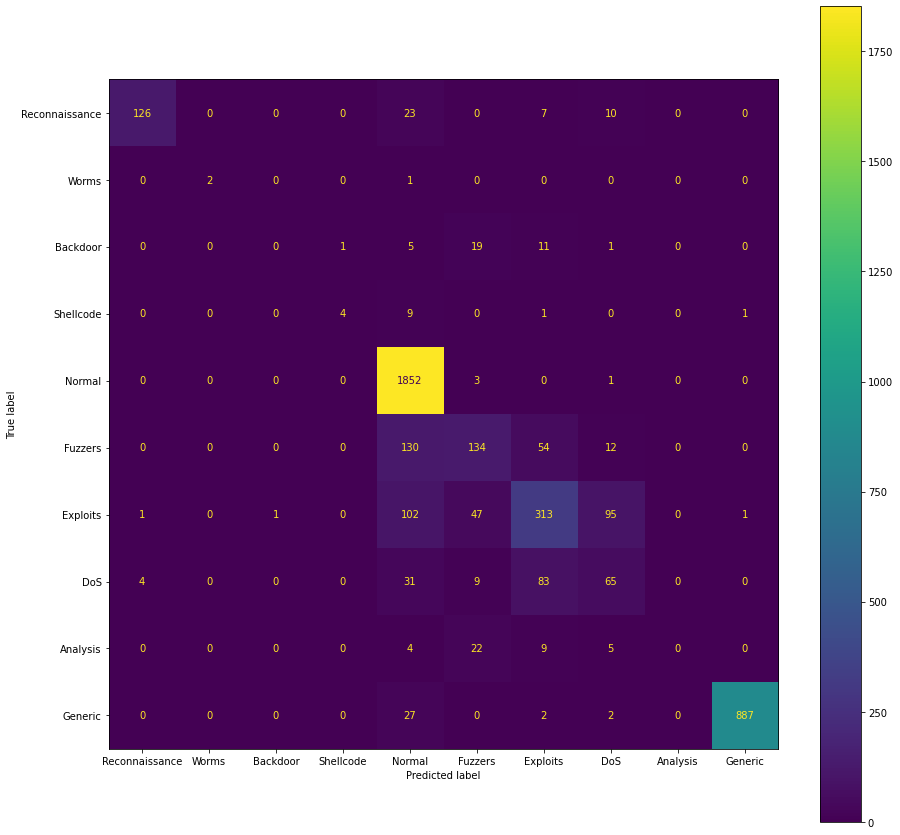
Henceforth the training and validation accuracy is plotted as follows.



In K-Nearest Neighbor (KNN) the obtained accuracy/precision score after execution of the code was 0.834590721. This algorithm labels an unlabeled sample according to the labels of its k-nearest neighborhood in the feature space. Specifically, the observed measurements ∈ S, ∀ i = 1, 2, . . . ,M, are taken as feature vectors. The most frequently observed class label is computed using majority voting among the class labels of the samples in the neighborhood, and assigned as the class label of . The confusion matrix of the model is plotted below.



In the bagging method which was an ensemble technique, the obtained accuracy/precision score after execution was 0.82171484090. In bagging , the models used were logistic regression , random forest , decision tree , SVM and KNN.



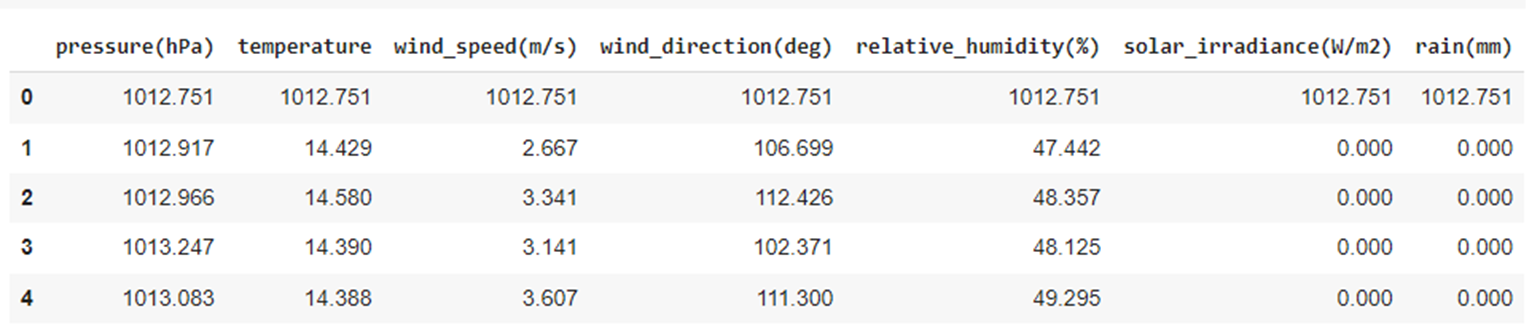
* 1. **Results**

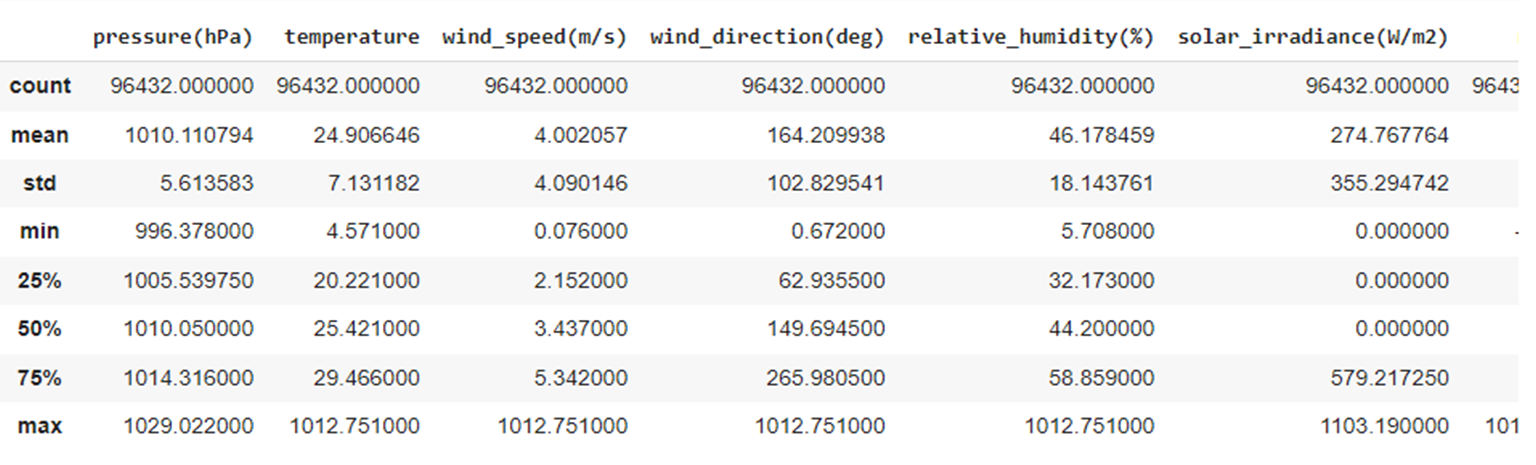
The time computations for the proposed approach based on KNN and other methods to perform correct classification is analyzed graphically by Figure 15. The proposed strategy has taken 12.37 s for the classification process while the time of 13.54 s, 52.53 s, and 17.79 s has been observed with SVM, RF, and LR. The results obtained show the high speed of our method with KNN outperforming other methods respectively.

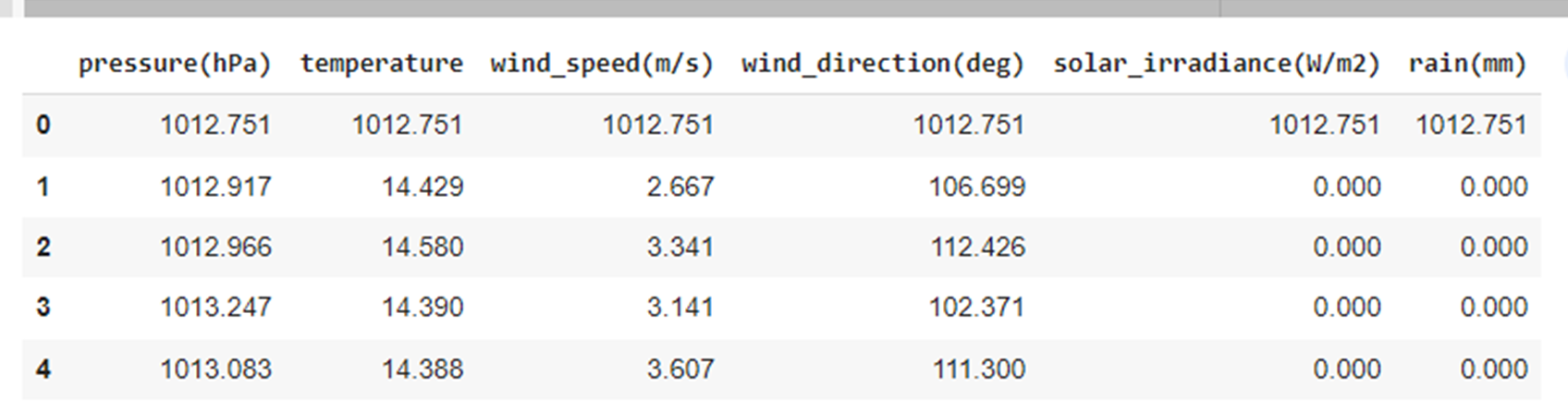
This section provides the output obtained for each model and makes analysis based on the results obtained. All the code snippets are provided in the Appendix section.

* + 1. **Prediction of Natural Disaster Outages**

The prediction tables of the extreme weather conditions and the natural disaster state are tabulated below.The data was split into two sets namely, training set and testing set with 80% and 20% data respectively. The accuracy obtained for the Logistic Regression Model was 0.99964, Decision Tree Classifier was 0.9994815 , Support Vector Machine(SVM) was 0.999222, and the Naive Bayes Classifier was 0.9892155.

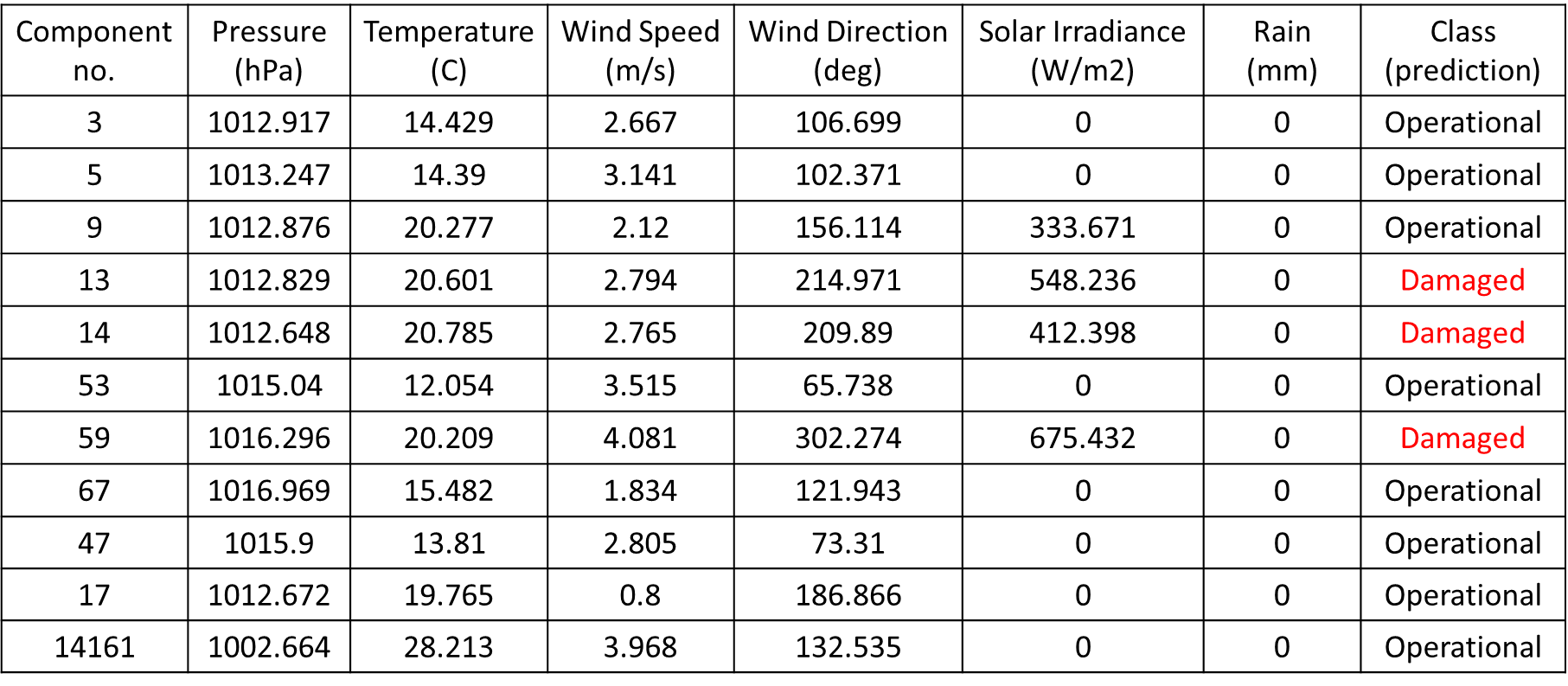
****

****

****

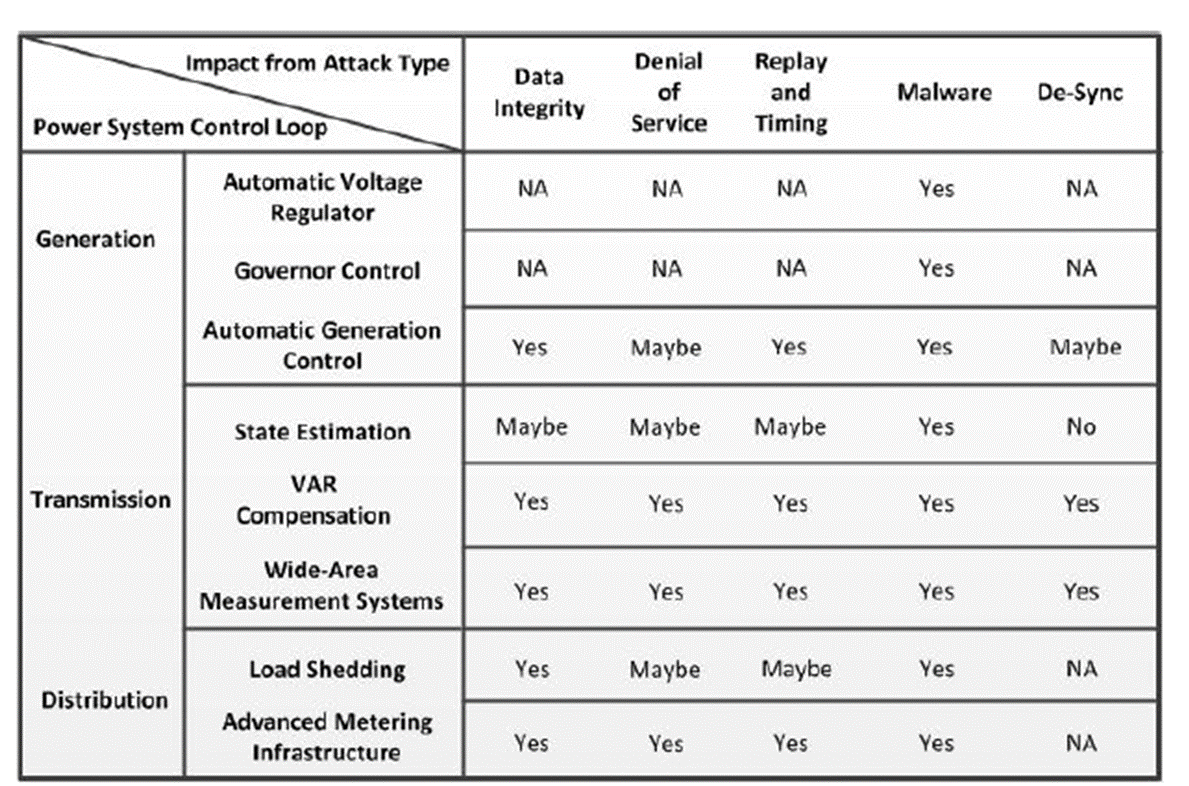
* + 1. **Prediction of Energy Component State**

The energy component states are classified into two categories, namely operational and damaged class. As the name suggests the former type depicts that the fault is operational and the grid is able to be repaired within delineated instructions while the later depicts that it has to be repaired.

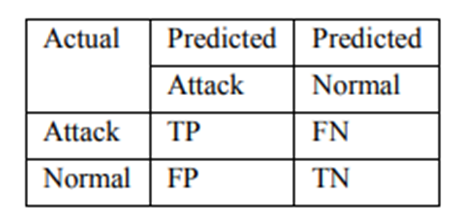


* + 1. **Classification of Attacks**

The attacks are mainly classified into data integrity attacks, denial of service(DoS) attacks, Replay attacks, Timing attacks, and Desynchronization attacks. Below is a tabulated form of the results obtained from the data set.



The evaluation matrix which was used for analyzing the data is given below



For evaluation of the confusion matrix the following formulas were used.

Classification Rate(CR) =

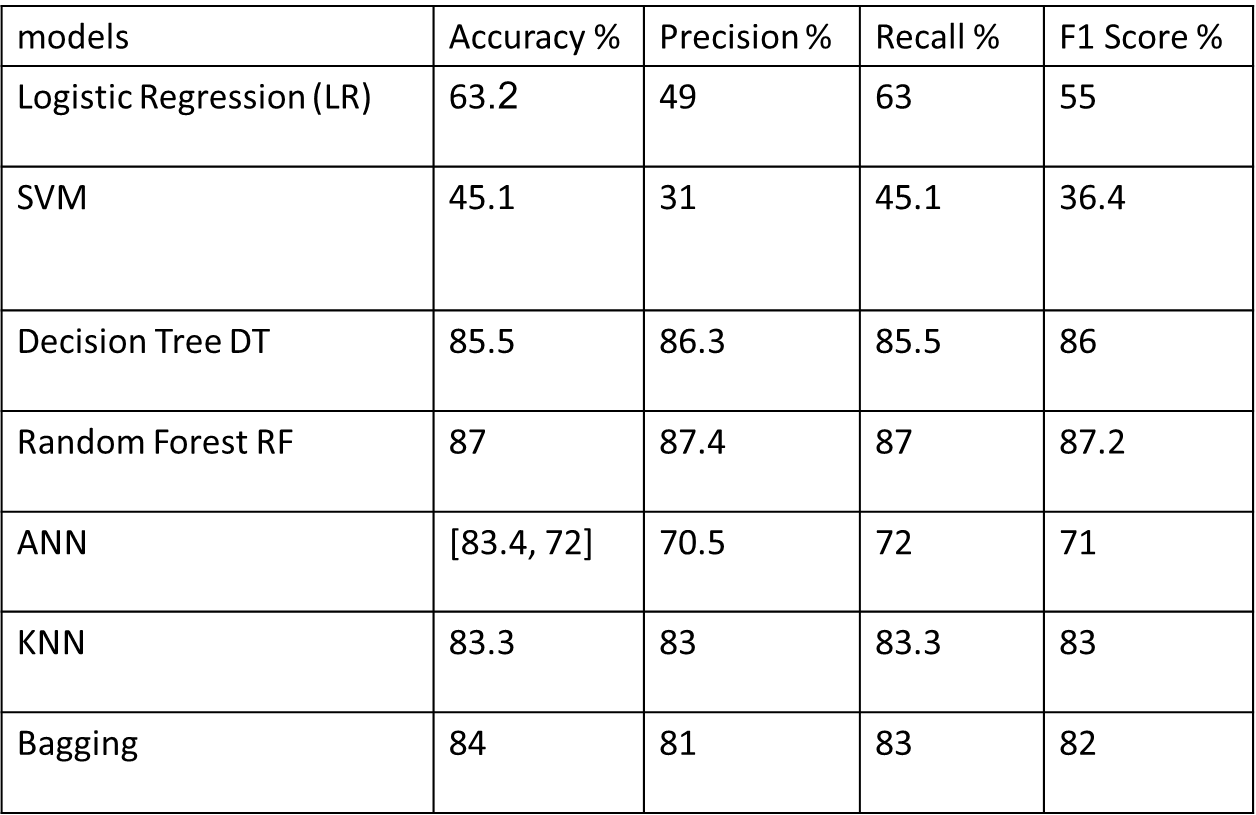
Detection Rate (DR)=

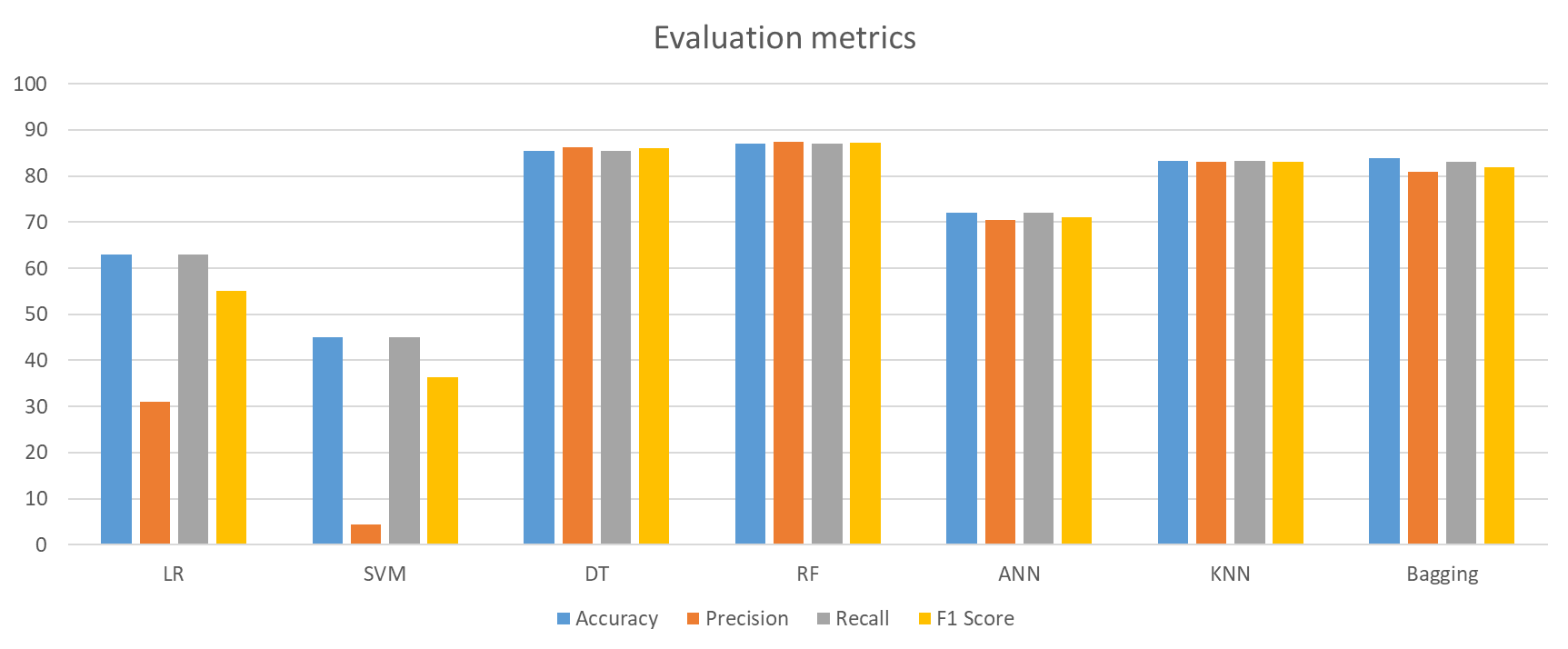
False positive rate(FPR)=

Precision(PR)=

F-measure(FM)=

Using these parameters the accuracy, precision, F1 score and Recall score of each model is tabulated below. The Random Forest Model provides the highest accuracy of 87% . In this case the Random Forest Classifier provided better accuracy than the others because it is able to handle outliers by essentially binning them. It also doesn't care about characteristics that aren't linear. In imbalanced data sets for class populations, it also provides techniques for balancing mistakes. When we have an imbalanced data set, the bigger class will obtain a low error rate while the smaller class will have a greater error rate since random forest strives to minimize the total error rate.





The above graph depicts that the Accuracy of the Decision Tree Classification and the Random Forest Classification is almost similar , but in the tabulated data the accuracy of the RF classifier is more, due to the reason that the Random forest algorithm avoids and prevents overfitting by using multiple trees. The results are not accurate. This gives accurate and precise results. Decision trees require low computation, thus reducing time to implement and carrying low accuracy.

**4.5 Summary**

This chapter briefly provides information about the platform and the tools used for the implementation of machine learning models. The performance evaluation measure used to evaluate the performance of three models i.e. Decision trees, linear regression, and SVM, ANN,KNN Decision Tree Classifier, RF Classifier and Bagging method are discussed. Further, the implementation results for the classification of the attacks and tabulating confusion matrix using each algorithm are discussed in detail one by one for better understanding. All the results on both test and train data are explained graphically with the help of plots for easy visualization and interpretation to make an effective analysis. At last, the performance of all models are evaluated and their accuracy is tabulated, Random Forest Classifier provides the best technique to achieve the purpose of this thesis work with minimum error in comparison to other models.

# CHAPTER 5

# CONCLUSION AND FUTURE WORK

# 5.1 Conclusion

Breakdowns and blackouts have been a major issue for people and governmental authorities since the inception of the conventional power infrastructure. Despite the advancement of traditional power networks, power outages and blackouts continue to be a concern with smart grids. Based on the research, we can infer that our technique might be employed precisely for predicting component status in reaction to an intense event, which is a difficult task in practice. To estimate the likely outage of power system components in reaction to an impending storm, an outage prediction model based on logistic regression was presented. This investigation confirmed the suggested model's acceptable performance.

# REFERENCES

[1] X. Fang, S. Misra, G. Xue, and D. Yang, “Smart grid - The new and improved power grid: A survey,” *IEEE Communications Surveys and Tutorials*. 2012.

[2] H. Jiang, J. J. Zhang, W. Gao, and Z. Wu, “Fault detection, identification, and location in smart grid based on data-driven computational methods,” *IEEE Trans. Smart Grid*, 2014.

[3] “European Context | EIP - Smart Cities and Communities Market Place,” *[Online]. Available:* [*https://eu-smartcities.eu/about/european\_context*](https://eu-smartcities.eu/about/european_context)*. [Accessed: ].* .

[4] M. Mousa, S. Abdelwahed, and J. Kluss, “Review of Fault Types, Impacts, and Management Solutions in Smart Grid Systems,” *Smart Grid Renew. Energy*, 2019.

[5] D. Hodgson, J. L. McDonald, and D. J. Hosken, “What do you mean, ‘resilient’?,” *Trends in Ecology and Evolution*. 2015.

[6] A. M. Madni and S. Jackson, “Towards a conceptual framework for resilience engineering,” *IEEE Eng. Manag. Rev.*, vol. 39, no. 4, 2011.

[7] J. Hare, X. Shi, S. Gupta, and A. Bazzi, “Fault diagnostics in smart micro-grids: A survey,” *Renewable and Sustainable Energy Reviews*, vol. 60. 2016.

[8] J. Liu, W. Zhang, R. Zhou, and J. Zhong, “Impacts of distributed renewable energy generations on smart grid operation and dispatch,” in *IEEE Power and Energy Society General Meeting*, 2012.

[9] F. Katiraei and M. R. Iravani, “Power management strategies for a microgrid with multiple distributed generation units,” *IEEE Trans. Power Syst.*, 2006.

[10] D. S. Popovic and E. E. Boskov, “Advanced fault management as a part of smart grid solution,” in *IET Seminar Digest*, 2008.

[11] N. Bhusal, M. Abdelmalak, M. Kamruzzaman, and M. Benidris, “Power system resilience: Current practices, challenges, and future directions,” *IEEE Access*, 2020.

[12] R. Arghandeh, A. Von Meier, L. Mehrmanesh, and L. Mili, “On the definition of cyber-physical resilience in power systems,” *Renewable and Sustainable Energy Reviews*. 2016.

[13] C. A. Blain *et al.*, “The Network Coordination Office of NHERI (Natural Hazards Engineering Research Infrastructure),” *Front. Built Environ.*, 2020.

[14] Cabinet Office, “Keeping the Country Running: Natural Hazards and Infrastructure,” *Environment*, 2011.

[15] F. H. Jufri, J. S. Kim, and J. Jung, “Analysis of determinants of the impact and the grid capability to evaluate and improve grid resilience from extreme weather event,” *Energies*, vol. 10, no. 11. 2017.

[16] M. Bruneau *et al.*, “A Framework to Quantitatively Assess and Enhance the Seismic Resilience of Communities,” *Earthquake Spectra*, vol. 19, no. 4. 2003.

[17] M. Panteli, D. N. Trakas, P. Mancarella, and N. D. Hatziargyriou, “Power Systems Resilience Assessment: Hardening and Smart Operational Enhancement Strategies,” *Proc. IEEE*, vol. 105, no. 7, 2017.

[18] M. Panteli, P. Mancarella, D. N. Trakas, E. Kyriakides, and N. D. Hatziargyriou, “Metrics and Quantification of Operational and Infrastructure Resilience in Power Systems,” *IEEE Trans. Power Syst.*, vol. 32, no. 6, 2017.

[19] M. Panteli and P. Mancarella, “The Grid: Stronger, Bigger, Smarter?,” *IEEE Power Energy Mag.*, no. May/June, 2015.

[20] Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. 2012.

[21] NIPP DHS, “Partnering for Critical Infrastructure Security and Resilience. Homeland Security,” *Dhs*, 2013.

[22] UN (United Nations), “Hyogo Framework for Action (HFA) 2005-2015 Building the Resilience of Nations and Communities to Disasters,” *United Nations*, no. January, 2005.

[23] M. Chaudry *et al.*, “Building a Resilient UK Energy System Research Report,” *Carbon N. Y.*, no. April, 2011.

[24] J. Jacobs, C. Kumar, H. Sarmiento, D. Sharafi, and A. Srivastava, “Power system operational resilience – what it means and where we stand,” *CIGRE Sci. Eng.*, vol. 18, no. June, 2020.

[25] S. Afzal, H. Mokhlis, H. Azil Lllias, N. Nadzirah Mansor, and H. Shareef, “State-of-the-art review on power system resilience and assessment techniques,” *IET Gener. Transm. Distrib.*, vol. 14, no. 25, 2020.

[26] M. Panteli and P. Mancarella, “Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies,” *Electric Power Systems Research*, vol. 127. 2015.

[27] Y. Wang, C. Chen, J. Wang, and R. Baldick, “Research on Resilience of Power Systems under Natural Disasters - A Review,” *IEEE Transactions on Power Systems*, vol. 31, no. 2. 2016.

[28] Q. Xie and R. Zhu, “Earth, wind, and ice,” *IEEE Power Energy Mag.*, vol. 9, no. 2, 2011.

[29] J. C. Araneda, H. Rudnick, S. Mocarquer, and P. Miquel, “Lessons from the 2010 Chilean earthquake and its impact on electricity supply,” in *2010 International Conference on Power System Technology: Technological Innovations Making Power Grid Smarter, POWERCON2010*, 2010.

[30] Y. Xu, C. C. Liu, K. P. Schneider, and D. T. Ton, “Toward a resilient distribution system,” in *IEEE Power and Energy Society General Meeting*, 2015, vol. 2015-September.

[31] N. C. Abi-Samra and W. P. Malcolm, “Extreme weather effects on power systems,” in *IEEE Power and Energy Society General Meeting*, 2011.

[32] N. C. Abi-Samra, K. R. Forsten, and R. Entriken, “Sample effects of extreme weather on power systems and components, part I: Sample effects on distribution systems,” in *IEEE PES General Meeting, PES 2010*, 2010.

[33] A. Castillo, “Risk analysis and management in power outage and restoration: A literature survey,” *Electr. Power Syst. Res.*, vol. 107, 2014.

[34] M. Panteli and P. Mancarella, “Modeling and evaluating the resilience of critical electrical power infrastructure to extreme weather events,” *IEEE Syst. J.*, vol. 11, no. 3, 2017.

[35] D. Yates *et al.*, “Stormy weather: Assessing climate change hazards to electric power infrastructure: A sandy case study,” *IEEE Power Energy Mag.*, vol. 12, no. 5, 2014.

[36] U. Nakarmi, M. Rahnamay-Naeini, and H. Khamfroush, “Critical Component Analysis in Cascading Failures for Power Grids Using Community Structures in Interaction Graphs,” *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 3, 2020.

[37] N. Singh and A. Gupta, “A study on mitigation of blackout risks in smart grid,” in *India International Conference on Power Electronics, IICPE*, 2016, vol. 2016-November.

[38] C. Chen, J. Wang, and D. Ton, “Modernizing Distribution System Restoration to Achieve Grid Resiliency Against Extreme Weather Events: An Integrated Solution,” *Proc. IEEE*, vol. 105, no. 7, 2017.

[39] P. Korada and A. R. Devidas, “Studying the impact of AC-microgrid on the main grid and it’s fault analysis,” in *2016 - Biennial International Conference on Power and Energy Systems: Towards Sustainable Energy, PESTSE 2016*, 2016.

[40] M. Mousa, S. Abdelwahed, and J. Kluss, “Impact of resistive superconductive fault current limiter’s location and its resistance value on the stability of micro grid system,” in *2019 IEEE Texas Power and Energy Conference, TPEC 2019*, 2019.

[41] W. Zhang, X. Xiao, K. Zhou, W. Xu, and Y. Jing, “Multicycle Incipient Fault Detection and Location for Medium Voltage Underground Cable,” *IEEE Trans. Power Deliv.*, 2017.

[42] H. C. Dubey, S. R. Mohanty, and N. Kishore, “Abrupt change detection of fault in power system using independent component analysis,” in *2011 - International Conference on Signal Processing, Communication, Computing and Networking Technologies, ICSCCN-2011*, 2011.

[43] C. Kim, “Detection and location of intermittent faults by monitoring carrier signal channel behavior of electrical interconnection system,” in *IEEE Electric Ship Technologies Symposium, ESTS 2009*, 2009.

[44] C. A. Andresen, B. N. Torsaeter, H. Haugdal, and K. Uhlen, “Fault Detection and Prediction in Smart Grids,” in *9th IEEE International Workshop on Applied Measurements for Power Systems, AMPS 2018 - Proceedings*, 2018.

[45] Enhancing the Resilience of the Nation’s Electricity System. 2017.

[46] Vastag, “Nuclear power plant remains offline after August earthquake,” *The Washington Post*.

[47] Elevation data for floodplain mapping. 2007.

[48] R. J. Campbell, “Weather-related power outages and electric system resiliency,” in *The Transformation of Electrical Power: Key Issues*, 2013.

[49] R. Eskandarpour and A. Khodaei, “Machine Learning Based Power Grid Outage Prediction in Response to Extreme Events,” *IEEE Trans. Power Syst.*, vol. 32, no. 4, 2017.

[50] S. I. Seneviratne *et al.*, “Changes in climate extremes and their impacts on the natural physical environment,” in *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change*, 2012.

[51] W. V. Sweet *et al.*, “Global and Regional Sea Level Rise Scenarios for the United States,” *NOAA Tech. Rep. NOS CO-OPS 083*, 2017.

[52] “Economic benefits of increasing electric grid resilience to weather outages,” in *Climate, Energy, and Environment: Issues, Analyses, and Developments*, vol. 2, 2014.

[53] P. Hines, J. Apt, and S. Talukdar, “Trends in the history of large blackouts in the United States,” in *IEEE Power and Energy Society 2008 General Meeting: Conversion and Delivery of Electrical Energy in the 21st Century, PES*, 2008.

[54] N. Wang, V. Aravinthan, and Y. Ding, “Feeder-level fault detection and classification with multiple sensors: A smart grid scenario,” in *IEEE Workshop on Statistical Signal Processing Proceedings*, 2014.

[55] A. M. Pasdar, Y. Sozer, and I. Husain, “Detecting and locating faulty nodes in smart grids based on high frequency signal injection,” *IEEE Trans. Smart Grid*, vol. 4, no. 2, 2013.

[56] J. Cordova and M. O. Faruque, “Fault location identification in smart distribution networks with Distributed Generation,” in *2015 North American Power Symposium, NAPS 2015*, 2015.

[57] IEEE, “IEEE Recommended Practice for 1 kV to 35 kV Medium-Voltage DC Power Systems on Ships; IEEE Std 1709-2010,” 2010.

**APPENDIX**

# Appendix-1:- Dataset Details

# This subsection provides the details regarding the dataset used in this research work. Some of the example data is presented as in the following:

# Using UNSW\_NB15 Dataset

**UNSW-NB15** is a network intrusion dataset. It contains nine different attacks, including DoS, worms, Backdoors, and Fuzzers. The dataset contains raw network packets. The number of records in the training set is 175,341 records and the testing set is 82,332 records from the different types, attack and normal.

# 

# 

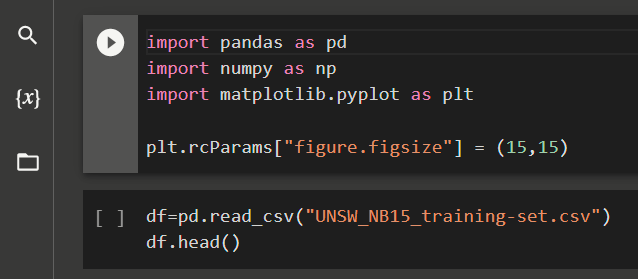
# 

# 

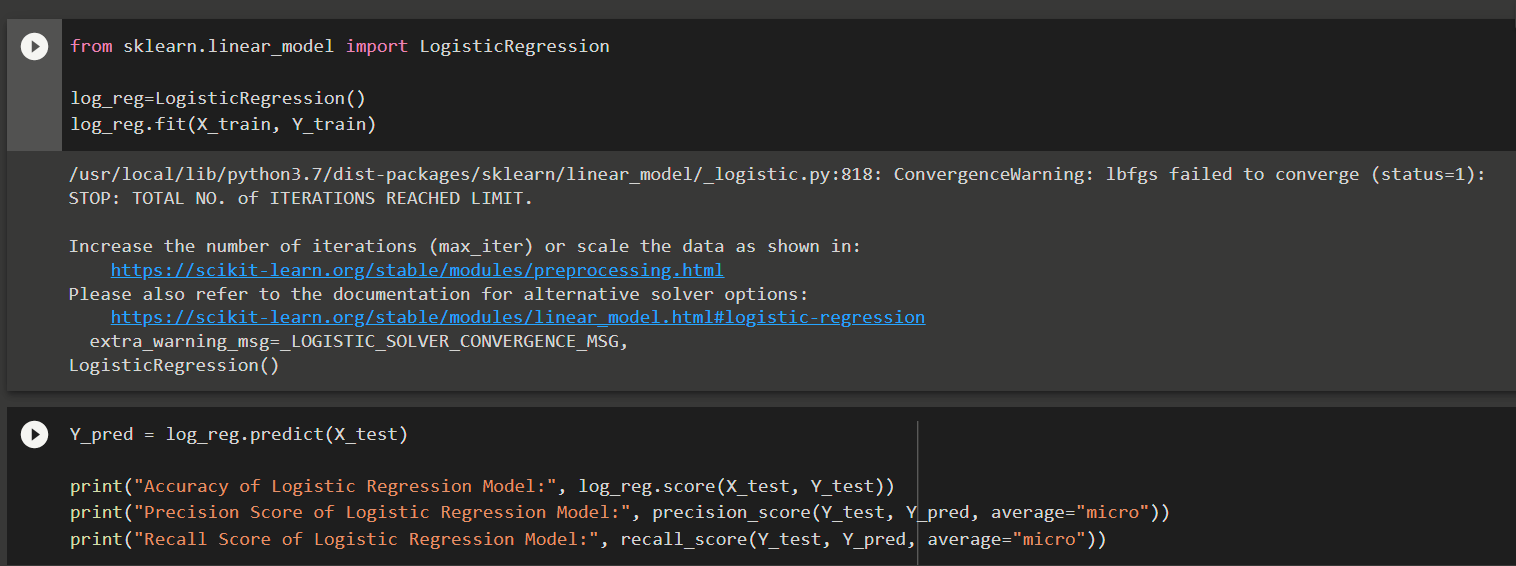
# Appendix-2:- Code Snippets

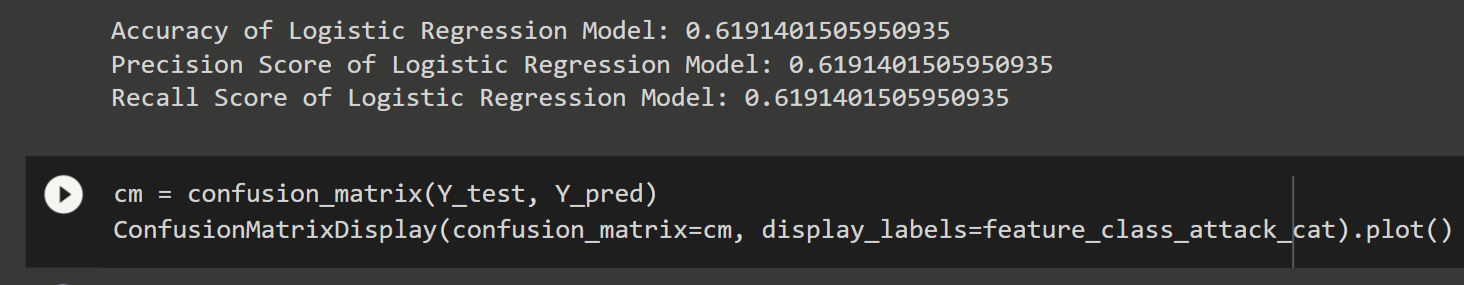
This subsection provides code snippets representing the implementation of the used machine learning and ensemble learning models and to produce the respective results and graphs.

* **For importing data set**

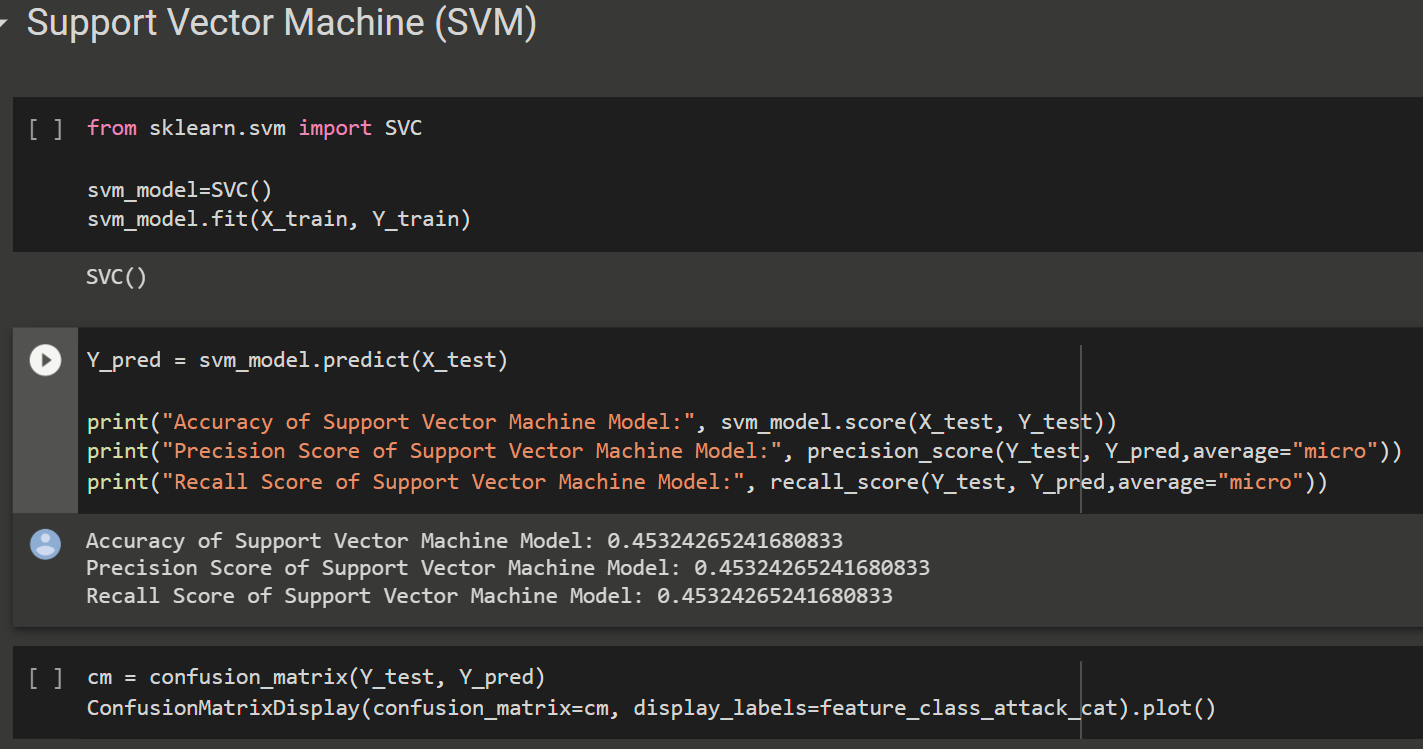


* **Logistic Regression (LR)**

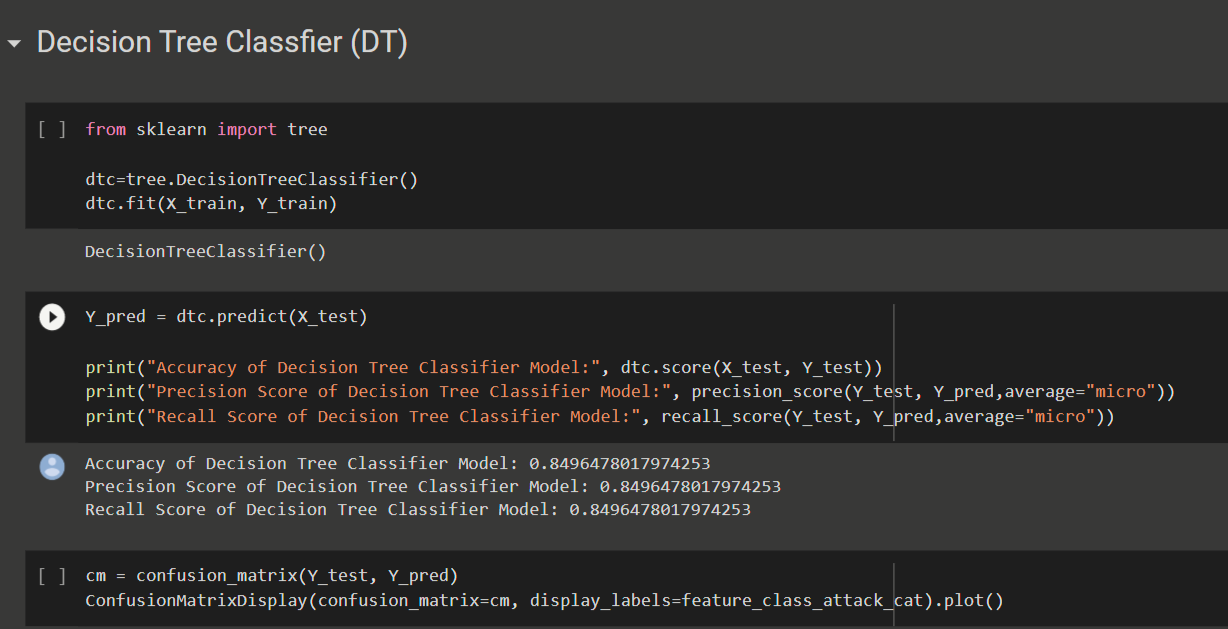




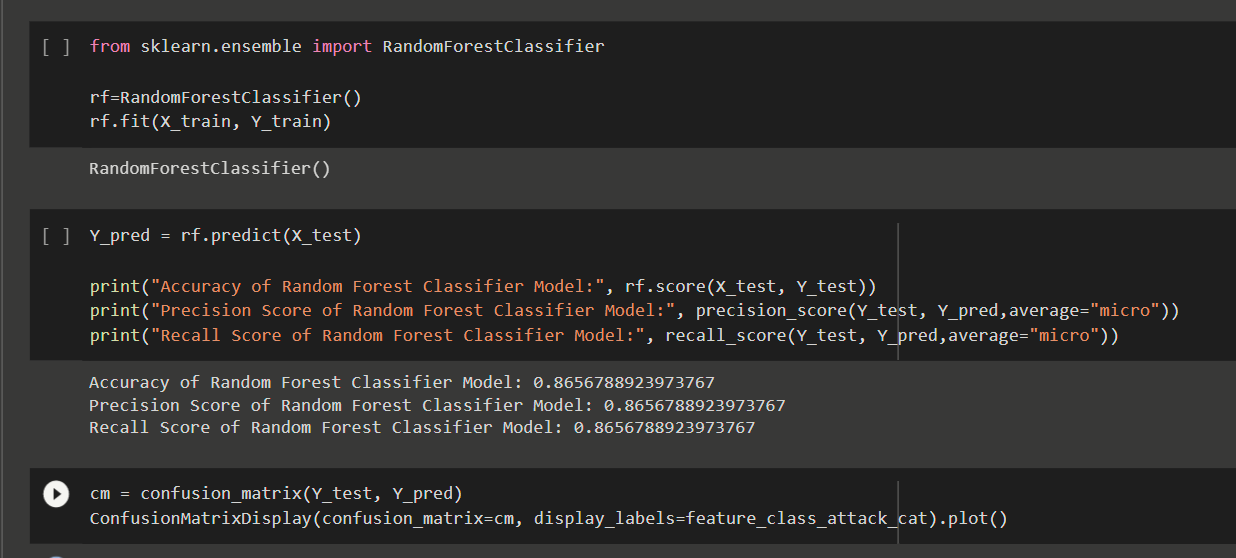
# Support Vector Machine (SVM)



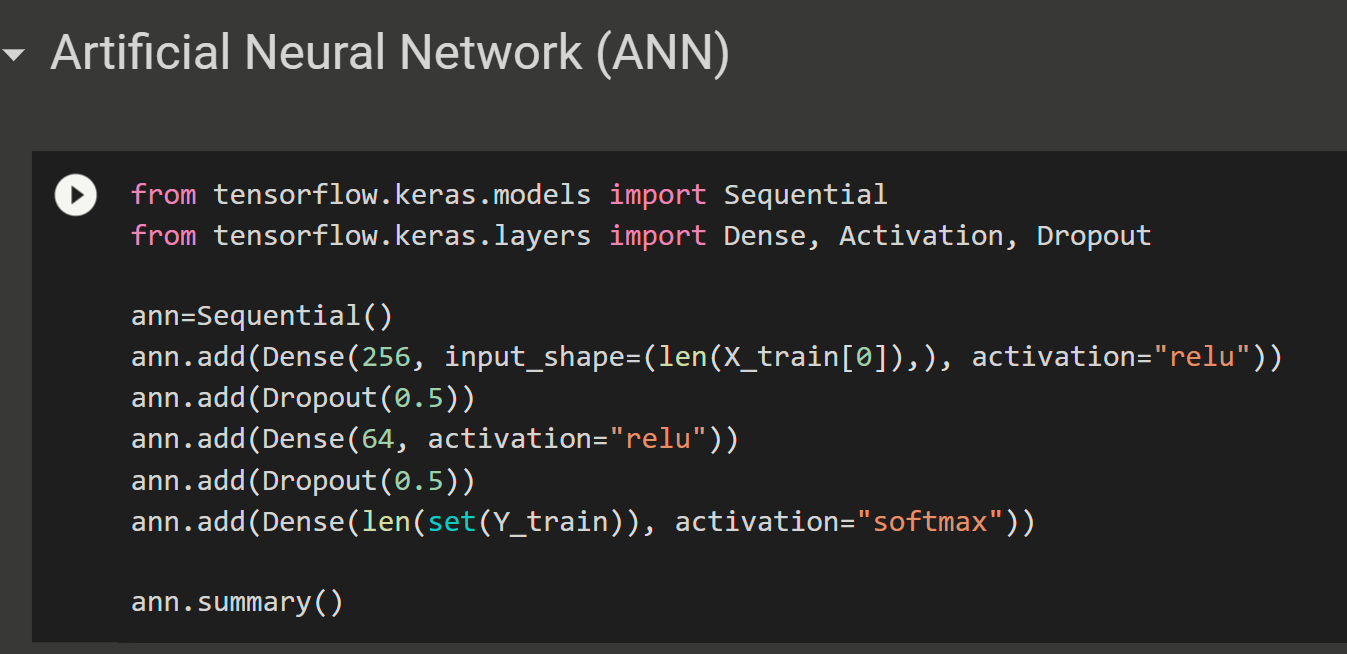
# Decision Tree (DT)

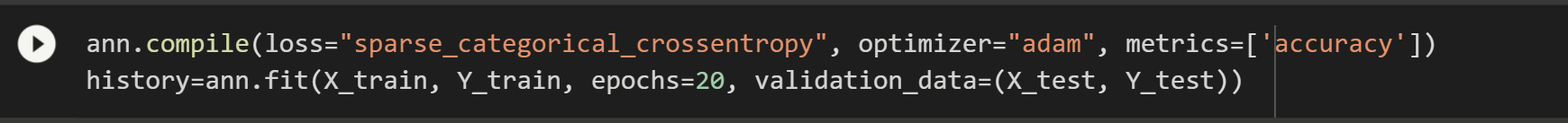


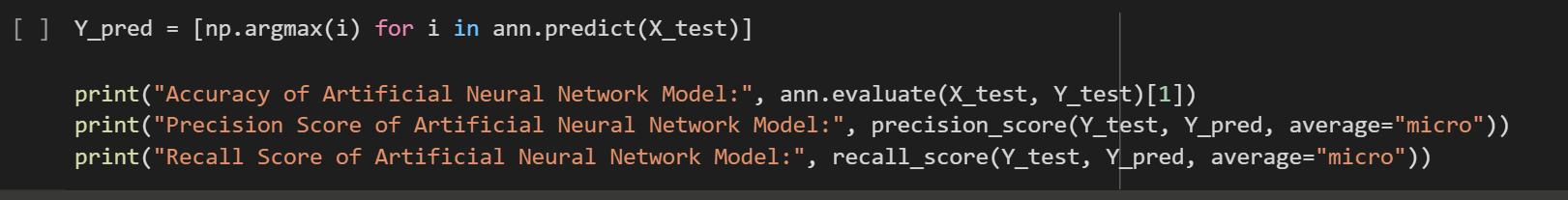
* **Random Forest(RF)**

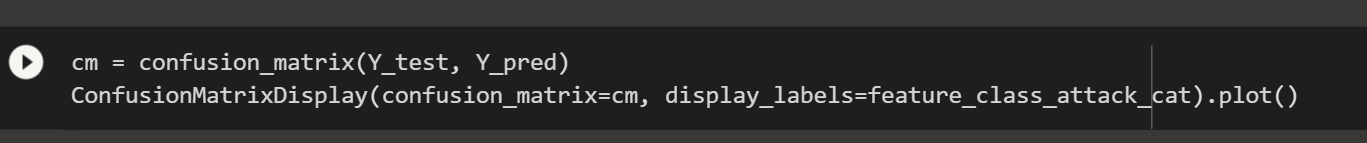


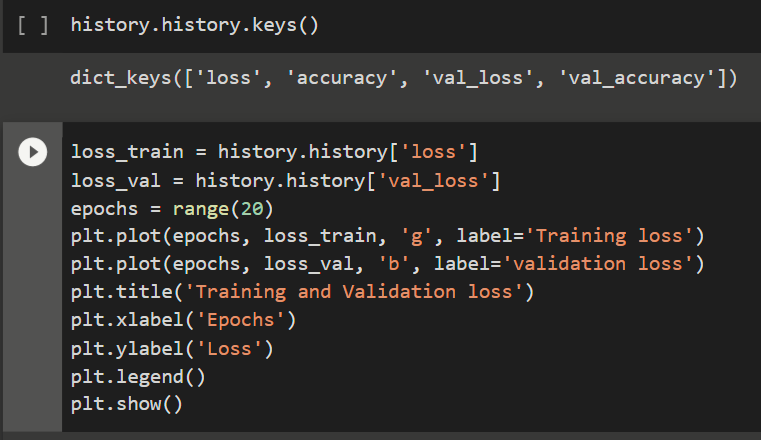
* **Artificial Neural Network**

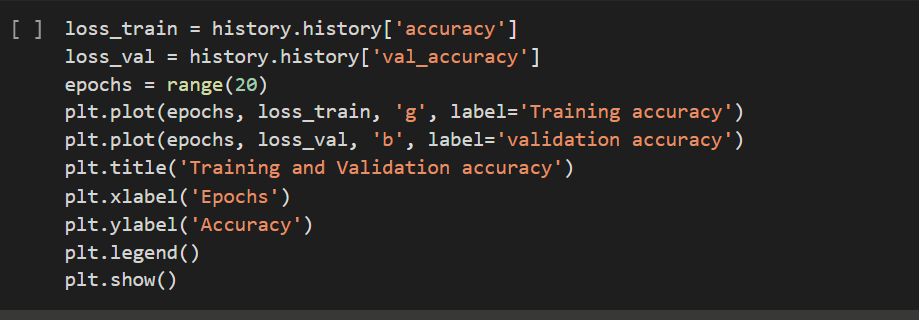




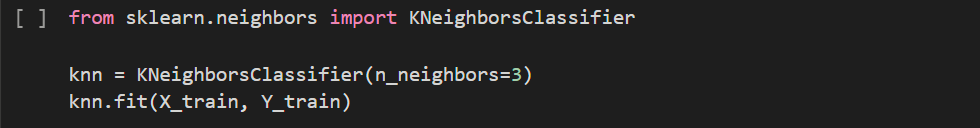


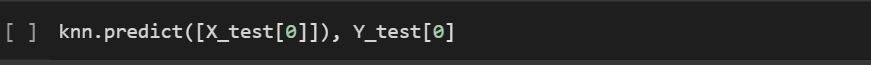


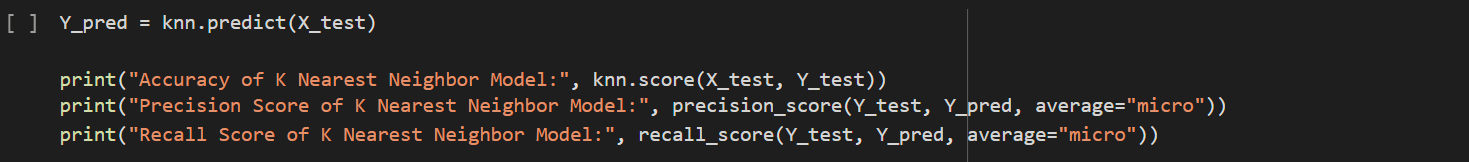


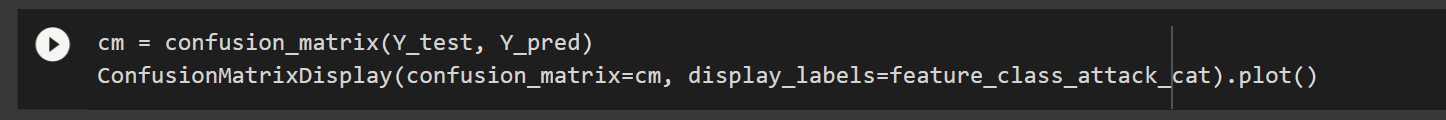


* **K-Nearest Neighbor(KNN)**









* **Bagging (Ensemble Method)**

