PERFORMANCE ASSESSMENT OF MACHINE LEARNING MODELS FOR TRANSMISSION NETWORK FAULT DIAGNOSIS AMIDST ONGOING RENEWABLE ENERGY SOURCE INTEGRATIONS

Thesis submitted in partial fulfillment

for the award of degree

Doctor of Philosophy

by

RACHNA VAISH

RAJIV GANDHI INSTITUTE OF PETROLEUM TECHNOLOGY

JAIS – 229304

Roll No.: PEE19001 **2024**

CERTIFICATE

It is certified that the work contained in the thesis titled ***"PERFORMANCE ASSESSMENT OF MACHINE LEARNING MODELS FOR TRANSMISSION NETWORK FAULT DIAGNOSIS AMIDST ONGOING RENEWABLE ENERGY SOURCE INTEGRATIONS"*** by ***Rachna Vaish*** has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive, Candidacy, SOTA, and Open Seminar.

Dr. U. D. Dwivedi

(Thesis Supervisor)

DECLARATION BY THE CANDIDATE

I, ***Rachna Vaish***, certify that the work embodied in this thesis is my own bona fide work and was carried out by me under the supervision of ***Dr. U. D. Dwivedi*** from ***July 2019*** **to** ***July 2024***, at the ***Department of Electrical & Electronics Engineering***, Rajiv Gandhi Institute of Petroleum Technology, Jais. The matter embodied in this thesis has not been submitted for the award of any other degree. I declare that I have faithfully acknowledged and given credit to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports, dissertations, theses, etc., or available at websites and have not included them in this thesis and have not cited as my own work.

Date:

Place: Jais, Amethi **Rachna Vaish**

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of my knowledge.

Dr. U. D. Dwivedi
(Thesis Supervisor)

Head of the Department
(Electrical & Electronics Engineering)

CERTIFICATE

CERTIFIED that the work contained in the thesis titled "***Performance Assessment of Machine Learning Models for Transmission Network Fault Diagnosis Amidst Ongoing Renewable Energy Source Integrations"*** by ***Mrs. Rachna Vaish*** has been carried out under my supervision. It is also certified that she fulfilled the mandatory requirement of TWO quality publications that arose out of her thesis work.

It is further certified that the two publications (copies enclosed) of the aforesaid ***Mrs. Rachna Vaish*** have been published in the Journals indexed by –

(a) SCI

(b) SCI Extended

(c) SCOPUS

(d) *Non-indexed journals

(only in special cases)

(*Please enclose DPGC resolution in this regard)

Dr. U. D. Dwivedi
(Thesis Supervisor)

Dr. Shivanshu Srivastava
(Convener, DPGC)

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: Performance Assessment of Machine Learning Models for Transmission Network Fault Diagnosis Amidst Ongoing Renewable Energy Source Integrations.

Name of the Student: Rachna Vaish

Copyright Transfer

The undersigned hereby assigns to the Rajiv Gandhi Institute of Petroleum Technology Jais all rights under copyright that may exist in and for the above thesis submitted for the award of the *"DOCTOR OF PHILOSOPHY"*.

Date:
Place: Jais, Amethi

Rachna Vaish
Roll No.: PEE19-001

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for the author's personal use, provided that the source and the Institute's copyright notice are indicated.

ACKNOWLEDGEMENT

Before commencing this thesis, I would like to profoundly thank individuals who have paved the path for my PhD journey with their unwavering support and encouragement. This expedition, with its highs and lows akin to a roller coaster, has been made manageable by the invaluable contributions of numerous individuals, both directly and indirectly involved. First and foremost, I extend my sincerest appreciation to my thesis supervisor, **Dr. Umakant Dhar Dwivedi**, whose guidance, encouragement, and persistent support have been instrumental in realizing this dissertation. Throughout my doctoral pursuit, Dr. Dwivedi has provided invaluable insights, constructive suggestions, and meticulous feedback, consistently upholding the highest academic rigor and technical proficiency standards essential for scholarly publication.

Additionally, I am deeply indebted to **Prof. A.S.K. Sinha**, Director of RGIPT, whose unwavering commitment to fostering a culture of excellence in research has inspired me. My heartfelt thanks also extend to **Dr. Shivanshu Srivastava**, DPGC Convenor, Dr. Abhishek Kumar Singh, and Dr. Bheemaiah Chikondra for their unwavering support, well wishes, understanding, and motivation. I want to extend my thanks to all the faculty members and RPEC members for their suggestions and to the administrative staff of the Department of Electrical & Electronics Engineering for their help and support.

My parents are equally deserving of my appreciation, whose boundless love, unwavering support, and steadfast belief in my abilities have been a constant source of strength. Furthermore, I am deeply grateful to my in-laws, sister, brothers, and beloved husband, Mr. Chitranshu Kesharwani, for their unwavering support, encouragement, and sacrifices that have enabled me to pursue my academic aspirations. Lastly, I extend my heartfelt thanks to my friends, Gargi, Vinamra, Yogendra, Shadab, and other fellow research scholars, Mukesh, Tanya, Amit, Mukuljeet, and my juniors within the department, whose unwavering support, camaraderie, and cherished memories have been a source of solace and inspiration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS/NOTATIONS

ACNN	Adaptive convolutional neural network
AIRS	Auto immune recognition system
API	Application programming interface
ART	Adaptive Resonant Theory
Bi-GRU	Bidirectional gated recurrent neural network
BPNN	Back-propagation neural network
BRR	Bayesian ridge regression
CART	Classification and regression tree
CCR	Correct Classification Rate
CNN	Convolutional neural network
db	Daubechies
DBF	Deep Belief Network
DER	Distributed energy resource
DG	Distributed generation
DNN	Deep neural network
DT	Decision Tree
DTNB	Decision table Naïve Bayesian
DWT	Discrete wavelet transform
EKF	Extended kalman filter
ELM	Extreme learning machine
ET	Extra Tree
FCNN	Fully connected neural network
FD	Fault data
FDOST	Fast discrete orthogonal S-transform
FDR	Fault data recorder
FIA	Fault inception angle
FL	Fuzzy Logic
FT	Fourier Transform
GA	Genetic algorithm
GCN	Graph convolutional network
Gen	Generator
GPR	Gaussian process regression
GPS	Global positioning system
HHT	Hilbert Huang transform
HIF	High impedance fault
HLCL	Human level concept learning
HMM	Hidden Markov model
HVDC	High voltage direct current

IEEE	Institute of electrical and electronics engineers
KLD	Kullback-Leibler Divergence
KNN	Know-nearest neighbors
kPCA	kernel principal component analysis
LDA	Linear discriminant analysis
LL	Line to line fault
LLG	Line to line to ground fault
LLL	Line to line to line fault
LLLG	Line to line to line to ground fault
LM	Levenberg Marquardt
LR	Logistic regression
MAP	Maximum posterior probability
ML	Machine learning
MLP	Multilayer Perceptron
MPPT	Maximum power point tracking
NB	Naïve Bayesian
NN	Neural network
NODR	Without dimensionality reduction
PCA	Principal component analysis
PDCs	Phasor data concentrators
PLC	Programmable logic controller
PMUs	Phasor measurement units
PNN	Probabilistic neural network
PV	Photovoltaic
QSSVM	Quarter sphere support vector machine
RBF	Radial Basis Function
RES	Renewable energy sources
RF	Random forest
RFR	Random forest regressor
RNN	Recurrent neural network
RT	Regression tree
RTU	Remote terminal unit
RVFL	Random vector functional link
RVM	Relevance vector machine
SCADA	Supervisory control and data acquisition
SGD	Stochastic gradient descent
SLG	Single line to ground fault
SSAE	Stacked sparse autoencoder
SVDD	Support Vector Data Description
SVM	Support vector machine
SVR	Support vector regression

SWT	Stationary wavelet transform
TA	Temporal attribute
TRANSCOMS	Transmission companies
TT	Time-time transform
VC	Vapnik-Chervonenkis
WAMs	Wide area management system
WT	Wavelet transform
XGBR	XGBoost regressor
N_{SER}	Number of series connected PV modules
$V_{DC\text{-}Link}$	DC link voltage at inverter input
V_{MP}	PV module voltage at maximum power point
V_{PHASE}	RMS value of phase voltage at inverter output
P_{STRING}	Maximum power from a series connected PV string
P_{MP}	Maximum power from a PV module
N_{PAR}	Number of parallel strings
P_{INV}	Inverter power rating
N_{ARRAYS}	Number of PV arrays
P_{PLANT}	Solar plant require rating
I_{INV_RMS}	RMS current through inverter
I_{INV_PEAK}	Peak current through inverter
$I_{INV\_RIPPLE\_PEAK}$	Peak-to-peak current ripple through inverter
L_{INV_PWM}	Inverter PWM coils inductance
f_{SW_INV}	Inverter switching frequency
I_{INV_AVG}	Average value of inverter current
V_{RIPPLE_PEAK}	Peak to peak DC Link voltage ripple
$C_{DC\text{-}Link}$	DC Link capacitor value
V_{RATED_CDC}	Rated voltage of DC Link capacitor
d_{max}	Maximum allowable duty cycle
I_{IGBT_MAX}	Maximum current through IGBT
I_{RIPPLE_PEAK}	IGBT peak to peak current ripple
L_{BOOST}	Boost converter coil inductance value
V_{MIN}	Minimum input voltage to Boost converter
C_{BOOST}	Boost converter capacitor value
C_{obj}	Classifier objective function
R_{obj}	Regressor objective function

PREFACE

The detection, classification, and localization of faults in power systems are essential to prevent prolonged power supply interruptions, mitigate cascading failures leading to blackouts, and ensure secure and uninterrupted bulk power transfer over long distances. Prompt fault localization facilitates swift maintenance actions by utility operators, thereby restoring normal operation quickly. Despite advancements in measuring instruments aiding fault detection, pinpointing the exact fault location remains challenging. Model-based techniques like impedance-based and traveling wave methods have been utilized for fault localization; however, each exhibits certain limitations, such as computational complexity and reliance on expensive equipment.

With the advent of sophisticated measuring instruments such as synchronized phasor measurement units and fault data recorders, real-time power system voltage and current data are accessible, which has led to an increasing interest in process history-based fault diagnosis techniques, particularly machine learning (ML) models. These models offer automated and rapid fault diagnosis capabilities, thus addressing the shortcomings of traditional methods. Despite numerous studies on ML-based power system fault diagnosis, there is a lack of a comprehensive review capturing the state-of-the-art advancements. This thesis aims to fill this gap by providing an inclusive review of ML-based fault diagnosis techniques. The review incorporates discussion supported by tabulated facts for fault detection, classification, location identification, and exact localization works, with techniques used, different simulation tools used, and their application systems. Further, the advantages and disadvantages of all the fault diagnosis techniques, the status of research, research trends, and perspectives for needed research have been highlighted.

Furthermore, from the critical analyses of the literature review, it was found that several models, such as Bayesian ridge regression, XGBoost, and Extra tree application, remain absent for power system fault diagnosis despite demonstrating superior performance in other power system related areas. Hence, in this thesis, these models have been proposed for conventional power system fault diagnosis, and the results demonstrate that they outperformed other compared models. To carry out the proposed work a diverse power system fault database was generated considering actual field variations of fault attributes utilizing the IEEE 9 Bus system.

Conventional power system protection schemes are encountering new challenges with the increasing integration of renewable energy sources (RES) into the continuously evolving power system architectures. The quick and precise power system fault diagnosis with RES integration is even more vital for transmission companies (TRANSCOMs) than it used to be for conventional power systems. The longer a RES plant stays out of service due to faults, the heavier the financial losses incurred by the associated TRANSCOM. TRANSCOMs not only lose revenue from unsupplied power but also incur expenses by compensating for the shortfall in energy for essential loads. Additionally, TRANSCOMs may face penalties for their inability to transmit power generated by RES during outage periods. Hence, timely fault classification and localization become more crucial for power systems with RES integration.

Adaptive and intelligent methodologies, incorporating ML techniques for fault diagnosis, have emerged as promising solutions. However, RES introduces bidirectional power flow in the lines and increases the short-circuit current level. With RES penetration levels reaching up to 40%, the short-circuit current can increase to as much as seven times the normal level. This significant rise can disrupt the normal operation of relays that were designed for normal short-circuit currents, necessitating the resetting of relays and the upgradation of protection devices. Further, the dependence of RES on weather conditions causes the power fed to the grid to keep changing, resulting in fluctuations in the increased short-circuit level of the transmission lines. As a result, increasing RES penetration will cause fluctuations in the fault signatures of various faults, despite any change in the fault resistance and inception angles, posing challenges to ML-based fault diagnosis. Thus, there is a need to assess the impact of RES integrations on fault diagnostic schemes. For this, a diverse fault dataset for the IEEE 9 Bus system as a conventional power system and RES integrated system as a solar photovoltaic integrated IEEE 9 Bus system was generated considering actual field variations of fault attributes and temperature and irradiance variations.

Thus, this dissertation assesses the impact of RES integration on the performance of ML models in the fault diagnosis of power systems. The integration of RES, such as solar power, introduces significant changes to system topology and fault characteristics, challenging the effectiveness of ML models trained using conventional power system fault data. The study proposes an analysis of ML model performance under two scenarios: impact analysis (absence of fault data for newly integrated RES) and adaptability analysis (availability of fault data over time). The impact analysis of different size RES integrations, varying in power generation capacity on the performance of these models, showed that the classification and localization performance of ML models degraded significantly. Further, the adaptability analysis reveals that Bayesian ridge regression is the most effective method, demonstrating efficient transfer of learning with minimal inclusion of new fault data for fault localization.

At the same time, XGBoost, Extra Tree, and Random Forest classifiers performed well for fault classification with gradual fault data availability.

The thesis further explores the challenges posed by the ongoing penetrations of RES on ML based fault diagnosis, as changed systems fault data with required diversity is immediately unavailable after an increase in RES penetration. An increase in the RES penetration level of an existing transmission network significantly alters its topology and fault characteristics, depending on the level of increase. Thus, there is a need to examine ML models' fault classification and localization performance for power systems under increasing RES penetrations of varying levels. Therefore, in this thesis, initially, ML models are explored for their adaptability to old penetration fault data, as fault data for increased penetration is immediately unavailable. Further, incremental learning approaches for ML models are adopted to test their continual learning ability as new fault data becomes gradually available for increased penetration level. The findings underscore the necessity of adaptable ML models for effective fault diagnosis in dynamically evolving power systems with increasing RES penetration. XGBoost was identified as a leading model for maintaining fault diagnosis accuracy amidst ongoing RES integrations, demonstrating superior performance in classifying and locating transmission line faults under varying RES penetrations and fluctuating weather conditions. The presented comprehensive review and empirical analysis within the thesis provide valuable insights into the future directions of ML-based power system fault diagnosis.

Chapter 1 INTRODUCTION

1.1 Introduction

The traditional structure of the power system has been characterized by a "vertical" framework comprising centralized generation, transmission, and distribution networks. Within this structure, power transmission networks hold particular significance as they facilitate the transportation of substantial quantities of high-voltage power from generating units to substations [1]. Consequently, ensuring consistent and highly reliable service to substations is the primary obligation of any electrical power system [2]. Modern society heavily relies on complex and widespread electrical networks for critical services like healthcare, residential and commercial loads, industrial manufacturing, and transportation [3]. Consequently, seamless energy delivery within urban infrastructure without interruption has become a major national concern.

Large deviations in power system parameters like current and voltage indicate that abnormal phenomena have occurred in the power system, such as faults. Faults are inevitable in power systems and are often more prevalent in overhead transmission and distribution lines than other components. Various factors, such as natural disasters, man-made accidents, and interference from animals or trees, cause faults in transmission and distribution lines [4]. These faults not only jeopardize the reliability of the system but also impact end-user satisfaction. In response to these challenges, researchers have intensified their efforts toward devising methodologies for the prompt detection, classification, and localization of faults.

Historically, fault localization techniques relied on manual inspections. Later model-based techniques, namely impedance and traveling waves methods, became popular for fault diagnosis [5], [6]. However, the impedance-based method, although effective, poses significant drawbacks. It is tedious, complicated, time-consuming, and computationally intensive, relying heavily on mathematical modeling and necessitating expertise in the domain. Conversely, traveling wave-based methods leverage transient information during the fault occurrence. Nevertheless, their implementation is cost-prohibitive due to the requirement of installing traveling wave fault locators along transmission lines to capture precise transient data for fault localization accuracy. While both impedance-based and traveling wave-based methods demonstrated effectiveness within their designated networks, any alterations in the system configuration jeopardized their accuracy. As a result, a pressing need emerges for more robust and adaptable fault localization techniques to address the evolving complexities of modern power systems.

The advent of advanced measuring infrastructure, such as phasor measurement units (PMUs) and fault data recorders deployed along transmission lines, offers valuable insights into transmission line disturbances and real-time fault data collection [7]. The widespread positioning of metering infrastructure throughout the power system has resulted in vast amounts of data, prompting the exploration of machine learning (ML) algorithms for smart fault detection, classification, and localization [8]. While numerous studies in the literature have explored ML-based approaches for power system fault diagnosis, there remains a need for a comprehensive and up-to-date review of ML-driven fault diagnosis in power systems. Thus, this dissertation aims to address this need by presenting a systematic review that encompasses all facets of ML-based power system fault diagnosis to address this research gap.

Further, many ML models have already been explored for conventional power system transmission and distributed lines, however, a few recent models have still not been explored. Thus, this dissertation proposes and explores previously unexplored ML models for transmission line fault classification and localization. To facilitate this research, a diverse fault database was generated using simulations of the IEEE 9 Bus System in MATLAB, incorporating real-world variations in fault attributes.

The scarcity of conventional energy resources and their hazardous effect on the environment have led to the integration of renewable energy-based power generation units, such as solar PV and wind plants, into the power system. These RES plants' integration can range from small-scale in kilowatts to large-scale in megawatts. Many RES have been integrated into the grid worldwide in the past few decades [9]. The integration of large-scale RES is done after an initial analysis of optimal sizing and placement; thus, the size may vary but is limited to the maximum allowable size while integrating into a conventional power system [10]. The recent increase in RES integration into transmission and distribution systems transforms traditional radial systems into meshed networks, thereby presenting novel challenges in fault detection, classification, and localization [11].

The integration of RES into power systems has become increasingly prevalent, however, with RES penetration levels reaching up to 40%, the short circuit current can increase to as much as seven times the normal level [12]. This significant rise can disrupt the normal operation of relays that were designed for normal short circuit currents, necessitating the upgradation of protection devices [13]. Furthermore, the increased short circuit level of the network necessitates a reevaluation of fault detection and localization methods [14]. The longer a RES plant stays out of service, the heavier the financial losses incurred by the associated power transmission company (TRANSCOM). Hence, timely fault localization is crucial for power systems, especially with RES integration. During the outage, TRANSCOMs lose payment for unsupplied power and must compensate for the curtailed energy to the essential loads. Furthermore, TRANSCOMs are required to pay some money as a penalty to RES plant owners for not being able to transfer the power generated from RES during the outage period. Thus, fast and accurate fault localization is crucial for TRANSCOMs [5].

The work available for RES integrated power system fault diagnosis is limited, especially for fault localization using regression models. Thus, this dissertation analyzes the impact of RES integration on ML algorithms, investigating whether models trained on conventional power system fault data can effectively classify and locate faults in RES-integrated systems. Also, adaptability analysis of several classification and regression models has been performed to find the most adaptable ML model that can quickly learn RES-integrated fault scenarios with the least available data.

Additionally, the dissertation assesses how ML classification and regression models perform as RES penetration levels increase and power fed to the transmission networks fluctuates due to varying weather conditions. The investigation into adaptability demonstrates the models' ability to leverage historical fault data from lower penetration levels to effectively classify and localize faults as penetration levels increase. Concurrently, the study on learning competence seeks to identify the most efficient model for quickly learning fault patterns associated with higher penetration levels, utilizing an incremental learning approach for fault classification and localization. These assessments contribute to a deeper understanding of the practical implementation of ML-based fault diagnosis in modern power systems undergoing increasing integration of RES.

1.2 Background

1.2.1 Conventional Power System

The conventional power system operates on traditional methods of power generation, typically relying on thermal, hydropower, or nuclear power plants, moreover, the power transfer from the generating units to consumers is unidirectional in nature, using transmission and distribution lines. Power generation in the conventional system relied on the consumption of fossil fuels such as coal and natural gas, or nuclear reactions. This process involves converting the heat energy from these sources into steam to run turbines for power generation. After generation at remotely located hydropower, thermal, or nuclear plants, the power is transmitted over long distances through high-voltage transmission lines to substations and distribution networks. Transmission lines are crucial for delivering bulk power from power plants to load areas. At the distribution level, the voltage is stepped down according to consumer requirements, such as those of residential, commercial, and industrial users.

The conventional power system is typically organized as interconnected grids, where multiple power plants are connected to a centralized network. The grid operators manage the flow of power to ensure a stable and reliable supply and balance power generation with demand in real-time. Although the conventional power system used to be relatively simple in operation and control, issues such as environmental pollution, greenhouse gas emissions from fossil fuel combustion, nuclear waste from nuclear power plants, and unrestrained consumption of conventional energy sources have raised alarming concerns about energy security and sustainability. Thus, there is a transitional global shift from conventional power generation towards cleaner and more sustainable energy sources, resulting in the evolution of modern power systems.

1.2.2 Modern Power System

Growing concerns about climate change, environmental pollution, and sustainability have prompted an increasing emphasis on transitioning away from conventional means of power generation towards renewable energy sources (RES) based power generation like solar, wind, tidal, biomass, and geothermal energy, as depicted in Figure 1.1. This transition involves integrating RES-based power plants into existing grids, alongside energy storage systems and smart grid technologies. This RES integrated power system represents a modern power system, combining both conventional and RES power plants to reduce dependence on conventional energy sources and minimize losses in transmission lines, thereby enhancing transmission efficiency and improving grid reliability.

India itself has planned the world’s largest solar power plant in Bhadla, Rajasthan, and the world’s third-largest in Pavagada, Karnataka. Large-scale PV plants are typically implemented in stages, starting with commissioning a fraction and gradually increasing penetration levels to full capacity [15]. However, along with several advantages come associated challenges for such generating units. Key challenges in RES integration include intermittency, variability, and grid stability, necessitating innovative solutions and advanced grid management techniques. The fluctuating nature of weather conditions throughout the day, such as temperature and irradiance value, significantly impacts power generation from these plants, resulting in fluctuating power fed into the grid. As a result, stability and power management in RES-integrated areas have become very challenging [16], [17]. Challenges arising from RES integration include power system operation and control, balancing supply and demand, and critical settings of protection schemes, as they alter the short circuit current level and its direction [18].

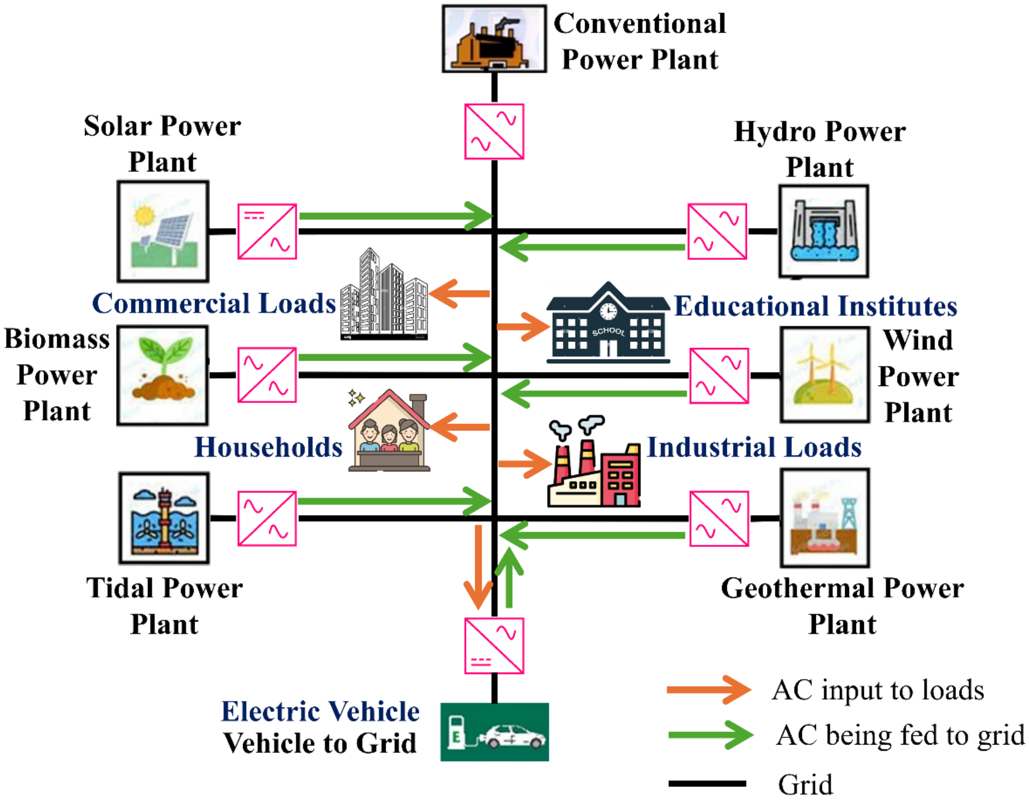


Figure 1.1 Types of renewable energy sources integrated into the grid.

The complexity of the power system increases with new RES integrations, rendering it more vulnerable to maloperations of protection devices during sudden load changes and switching actions. The ever-increasing integrations of RES, distributed generations (DGs), microgrids, and newer generation loads in existing power systems pose new challenges to conventional protection schemes. Most of these newer generation sources and loads utilize power electronics interfaces and allow bidirectional power flows, leading to various protection issues such as blinding of protection, false tripping, islanding problems, loss of coordination, and auto recloser problems [19]–[21]. Consequently, power system operation and control have become increasingly challenging for operators, and manual monitoring for maintenance purposes has become difficult [22]. Thus, the transition from conventional power to a modern power system cannot be achieved without an equal and parallel transition from conventional protection schemes to more advanced schemes.

1.2.3 Power System Faults

Faults in the power system can manifest as either transient or permanent faults, depending on the duration of the disturbances. Transients occur for short intervals, typically ranging from microseconds to milliseconds, and are often self-quenched, with minimal impact on power system operations [23]. However, they signify abnormalities in power system equipment and should not be disregarded, as they may escalate into permanent faults [24]. Transients are caused by various factors, such as degradation of system components, load changes, lightning strikes, or switching actions. Incipient faults, a type of internal transient disturbance in the power system, occur for short durations ranging from ¼ cycle to multi-cycle with low fault current values [23]. Incipient faults commonly arise from cable aging, as insulation deteriorates over time due to electrical and mechanical overstress, chemical pollution, and environmental conditions [25], [26]. Repeated occsurrences of these faults may lead to permanent faults in electrical transmission and distribution systems. Real-time monitoring of pre-fault symptoms aids in detecting abnormalities in power system operation, allowing operators to anticipate potential permanent faults to some extent [27].

Permanent faults can be broadly classified into symmetrical and asymmetrical faults, further divided into five sub-categories: single line to ground (SLG) fault, double line to ground (LLG) fault, line to line (LL) fault, three-phase to ground (LLLG) fault, and three-phase (LLL) fault [28]. Their occurrences are further classified into eleven possibilities depending upon the phases involved, as depicted in Figure 1.2.

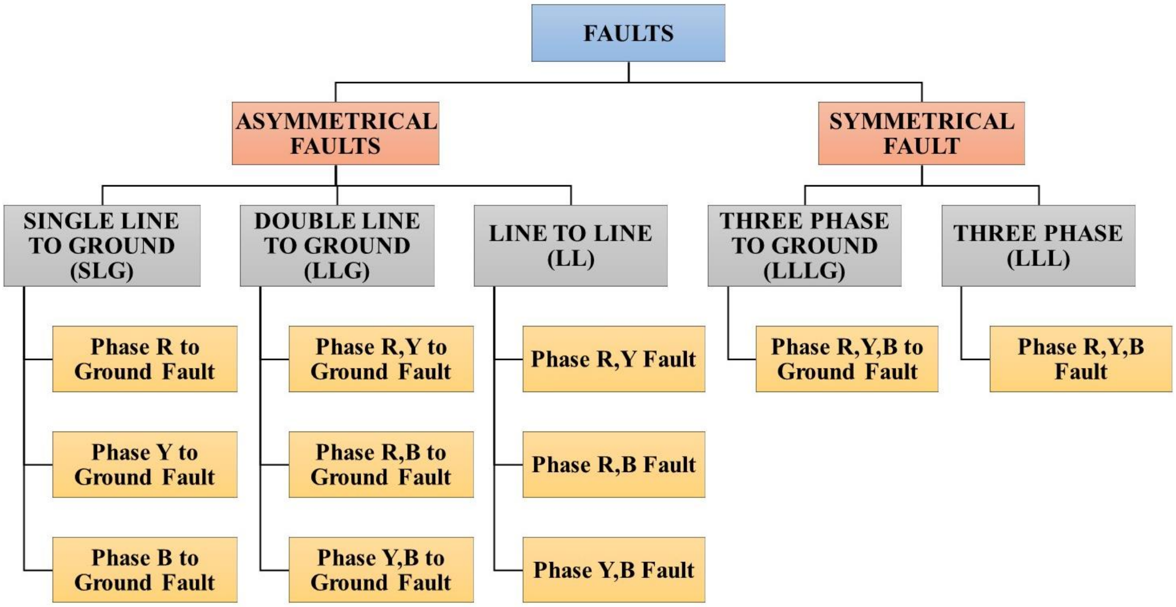
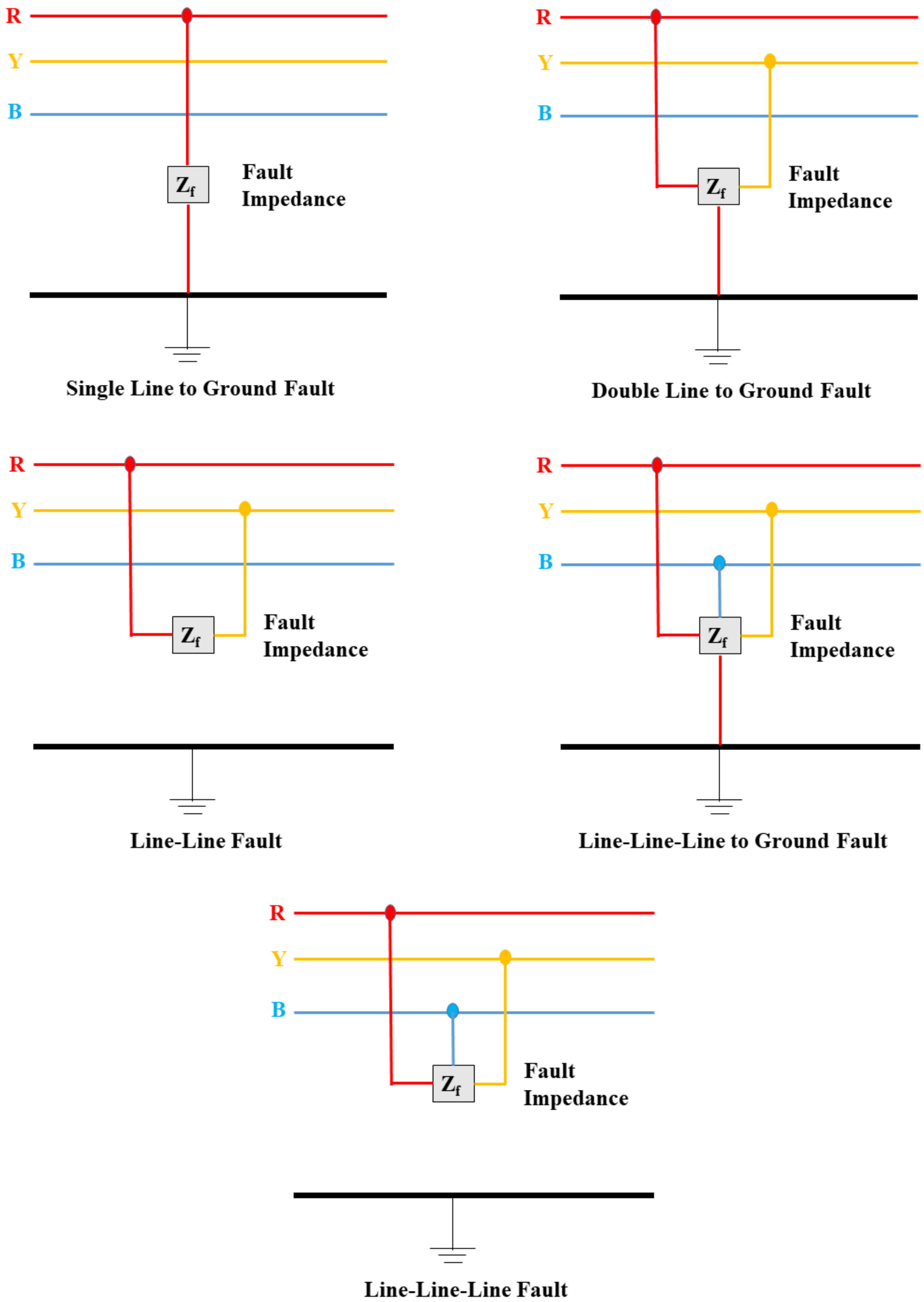


Figure 1.2 Types of permanent short-circuit power system faults.

The alteration in voltage and current characteristics during the transition from normal to faulty operation is significantly influenced by fault attributes such as fault type, impedance, distance, and inception angle. Figure 1.3 represents the fault scenarios of each fault type. Where Z_f represents the fault impedance. Fault impedance is typically categorized as either high or low impedance faults. A detailed discussion about fault attributes and their typical range of values has been covered in chapter 3 of this dissertation. High impedance faults pose challenges in fault diagnosis due to their extremely low fault current values.



Single Line to Ground Fault

Double Line to Ground Fault

Line-Line Fault

Line-Line-Line to Ground Fault

Line-Line-Line Fault

Figure 1.3 Illustration of fault scenarios for each type of fault.

Consequently, numerous research efforts have specifically focused on high-impedance fault detection, classification, and localization. Timely detection, classification, and localization of faults are crucial to prevent prolonged power supply interruptions, cascading failures, and blackouts, thereby ensuring the reliable operation of the power system [29], [30].

The likelihood of faults occurring in RES-integrated power systems is relatively high due to various factors. These include failures or malfunctions of protective devices, degradation of power system components such as insulation breakdown [27], natural disasters like floods, storms, and lightning strikes with heavy rainfall causing cable breakage [4], sudden load changes, switching actions, and high temperatures leading to equipment overheating [31]. Additionally, RES can contribute to fault currents during faults, resulting in increased fault currents in lines [44]. As a result, the signature of a fault occurring at a location will change with the incorporation of RES units, even if the fault attributes remain constant. Figure 1.4 illustrates the increase in fault current levels with increasing RES penetration levels, demonstrated by the Bus 7 current waveform obtained from MATLAB Simulink models of the IEEE 9 Bus System and the IEEE 9 Bus System with RES at Bus 7. The considered temperature and irradiance values for the given RES integrated system are 1000 W/m^2 and 25 ^{o}C. The I_f in the figure represents the peak value of the fault current. Moreover, the wide variability in power generation from RES plants due to weather conditions, particularly temperature and irradiance fluctuations throughout the day and year, directly impacts the power output from solar PV-based RES [48]. Consequently, the power fed into the grid fluctuates, leading to variations in fault current levels [49]. Therefore, it is imperative to consider temperature and irradiance variations while analyzing solar PV-based, RES-integrated transmission networks [50].

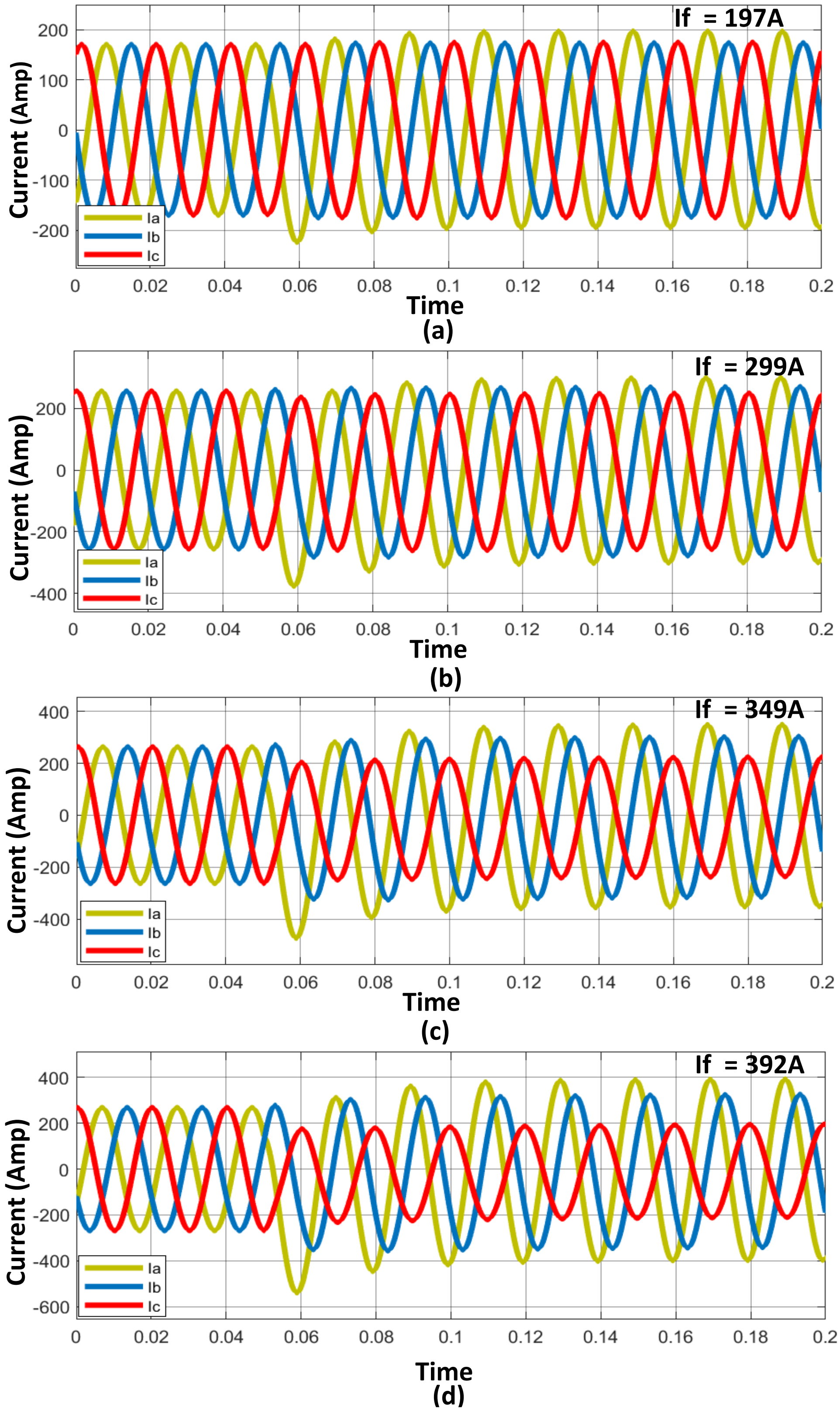


Figure 1.4 Illustration of change in fault current level of Bus-7 of (a) Standard IEEE 9 Bus system, (b) 10MW RES at Bus-7, (c) 20MW RES at Bus-7 and (d) 30MW RES at Bus-7 of IEEE 9 Bus system.

1.3 Literature Review

1.3.1 Machine Learning-Based Conventional Power System Fault Diagnosis

In recent times, numerous researchers have embraced ML methods for diagnosing faults in power systems. These techniques utilize historical data of both faulty and non-faulty power systems to train models that can automatically detect, classify, and locate faults. This approach grants the system self-healing capabilities. Once trained, these models deliver rapid and precise decisions, eliminating the need for mathematical modeling or domain expertise [32]. Rapidly locating faults aid utility operators in promptly conducting maintenance to restore normal power system operations [33]. Successful fault detection and swift localization not only expedite line restoration and ensure uninterrupted power supply but also yield economic advantages by reducing revenue loss due to power outages and saving on labor and vehicle expenses associated with manual fault location searches [22].

Several works on ML-based power system fault detection, classification, and localization are reported in the literature. Models such as Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Neural Network (NN), Naïve Bayesian (NB), Decision Tree (DT), Gaussian Process Regression (GPR), and Ensemble methods have been utilized for power system fault diagnosis in the literature [6].

The work proposed by Galvez and Abur used k-means clustering and weighted directed trees for fault localization using the time of arrival of faulted voltage transients at buses to identify the location of faults. The recorded voltage signal was decomposed using discrete wavelet transform (DWT) for extracting valuable information [11]. Chen et al. also leveraged measurements from various buses to locate faults in distribution networks. Their scheme employed Deep Graph Convolutional Networks for fault localization within the

IEEE 123 bus system, focusing more on faulty bus identification through classification techniques than pinpointing exact fault locations [34]. Zhang et al. introduced a neural network-based fault localization approach for the IEEE 39 bus system, utilizing a bi-directional gated recurrent unit (Bi-GRU) neural network. They incorporated an attention mechanism to extract fault features from the data, comparing results with various regression techniques such as regression tree (RT), linear regression, random forest (RF), support vector regression (SVR), and back-propagation neural network (BPNN) [35]. Liang et al. proposed an adaptive convolutional neural network (ACNN) for faulty line identification in the IEEE 33 bus distribution network. Their algorithm utilized two-terminal line fault data for fault localization on the selected line [36]. Gashteroodkhani et al. employed SVM for faulty section identification and the Bewley lattice diagram for fault localization in a hybrid overhead and underground transmission line network. Their study utilized three transformation techniques—wavelet, S-transform, and time-time (TT) transformation highlighting that TT transformation yielded the best results for faulty section identification with the SVM model [37]. Sahani and Dash developed a fault localization scheme for a series compensated double-circuit transmission line using a weighted random vector functional link (RVFL) network, least squares SVM, extreme learning machine (ELM), and RT. They utilized empirical wavelet and Hilbert transforms for feature extraction [38]. Shafiullah et al. utilized wavelet transformation for feature extraction from fault datasets for fault localization in two-bus systems using SVR, NN, and ELM techniques [39].

LR has been used as a base classifier model [40] for fault detection and classification. KNN has been applied as a classifier for fault detection in the microgrid [41], fault classification in a two-bus system [42], fault classification in the IEEE 9 Bus system [43], and faulty zone detection in the IEEE 9 Bus system [44]. Many researchers have included NB as a base classifier model for fault classification in a two-bus system [45] and microgrid fault detection and classification [46]. DT has been used as a base model for both fault detection and classification [47], [48], and faulty zone detection in the two-bus system [44]. SVM was applied for high-impedance fault detection in the IEEE 13 Bus System [49] and distribution feeders [50]. SVM is an excellent classifier for the classification of faults in different transmission networks [46], [51]–[54]. Similarly, the work done by [35], [39], [55], [56] has all used SVM regression (SVR) for fault localization of different power networks. GPR has also been used for fault localization [57]–[59]. Interest in ensemble methods is growing in the literature, and works are frequently published on one or more ensemble methods. RF has been used for fault classification [35], classification and localization [60], and fault location detection [44]. Other ensemble models such as Bagging, Boosting, and AdaBoost can also be found in the literature for transmission and distribution network fault classification [43], [61], classification and faulty branch identification [62]. The presented work by Mrabet et al. for detecting fault location on IEEE 9 bus systems used RF for detecting fault location and duration from phasor measurement unit (PMU) data. It compares the results of RF with other ML models such as deep neural network (DNN), NN, DT, hoeffding tree, SVM, NB, and KNN [44]. Many fault classification and faulty line/zone detection works have been reported, giving outstanding performances on the studied power system networks. However, relatively fewer studies on fault localization using ML regression are available in the literature.

1.3.2 Machine Learning Based RES Integrated Power System Fault Diagnosis

Limited work has been reported in the literature for ML-based power system fault diagnosis with RES integration. Recently, a faulty line identification work has been presented by Wang et al. with RES integration using a deep learning framework where CNN layers have been used for feature extraction from voltage and current waveforms [63]. Shah et al. presented a wind energy-based RES integrated IEEE 9 Bus system fault detection and classification using NN and SVM in [64]. The IEEE 9 Bus system consists of three generators, and in this study, the RES has been integrated by replacing the third generator. Thus, no new source has been included in the system; hence, optimal placement and sizing analysis of the IEEE 9 Bus have not been considered in the study. Rai et al. presented a fault classification work for a distribution network having two DGs using a convolutional neural network (CNN) [17]. Other works reported by Lin et al. for distributed energy sources (DER) integrated systems using ML-based fault diagnosis are Support Vector Data Description-based faulty region identification [65] and SLG fault detection for varying penetration levels of DER in a distribution network [16]. Srivastava and Parida presented fault classification and localization work in [66] for a five-bus test system of 11 kV medium voltage distribution line with two DGs using linear SVM, KNN, and Bagging for fault classification and GPR, RT, SVR, and linear regression for fault localization. However, the work assumed DG as a fixed source and sufficient fault data availability. Maruf et al. presented a study on the identification of an SLG fault phase and faulty segment using ANN, SVM, Bagging, and AdaBoost [67]. The study utilizes the IEEE 13 Bus test system, incorporating two DGs of 1800 kVA and 2600 kVA located at two feeder buses. The performance of the faulty phase identification classification models is evaluated under varying DG penetrations, specifically

20%, 30%, and 50%. The results indicate that as the penetration of DG increases, the accuracy of faulty phase identification decreases. The work also used k-means clustering to predict missing data.

1.3.3 State – of – the – art Available on ML-Based Power System Fault Diagnosis

Owing to the increasing concerns for fault diagnosis, several research articles have been published for enhancing automatic fault detection, classification, and localization using ML techniques. A few review papers have been published by esteemed researchers, highlighting the work carried out in this area. Haque and Kashtiban emphasize the NN applications in electrical power systems for various closely related topics such as load forecasting, fault diagnosis, economic dispatch, and security assessment. It briefly discussed the contribution of the NN alone in the abovementioned electrical power system applications [68]. Prakash et al. presented another review focusing on techniques such as wavelet transform, fuzzy logic (FL), NN, genetic algorithm (GA), and SVM for fault detection in distribution networks. It covered two ML techniques, viz., NN and SVM [69]. The review presented by Aleem et al. mainly emphasizes the concepts and challenges associated with model-based and data-driven fault diagnosis techniques [70]. Ferreira et al. presented a survey on intelligent system-based fault diagnosis of power system transmission lines, discussing several classical and computational intelligence techniques that have been applied to fault diagnosis in transmission lines [71]. Tirnovan and Cristea reported the works of certain popular classifiers, namely, NN, KNN, NB, and SVM, used for fault diagnosis [31]. Chen et al. gave an overview of fault diagnosis with some popular classifiers like NN, SVM, FL, and DT, with modifications to emphasize fault location estimations from these techniques [72].

Gururajapathy et al. work covered fault diagnosis for distribution systems with DGs containing various conventional and artificial intelligence techniques (NN, SVM, FL, GA, and matching approach) for the estimation of the fault location. The authors have discussed in length the types of transmission line faults and traditional fault localization methods, namely, the traveling wave method and impedance-based methods (viz., Takagi algorithm, reactive component method, and Girgis equation) [73]. Mishra and Ray presented multiple NN configurations like backpropagation, probabilistic feed-forward, and radial basis function, along with FL, SVM, ELM, and RF. It also discussed the strengths and weaknesses of the NN, SVM, ELM, and RF [74]. Prasad et al. have reported the applications of various classifiers, namely, SVM, GA, DT, and ELM, for fault classification [75]. The review works presented by Mishra and Ray and Raza et al. are more comprehensive and chronologically tabulated [74], [76].

1.4 Research Gap

The existing literature lacks a recent, up-to-date comprehensive review on ML-based power system fault diagnosis, failing to systematically discuss several key aspects. Firstly, there is a dearth of analysis on the issues inherent in conventional fault diagnosis methods, which prompted the surge in the popularity of ML techniques. Moreover, a baseline framework and workflow for ML-based fault diagnosis are noticeably absent in the literature. Additionally, the literature fails to provide a structured presentation of various unsupervised and supervised learning techniques utilized for fault detection, classification, and localization. Tabulated works outlining the techniques employed, simulation tools utilized, and their applications within different systems are notably missing. Moreover, there is a notable absence of comprehensive discussions on the advantages and disadvantages of these fault diagnosis techniques, the status of research, and recent research trends, which are vital for their selection.

In tandem with the burgeoning trend of RES integration into power networks, there is a noticeable gap in studies analyzing the impact of RES integration into transmission networks on ML models' performance for fault diagnosis. Prior studies on RES often treated them as fixed power sources, neglecting the significant effect of fluctuations in power generation due to varying weather conditions. Moreover, assumptions about readily available fault data for training ML models do not align with practical scenarios, where acquiring diverse fault data can take years. Furthermore, analyses of ML models’ performance for transmission line fault diagnosis under the practical constraint of limited and unavailable fault data in the context of increasing RES penetration levels is missing in the literature. Analyzing the performance of fault diagnostic techniques under these real-world constraints is essential to ensure uninterrupted power transfer through transmission lines. This dissertation seeks to address these research gaps by specifically exploring the performance of ML models for transmission network fault classification and localization under the discussed scenarios.

Furthermore, while most ML-based fault localization schemes focus on identifying faulty lines or sections, the literature lacks adequate exploration of ML regressor models for pinpointing the exact fault locations. Identifying precise fault locations using ML regression techniques holds significant value for expediting maintenance in transmission networks. Few studies have tested regressor models for fault localization, however, none have addressed localization in RES-integrated transmission systems using either ML-based classification or regression techniques. Thus, there is an urgent need to explore other regressor models for fault localization to identify more suitable regression models for fault localization in both conventional and RES-integrated power systems.

1.5 Dissertation Motivation

The increasing integration of RES plants into transmission and distribution networks has spurred intensified research into the maintenance, protection, and control of RES-integrated power systems. The integration of RES introduces complexity and vulnerability, thereby increasing the likelihood of faults. Bidirectional power flow and increased short-circuit current levels necessitate the upgradation of protection devices [12]. Through an extensive literature review, it has become apparent that fault localization methods based on the impedance method entail substantial system remodeling whenever a new source or load is integrated. Conversely, while the traveling wave method proves effective, its high cost and susceptibility to accuracy degradation stemming from line tappings pose significant challenges. Recognizing the advantages of ML models over conventional fault diagnostic techniques, I am propelled to delve into investigating their suitability for RES-integrated systems. Therefore, I have undertaken an investigation to conduct an impact analysis of RES integration into transmission networks on the fault classification and localization performance of ML models. This endeavor aims to ascertain their efficacy for newly integrated RES in power systems, particularly in scenarios where fault data specific to RES-integrated systems is unavailable. Additionally, an adaptability analysis of selected ML models for RES-integrated power system fault diagnosis has been conducted, aiding in the selection of appropriate ML models based on their learning capability to address the changed system topology under the practical condition of limited fault data availability over time.

Moreover, an investigation on the performance of ML models for increasing penetration level of existing RES has also been performed, considering the practical scenario of fault data unavailability and limited data availability gathered over time.

Gathering diverse fault data encompassing various locations, all types of faults, and other attributes takes years. Additionally, the occurrence of different fault types with significant variations in attributes is occasional [77]. Hence, it becomes indispensable to evaluate the performance of ML models while considering these practical issues. Further, the integration of RES units significantly alters power system topology and fault characteristics [21]. This variability is further exacerbated by weather fluctuations. The deviation in fault current level can lead to increased misclassification rates and errors in fault location estimation by ML models. Therefore, it becomes imperative to consider factors such as temperature and irradiance variations for solar plants when analyzing RES-integrated transmission networks [13], [78], [79].

1.6 Research Questions

Prior to initiating the research endeavors of this dissertation, the following research questions were formulated:

- What factors contribute to the increasing popularity of ML-based power system fault diagnosis over conventional techniques?
- Is there comprehensive literature available that outlines the advantages and workflow for ML-based power system fault diagnosis?
- Are there any better ML models not yet explored in the literature for the power system fault classification and localization?

- What challenges, if any, are posed to ML-based power system fault diagnostic models by the integration of RES into power systems?
- Is there literature available that examines the impact of RES integration on fault diagnostic techniques in power systems?
- Are there existing studies discussing the adaptability of ML models to changing system topologies, such as the integration of RES into power systems?
- Will the increasing penetration of existing RES present challenges to power system fault diagnostic techniques, and if so, what are those challenges?
- Will ML models be able to adapt to increasing RES penetrations trained from old penetration level fault data?

1.7 Research Objectives

Considering the identified literature gaps, the primary objectives of this dissertation research are outlined as follows:

- To conduct simulations of both the "standard IEEE 9 Bus System" and the "RES-integrated IEEE 9 Bus System," and generate faults within these simulations for fault database formation. This entails considering various fault attributes to reflect real-world variations, including the influence of temperature and irradiance on power generation from different sizes of solar PV-based RES integrated into the transmission network under examination.
- To analyze and compare the fault classification performance of XGBoost and Extra Tree with other potential ML models utilized for classifying conventional power system faults.

- To analyze and compare the localization performance of Extra Tree and Bayesian Ridge regression with other potential ML regression models for estimating the locations of conventional power system faults.
- To assess and compare the impact of new RES integration on the classification and localization performance of ML models trained for conventional power systems.
- To examine and compare the adaptability performance of ML models for fault classification and localization post-RES integration, considering real-world scenarios of gradual fault data availability over time.
- To identify ML models that demonstrate rapid learning with limited data, by analyzing the adaptability trends of the studied ML models for classifying and localizing faults in RES-integrated transmission networks.
- To investigate the adaptability of ML classifiers and regressors for classifying and localizing transmission line faults as RES penetration increases, particularly when fault data for the current penetration level is unavailable.
- To explore the learning capability of ML classifiers and regressors for classifying and localizing transmission line faults as RES penetration increases, particularly when fault data for the current penetration level becomes available gradually over time.

1.8 Hypothesis of Work

The hypothesis section of the dissertation proposes the following conjectures:

- Beyond the commonly used SVR, GPR, and RT models, there may exist other regression models better suited for fault localization within power systems.

- The integration of RES plants into power systems is expected to profoundly affect the performance of ML classification and localization models for fault classification and localization, given the absence of fault data specific to RES-integrated systems.
- ML models will demonstrate an inherent capability to adapt to faults occurring within RES-integrated power systems as fault data gradually becomes available.
- The performance of ML models will be affected by the increasing penetration of existing RES, particularly in the absence of fault data pertaining to this increased penetration.
- Despite the unavailability of fault data corresponding to increased RES penetration, ML models will exhibit some degree of adaptation to the subsequent increase in RES penetration.
- ML models will display a tendency to learn and adapt to the increasing penetration of existing RES as fault data becomes gradually available.

1.9 Research Challenges

The challenges faced during the course of this dissertation are as follows:

- Finding the research gap from the enormous amount of literature available on ML-based power system fault diagnosis.
- Unavailability of authentic power system fault datasets on known standard data repositories such as IEEE Data Port, Kaggle, and UCI Machine Learning Repository.
- Selection of a suitable IEEE standard test system for RES integration and fault database formation.
- The creation and formation of a fault database is a highly time-consuming process.

- Selection of suitable ML models that may be least affected by the new RES integration into the power system.

1.10 Problem Statement

Machine learning models have gained extensive attention for power system fault diagnosis. However, the evolving panorama of power systems, marked by the integration of new loads, distributed generations, and renewable energy sources, has shifted research focus toward fault diagnosis in RES-integrated power systems. Despite this, there is a noticeable scarcity of research on fault diagnosis in RES-integrated power systems, particularly concerning fault localization using regression models.

Existing literature on RES integrated power system fault diagnosis often assumes the availability of fault data for training and testing ML models, treating RES sources as fixed power sources. However, the challenges posed by data unavailability and weather-induced power generation fluctuations in RES-integrated systems remain largely unexplored. The absence of fault data for newly integrated RES poses a challenge in training ML models tailored for RES-integrated systems. As a result, there is uncertainty regarding the performance of pre-trained ML models, originally trained for conventional power systems, when applied to RES-integrated systems. Additionally, information about the adaptability of ML models in this context is lacking in the existing literature.

Furthermore, the installation of large-scale PV plants typically occurs in stages, from partial commissioning to reaching full capacity. Consequently, fault data collected at each penetration level may be insufficient to capture the diverse fault attributes. Despite this, there is a dearth of analysis on power system fault classification and localization with increasing

RES penetration, considering the limitations of data availability. Thus, it remains unclear whether pre-trained ML models, based on fault data from lower penetration levels, can effectively adapt to increased RES penetration levels to continue diagnosing faults in scenarios of data unavailability.

1.11 Dataset Challenge and Description

Research scholars encounter numerous challenges in acquiring real-time power system data for their studies. These challenges encompass various aspects: Firstly, access to real-time data from power system operators and utilities are often restricted due to security and confidentiality concerns. These entities are cautious about sharing sensitive data due to regulatory compliance and confidentiality requirements. As a result, obtaining permission to access such data can be challenging and time-consuming for researchers. Obtaining permission to access such data can be difficult and time-consuming for researchers. Secondly, real-time power system data may not always be readily accessible to researchers, particularly from large-scale power grids or transmission networks. This limited availability of data sources can impede research work. Furthermore, accessing real-time power system data often entails significant costs, including subscription fees for data providers or expenses associated with setting up data acquisition systems. Ensuring the quality and reliability of real-time power system data presents another challenge. Data may contain errors, missing values, or inconsistencies, thereby affecting the accuracy and validity of research findings. Moreover, acquiring and processing real-time power system data demands specialized technical expertise and resources. Researchers may need to develop or customize data acquisition systems and software tools to align with their research objectives. Adherence to ethical guidelines and legal regulations governing the use of real-time power system data is paramount. This includes compliance with privacy and data protection laws to safeguard the rights and privacy of individuals and organizations involved. Overall, navigating these challenges necessitates careful planning, collaboration with industry partners, and leveraging available resources to ensure successful research outcomes in the field of power systems.

Hence, research scholars generally resort to simulated models to obtain data for their research endeavors, thus avoiding the challenges mentioned earlier and saving considerable time in obtaining approval to access real power system data. Moreover, simulation software such as Matrix Laboratory (MATLAB), MAT-POWER, Alternative Transient Program-Electromagnetic Transient Program (ATP-EMTP), Grid LAB Power Distribution System Simulation Software (GridLAB-D), Open Distribution System Simulator (OpenDSS), Power System Analysis Software Package (PSASP), Power System Computer Aided Design (PSCAD), Positive Sequence Load Flow Software (PSLF), Power System Simulator for Engineering (PSS/E), Electrical Transient and Analyzer Program (ETAP), DIgSILENT, and Real-Time Digital Simulators (RTDS) offer models that closely emulate the performance of real power systems [80]. Among these, MATLAB [81], PSCAD [82], PSS/E [83], and DIgSILENT [84] tools have been extensively utilized in the literature for power system fault diagnosis applications. In line with the objectives of this dissertation, the MATLAB Simulink 2021b environment has been employed to simulate the IEEE 9 Bus system and the integrated solar PV plant system, facilitating the generation of fault data required for the proposed research.

1.12 Contribution

With the accomplishment of the presented dissertation work, the contributions made have been listed below:

- A comprehensive literature review on ML-based power system fault diagnosis, its advantages and disadvantages, research trends, some key issues, and perspectives for needed research have been summarized.
- Analysis and comparison of XGBoost and Extra Tree classifiers with the potential ML classifiers and Extra Tree and Bayesian Ridge regression with the potential ML regressors for fault classification and localization of the transmission line faults.
- Analyzing and comparing the impact of new RES integration on transmission line fault classification and localization performance of various ML models.
- Analyzing and comparing the adaptability performance of the ML models for classification and localization of fault post-RES integration considering the real-world power systems scenarios of gradual fault data availability over time.
- Analyzing the adaptability of ML models for transmission line fault classification and localization amidst increasing penetration of RES when fault data for the current penetration is unavailable.
- Investigating the learning capability of ML models for transmission line fault classification and localization through incremental learning approach amidst increasing penetration of RES when fault data for current penetration is available gradually over time.

1.13 Procedural Framework of the Proposed Research

The proposed research's conceptual framework is depicted in Figure 1.5, illustrating the progression of the dissertation's work. This framework provides an overview of each chapter's contents and outlines the steps involved in training and testing ML models across Chapters 4, 5, and 6. Within each chapter, a detailed workflow is provided for scenarios under study. Thus, this presented workflow serves as a graphical summary of the dissertation, bridging the transition from Chapter 1 to subsequent chapters. It also briefs about the steps followed in chapters 4, 5, and 6 for analyzing and comparing several ML models' performance from loading the data, to preprocessing the data to make it compatible with the models' trainable format. Further, the splitting of the complete dataset is done to use the part of data for training and the rest for testing. Each study utilizes a different train-test split, discussed in detail in subsequent chapters. Hyperparameters of the models are then tuned for optimal performance, followed by model training and testing. Finally, the performance of the models is analyzed and compared to identify the best and worst performers for each scenario under study.

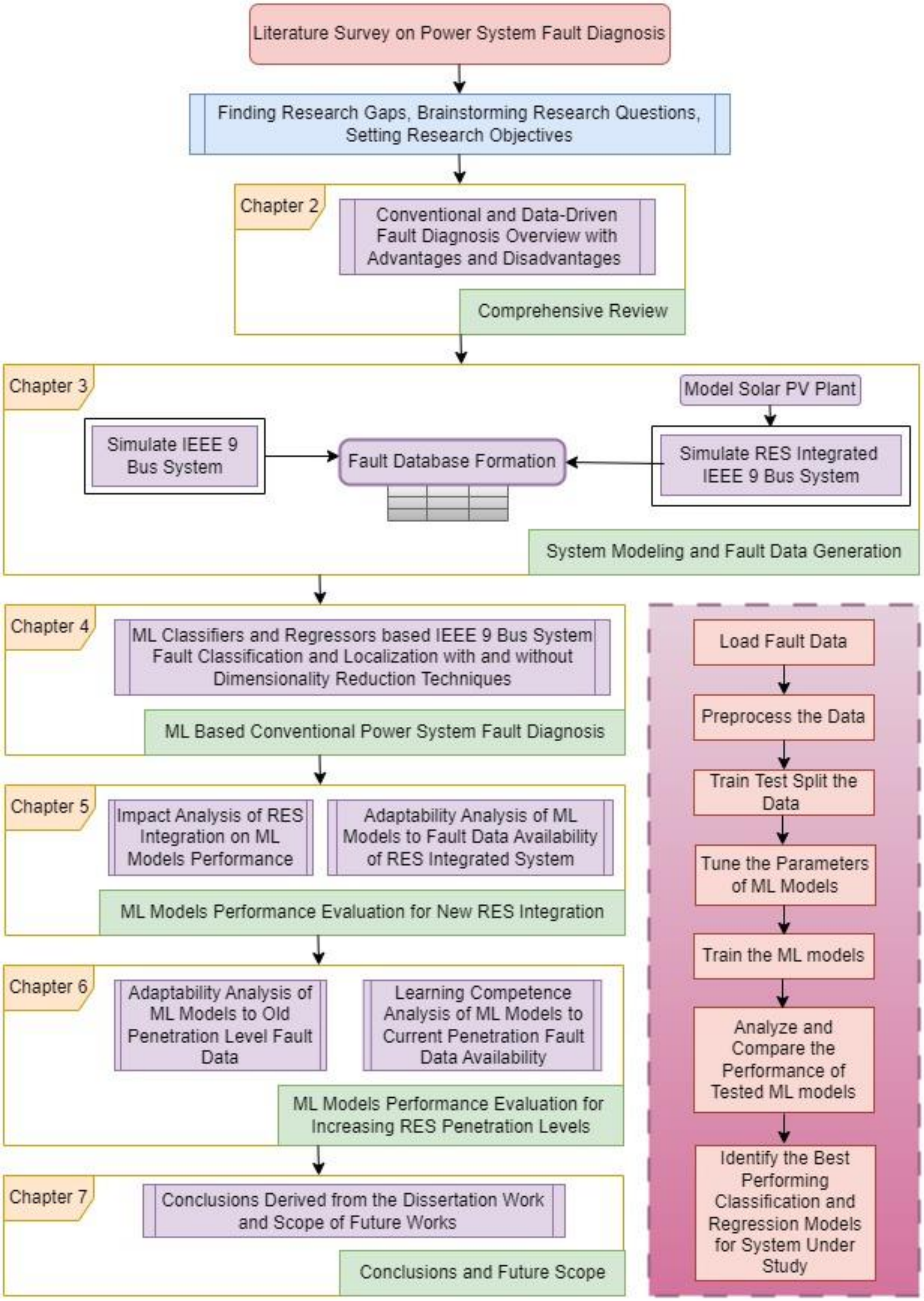

Literature Survey on Power System Fault Diagnosis
Finding Research Gaps, Brainstorming Research Questions, Setting Research Objectives
Chapter 2
Conventional and Data-Driven Fault Diagnosis Overview with Advantages and Disadvantages
Comprehensive Review
Chapter 3
Model Solar PV Plant
Simulate IEEE 9 Bus System
Fault Database Formation
Simulate RES Integrated IEEE 9 Bus System
System Modeling and Fault Data Generation
Chapter 4
ML Classifiers and Regressors based IEEE 9 Bus System Fault Classification and Localization with and without Dimensionality Reduction Techniques
ML Based Conventional Power System Fault Diagnosis
Chapter 5
Impact Analysis of RES Integration on ML Models Performance
Adaptability Analysis of ML Models to Fault Data Availability of RES Integrated System
ML Models Performance Evaluation for New RES Integration
Chapter 6
Adaptability Analysis of ML Models to Old Penetration Level Fault Data
Learning Competence Analysis of ML Models to Current Penetration Fault Data Availability
ML Models Performance Evaluation for Increasing RES Penetration Levels
Chapter 7
Conclusions Derived from the Dissertation Work and Scope of Future Works
Conclusions and Future Scope
Load Fault Data
Preprocess the Data
Train Test Split the Data
Tune the Parameters of ML Models
Train the ML models
Analyze and Compare the Performance of Tested ML models
Identify the Best Performing Classification and Regression Models for System Under Study

Figure 1.5 Conceptual framework of the proposed dissertation.

1.14 Dissertation Outline

Chapter 2 provides a comprehensive review of power system fault diagnosis challenges, particularly in the context of conventional methods, which have driven the widespread adoption of ML techniques. It establishes a foundational framework and workflow for ML-based fault diagnosis, accompanied by an extensive discussion on both unsupervised and supervised learning techniques. This discussion is further supported by tabulated facts, covering fault detection, classification, and localization methods. Additionally, various simulation tools and their applications in fault diagnosis are analyzed. The advantages and disadvantages of different fault diagnosis techniques are critically assessed, and the research trends and perspectives about needed investigation in power system fault diagnosis are outlined.

Chapter 3 presents an overview of the IEEE 9 Bus system selected for the research undertaken in this dissertation. It provides detailed descriptions of the system parameters and its modeling using MATLAB Simulink. Additionally, it discusses the fault attributes for which fault datasets have been generated. Furthermore, the chapter addresses the integration of RES into the IEEE 9 bus system at optimal locations and sizes, along with their modeling in MATLAB. Lastly, it discusses the conditions under which fault data for IEEE 9 Bus system and RES-integrated systems has been generated.

Chapter 4 presents standard IEEE 9 Bus system fault classification and localization using various ML models. While many of these models have been previously explored in the literature, XGBoost and Extra Trees (ET) as classifiers and Extra Trees and Bayesian Ridge Regression as regressors have not been utilized in the power system fault diagnosis literature. Hence, these models are incorporated in this dissertation to assess their effectiveness for fault classification and localization. Additionally, dimensionality reduction techniques such as PCA, Kernel PCA, and LDA for fault classification and PCA and Kernel PCA for fault localization are employed to compare the performance of models with and without dimensionality reduction techniques.

Chapter 5 examines the impact of newly integrated RES into transmission lines on the performance of ML models when fault data for RES-integrated systems is unavailable. In this scenario, fault classification and localization responses are predicted using pre-trained ML models trained on fault data from conventional power systems. Additionally, an investigation into the adaptability of the models is conducted to identify the most adaptable ML classifier and regressor capable of quickly adjusting to RES-integrated system fault data with minimal available data for fault classification and localization.

Chapter 6 investigates the adaptability of ML models to the increasing penetration level of existing RES concerning fault classification and localization, particularly based on the limited availability of fault data from previous penetration levels for training. The objective is to predict fault type and location for the current penetration level in the absence of data specific to that level. Additionally, an examination of the learning capabilities of the models is undertaken, focusing on the limited availability of fault data from the current penetration level. The aim is to identify the ML classifier and regressor models that exhibit the quickest learning for fault classification and localization in systems with increased RES integration.

Chapter 7 summarizes the main conclusions derived from this dissertation work. The potential future propositions of these studies, along with the future research scope, are also highlighted in this chapter.

1.15 Chapter Summary and Transition

This chapter serves as an introduction to the dissertation, providing an overview of the proposed research, its background, significance, and rationale. It begins by elucidating the necessity and importance of the study, outlining the existing literature in the field, and identifying research gaps that motivated the investigation. The identification of research gaps led to the formulation of research questions, which in turn drove the commencement of the dissertation, resulting in the establishment of precise research objectives. Before delving into the empirical research, certain assumptions were made, delineated as hypotheses of work. Additionally, the chapter addresses the challenges encountered during the research process, particularly those related to data acquisition. Furthermore, the chapter outlines the contributions of the dissertation and presents a conceptual framework for conducting the proposed research. It also provides an outline of the dissertation, delineating the content and scope of each chapter.

The dissertation outline delineates the structure of the subsequent chapters. Firstly, Chapter 2 gives a comprehensive review of ML-based fault diagnosis, elucidating various ML models in detail. Following chapter 3, presents details about the modeling of the IEEE 9 bus system and solar PV plant integration to the IEEE 9 Bus system, along with discussions on fault data generation. Moving on, chapter 4 delves into ML models for conventional power system fault classification and localization, both with and without dimensionality reduction techniques. Further, Chapter 5 examines the impact of RES integration on ML models' performance, analyzing their adaptability to RES-integrated power system fault data. Chapter 6 investigates adaptability and learning competence concerning increasing

RES penetration levels. Lastly, chapter 7 provides concluding remarks derived from the findings presented throughout the dissertation and the scope of future research work.

Chapter 2 MACHINE LEARNING BASED POWER SYSTEM FAULT DIAGNOSIS: RESEARCH ADVANCEMENTS AND PERSPECTIVES

2.1 Introduction

Many research studies have been reported on machine learning (ML) based power system fault diagnosis. However, ML techniques are evolving rapidly, and an inclusive and state-of-the-art review of ML-based power system fault diagnosis is not available in the literature. The existing reviews are excellent works that are useful to the research community. However, none of the reviews give a collective overview of power system fault types, conventional fault diagnosis methods and their advancements up to ML models, a brief overview of ML applications and stepwise procedures for problem-solving, the pros and cons of prevailing methods, and their future implications. Therefore, there was a need for a comprehensive and systematic review to cover all aspects of ML-based power system fault diagnosis to bridge the existing research gap.

Given this need and the growing trend towards ML, this chapter provides a comprehensive review of ML-based power system fault diagnosis and various models employed for power system fault diagnosis in the literature and proposed in this dissertation. At first, efforts have been made to address the issues present in model-based fault diagnosis techniques, leading to the popularity of ML techniques. Also, a baseline framework and workflow for ML-based fault diagnosis are presented. Next, various unsupervised and supervised learning techniques were discussed separately based on the work available for fault diagnosis. The advantages and disadvantages of fault diagnosis techniques have also been discussed, which can help select techniques for fault diagnosis research. Further, the works reviewed have been compiled in tabular form, showing the status of research on fault detection, classification, location identification, and exact localization with ML techniques, different simulation tools, and their application systems. Moreover, the status of the research has been profoundly analyzed to draw meaningful research trends from the literature review. Finally, the needed research on power system fault diagnosis has been given as research perspectives.

2.2 Power System Monitoring

Power system monitoring is vital for maintaining the reliability, stability, and efficiency of electrical systems. It involves real-time monitoring and control of key parameters like voltage, current, frequency, and load. By continuously monitoring voltage and current levels, the system ensures they remain within acceptable limits. Any fluctuations in these values signify potential issues such as overloading or faults. Frequency fluctuations indicate imbalances between generation and demand, which are critical for system stability. Load monitoring is essential for optimizing power distribution and preventing overloading. It also helps in planning load shedding during peak demand periods. Additionally, continuous monitoring facilitates the early detection of faults in lines or equipment failures. In essence, power system monitoring allows for the proactive management of electrical systems, ensuring smooth operation and minimizing the risk of interruptions.

With the advancement of measuring and communication devices in the late twentieth century, supervisory control and data acquisition system (SCADA) systems were developed, which provided local or remote observation and control of power system networks. Programmable logic controllers (PLCs) and remote terminal units (RTUs) are the basic components of SCADA; they communicate with one another and send information to

SCADA panels at control centers. The SCADA software processes and displays data, assisting operators in analyzing the data and making critical decisions. The SCADA system collects data at 1 or 2-second intervals. The data acquired from SCADA lacks an accurate depiction of power system subtleties. Thus, the SCADA system is insufficient to guarantee the power system's stable and secure functioning. The SCADA is frequently incapable of measuring data from all buses at the same time [85].

Wide area measurement systems (WAMs) have been developed to overcome the challenges of the SCADA system by providing more detailed power system monitoring. WAMs collect time-stamped data at a rate of 20 to 60 times per second. The main components of the WAMs system are phasor measurement units (PMUs) and phasor data concentrators (PDCs), which were first installed in China in 1995. Thus, WAMs provide data collection and transmission over high-speed communication links, allowing for real-time data analysis. WAMs collect current and voltage phasors, and frequency real-time timestamped data using PMUs coordinated with PDCs via the global positioning system (GPS). These data measurements aid in monitoring both the stability and dynamics of the power system [86]–[88]. By analyzing these measurements, operators can detect disturbances, identify potential stability issues, and take corrective actions to prevent cascading failures or blackouts.

SCADA offered a superficial view of the system stats; however, PMUs enabled a more profound study of the system stats by providing real-time data on WAMs. Such time-stamped measurements of the system data help to monitor not only the steady state but also the dynamics of the power system. One of the key advantages of WAMs is their ability to provide situational awareness over large areas, enabling operators to make real-time decisions for grid operation and control. It enhances grid reliability, stability, and resilience by facilitating faster responses to dynamic events and disturbances.

2.3 Fault Diagnosis Techniques

Historically, in usual practice, the operators relied on reports and complaints from consumers regarding trips and faults for maintenance. However, the advancements in measuring instruments and their communication ability with monitoring units enabled real-time monitoring of power systems from control centers [89]. Thus, through monitoring of power system stats at control centers, power system fault diagnosis can be done through two approaches, i.e., model-based techniques or process history-based techniques, which have been illustrated in Figure 2.1. Potential and current transformers are installed at various locations within the power system transmission and distribution line, providing voltage and current data at control centers for monitoring. Fault data recorders (FDR) are specifically installed in power systems to capture and store faults and disturbances related to voltage, current, frequency, and phase angle data at high sampling rates ranging from several hundred to thousands of samples per second. The FDRs are equipped with a triggering mechanism that initiates data recording based on certain pre-defined voltage or current thresholds or frequency deviations. The data stored in FDRs is time-synchronized with GPS. On the occurrence of a fault, data from FDRs can be retrieved for fault analysis and diagnosis. Though the data from the fault recorders helps to retrieve the fault data, however, the estimation of the exact fault location remains a challenging task and requires traveling wave or other fault diagnosis techniques [90], [91].

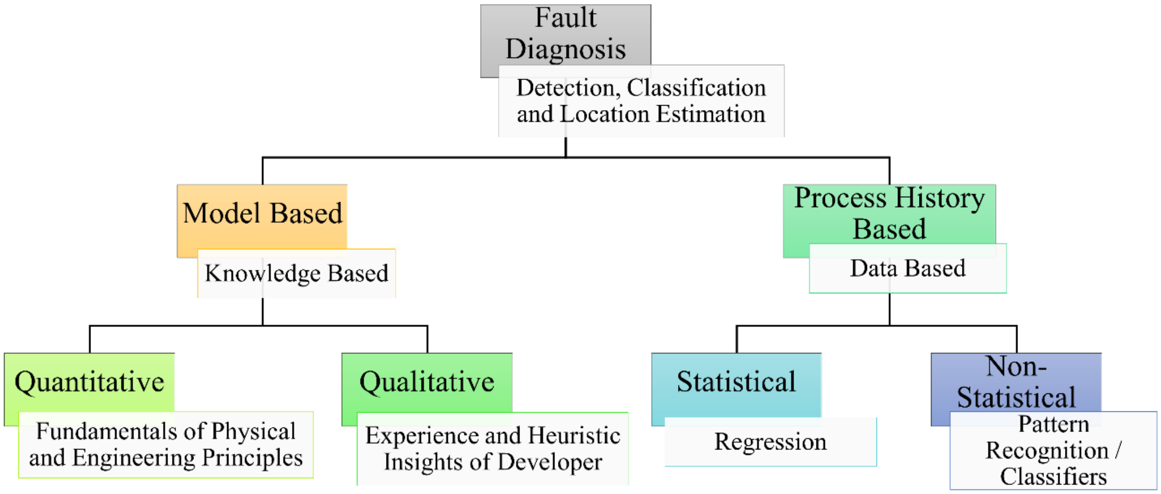

Fault Diagnosis
Detection, Classification and Location Estimation
Model Based
Knowledge Based
Process History Based
Data Based
Quantitative
Fundamentals of Physical and Engineering Principles
Qualitative
Experience and Heuristic Insights of Developer
Statistical
Regression
Non-Statistical
Pattern Recognition / Classifiers

Figure 2.1 Power system fault diagnosis schemes.

Model-based techniques, namely the impedance-based method and traveling wave methods, were used for fault localization. However, impedance-based methods are tedious, time-consuming, and computationally extensive; they rely on mathematical modeling and require domain expertise, whereas traveling wave methods require the installation of costly instruments at the ends of transmission lines, or the operator needs to go to one end of the line with a fault locator for fault localization. The successful detection and quick localization of the faults ensure not only the accelerated line restoration and continuity of the power supply but also economic benefits by minimizing the revenue loss owing to non-supplied power and saving labor and vehicle costs spent in manual searching of fault locations. The longer it takes to classify and repair a fault, the more damage may be done to the electrical power system, especially during peak loads as healthy lines of the system become overloaded. The continued overloading of lines can lead to the system collapsing, causing the power outage to last longer and affecting a larger portion of the electrical network. As a result, accurate fault classification and localization improve the operational stability and reliability of the power system and help to avoid catastrophic power outages [92]. Modern power systems are complex networks that necessitate a high-speed, precise, and reliable protective system for stable operation and consumer satisfaction.

2.3.1 Model-Based Fault Diagnosis Techniques

Conventional model-based techniques like impedance-based methods and traveling wave methods are widely used; however, they have certain drawbacks. Impedance-based methods rely heavily on accurate knowledge of system parameters and operate under various assumptions, which can lead to errors in fault location estimation. Also, it requires complex mathematical modeling and domain expertise and consumes significant time, causing its popularity to decline for real-world power systems. The traveling wave method, while effective for long transmission lines, can be costly and may yield inaccurate results when there are tappings on the line.

Furthermore, the integration of RES into power systems poses challenges for both impedance-based and traveling wave methods. The traditional impedance-based methods work well for the modeled network, but any changes in network configuration affect their accuracy. The uncertain power generation from RES units and the characteristics of power electronic devices used for their integration can further complicate fault location using these techniques [63]. As a result, traditional methods are not adaptable to system variations due to RES, DGs, microgrids, and new load integrations. As a result, intelligent paradigms are being researched for early fault identification and localization to avoid long-term power supply interruptions without relying on mathematical modeling and proficiency.

2.3.1.1 Impedance-Based Fault Localization

The impedance-based techniques of fault localization require accurate knowledge of system parameters such as line sequence resistances and reactance, load parameters, and source parameters. The values of system parameters can be undermined due to the degradation of lines, system non-homogeneity, and tapping [93]. Moreover, each type of fault requires a different procedure to locate the fault. Therefore, prior knowledge of fault classification is a must for fault localization. Further, it works under several assumptions, making fault location estimation erroneous. Impedance-based methods such as the simple-reactance method work under the assumption that fault current and terminal line current are in phase, which is not always true. The transmission network should be homogenous for the Takagi method. The modified Takagi method was developed to avoid the homogeneity assumption, but it needs information about sequence components. The Eriksson method requires knowledge of source impedance, while the Novosel et al. method applies only to short and radial transmission lines. Thereby, the impedance-based method requires complex mathematical modeling, system parameters, domain expertise, time, and several assumptions in its operation, which makes it inaccurate and less reliable for real-world power systems [94].

The fault identification and localization using SCADA and WAMs need expertise, as the operator has to identify faults based on the data from the RTUs or PMUs, while for localization they rely on conventional impedance-based methods. Rangel-Damian et al. present a SCADA system-based fault localization using an impedance-based algorithm for locating faults on the IEEE-34 bus system [85]. A WAMs-based fault localization work using an impedance-based method is presented in [86] by Wang et al., which uses synchronized fault voltages from PMUs of faulted nodes and their neighboring nodes. The change in node voltages of faulted lines and neighboring nodes is calculated from pre-fault to faulted conditions. This change in voltage at nodes helps to calculate the change in current flowing at both terminals of faulted line nodes, which thereby helps in fault localization. The method uses the value of line reactance for the current calculation to locate the fault. The presented work used the IEEE-14 bus system and a 500kV transmission line simulated on EMTP as a test system. The method's downside is that it relies on the value of line reactance for localization. However, with aging lines, reactance may change, which may increase localization errors [86]. Another WAMs based fault diagnosis scheme was presented by Zhang and Wang using the maximum posterior probability (MAP) principle for fault classification using positive and negative sequence currents at the faulted line [87]. Fan and Liao also presented WAMs monitoring-based faulty zone localization by computing driving point impedances (principal diagonal elements) and transfer impedances (off-diagonal elements) of the bus impedance matrix which can represent fault locations using measured fault voltage [88]. Thus, it can be seen from the preceding works, that both SCADA and WAMs alone cannot exactly estimate the fault location but rather rely on traditional impedance-based techniques for fault localization.

2.3.1.2 Travelling Wave-Based Fault Localization

The traveling wave method offers numerous advantages compared to impedance-based methods. It eliminates the need for complex mathematical modeling, does not work on any assumptions, and ensures high accuracy. Moreover, it operates independently of system parameters and doesn't require prior knowledge about fault types, making it suitable for both high and low-impedance faults. However, the method does have its drawbacks. Its implementation necessitates costly instruments for fault location estimation. Additionally, the presence of tappings along the line may lead to erroneous fault location estimations thus it is suitable for only long transmission lines with no or minimal tappings. Furthermore, analyzing traveling wave signals needs signal processing expertise due to their complexity.

The method's costliness stems from the requirement of installing fault data recorders at each site to capture high-precision transient information, crucial for accurate fault localization [11]. When faults occur on transmission lines, high-frequency transients or electromagnetic waves are generated at the fault point, traveling along the line at high speeds. By measuring the arrival time of these waves at both ends of the line, the fault location can be determined. Thus, the traveling wave method mandates the deployment of fault recorders at both ends of the transmission line to capture transients during faults. Alternatively, operators can position themselves at one end of the line, sending a high-frequency signal and calculating the time taken for the wave to travel back and forth.

2.3.2 Process History-Based Fault Diagnosis Techniques

In recent years, there has been a growing trend among researchers to utilize machine learning techniques for diagnosing faults in power systems. ML techniques boast robust learning capabilities, effortlessly grasping system changes and making decisions based solely on pre-fault and post-fault data [89]. Given their reliance on historical fault data for model training, these techniques are often termed process history-based fault diagnosis. Models rely on real-time power system data collected from GPS-equipped fault data recorders for online decision-making in fault detection, classification, and localization [31], [70]. Thus, the integration of SCADA and WAMs with artificial intelligence is seen as a promising avenue toward achieving a self-managing and self-healing power system [32]. Time-stamped data from PMUs and fault data recorders or voltage and current data from potential and current transformers can be utilized in process history-based fault diagnosis, empowering operators to make real-time decisions concerning power system issues such as faults. Once trained, these models deliver swift and accurate decisions without the need for intricate mathematical modeling or specialized expertise [89].

The ability of ML models to swiftly locate faults allows utility operators to promptly conduct maintenance, thus restoring normal power system operations. Thus, an effective automatic fault diagnosis scheme must ensure efficient, consistent, quick, and secure relay operations, while also minimizing maintenance costs associated with vehicle and labor deployment for fault localization. This approach accelerates restoration processes and reduces monetary losses related to unmet power demands. Machine learning has been extensively used in the literature for power system fault detection, classification, and localization. Each holds a certain significance in fault diagnostic procedures.

Fault Detection: Although circuit breakers and relays are installed along the transmission lines, a significant amount of literature is dedicated to fault detection. In instances where protection devices fail, ML-based fault diagnosis schemes can offer supplementary verification. This ensures accurate isolation of the faulted sections, serving as a backup mechanism for fault isolation.

Fault Classification: Understanding the type of fault aids operators in assessing its severity and nature, facilitating the implementation of suitable responses and corrective measures. Additionally, employing impedance-based methods for fault localization necessitates prior knowledge of the fault type. Thus, ML classifiers and impedance-based methods can be collectively used for identifying fault types, followed by fault localization, respectively.

Fault Localization: Rapid fault localization is crucial for minimizing maintenance downtime and mitigating the impact of faults on power system operations. ML-based precise fault localization helps save on labor and vehicle costs otherwise incurred in locating faults. Furthermore, swift maintenance prevents prolonged overloading of healthy lines, mitigates the risk of cascading failures, reduces losses associated with unmet demands, and improves system reliability and consumer satisfaction.

2.4 Machine Learning Overview

Machine learning trains algorithms to extract information from historical data without relying on explicit mathematical modeling [95], [96]. These algorithms enable computers to learn autonomously, making decisions and predictions based on acquired intelligence. As the dataset grows, the performance of ML algorithms dynamically improves [97]. ML methodologies find applications across various domains, including medical diagnosis, disease spread forecasting [98], stock market analysis, search engines, recommender systems, gaming, navigation, weather forecasting (e.g., wind speed prediction) [99], image and speech recognition, energy load forecasting, fraud detection [100], and parameter estimation of unknown non-linear systems [101]. ML is particularly adept at tackling complex problems involving large training datasets with multiple variables. It can adopt unsupervised, supervised, or reinforcement learning paradigms depending on data availability and target responses [102]. These learning methods encompass intelligent models suitable for classification and regression tasks, as illustrated in Figure 2.2.

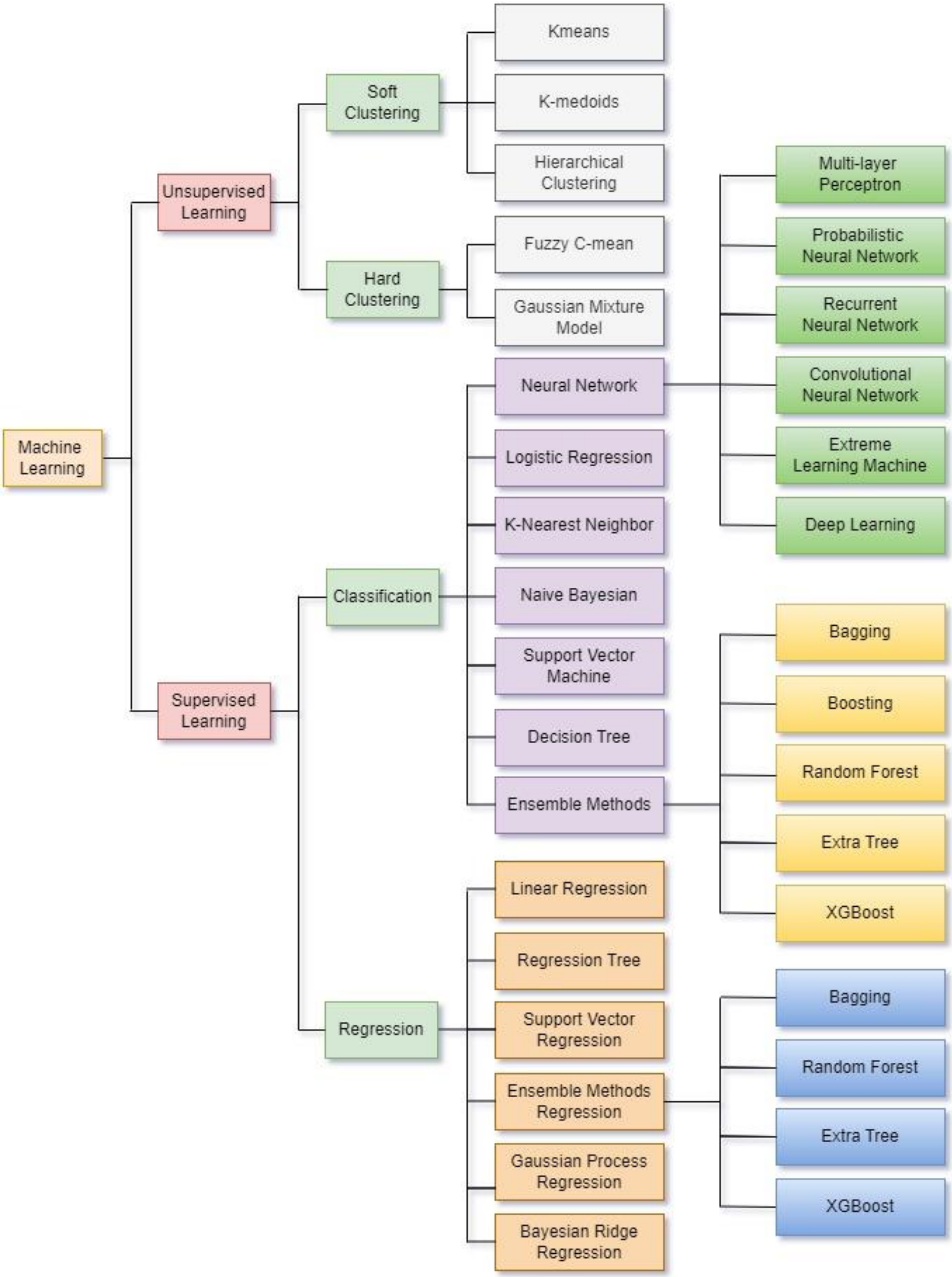


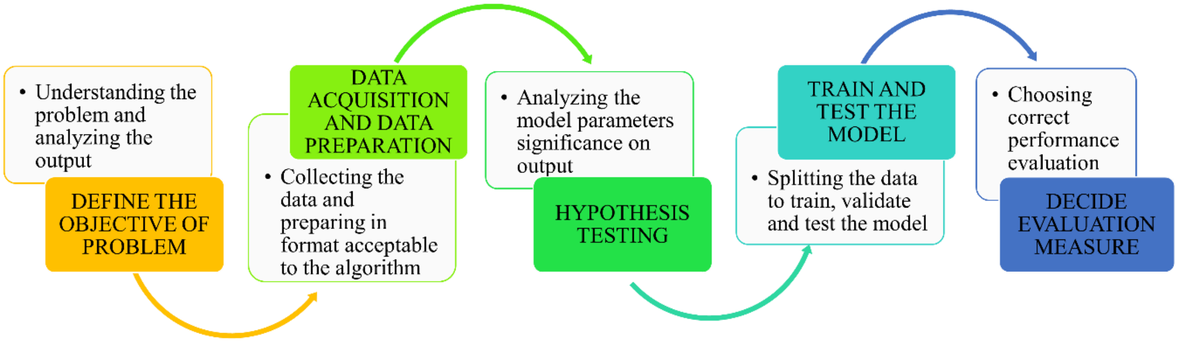
Figure 2.2 Variants of machine learning models.

Additionally, ML encompasses variants like semi-supervised learning (combining aspects of supervised and unsupervised learning), self-learning (a goal-seeking approach without external feedback), feature learning (extracting features from data), sparse dictionary learning, anomaly detection, robot learning, transfer learning, and association rule learning [29], [102], [111]–[113], [103]–[110]. Despite the diversity of ML algorithms, there are no definitive guidelines for selecting a specific algorithm for a given problem. Selection often involves multiple trial-and-error iterations, with the algorithm demonstrating the best performance on the dataset ultimately deemed most suitable. Choices between clustering, classification, or regression techniques depend on the nature of the available data and desired outcomes [95].

ML techniques have found widespread application across various engineering research domains, including chemical engineering, electrical engineering, civil engineering, industrial engineering, image processing, mechanical engineering, automobile engineering, and others. The use of ML for fault diagnosis in these fields has gained significant popularity. Examples of applications include fault diagnosis in industrial processes based on sensor data [114]; fault diagnosis in nuclear power plants and their components using sensor time-series data [115]; crack detection in plates or objects [116]; identification and tracking of cracks in civil structures [117]; detection and estimation of damage extent in composite structures [118]; image profile recognition-based tracking system [119]; fault diagnosis in electric motor drives for applications such as electric traction and propulsion [120]; diagnosis of faults in rotating equipment such as hydraulic pumps and motor parts like gearboxes, rotors, and rolling bearings [121], [122]; fault diagnosis in aircraft engines and their components [123]; prediction of faults in electrical power insulators [124]; classification of electrical insulator condition [125]; and many others. Thus, it is evident that fault diagnosis is a critical application across various engineering domains where ML techniques can be effectively employed. However, this chapter focuses specifically on ML-based fault diagnosis in power systems.

2.5 Workflow for Solving Problems Using Machine Learning

To tackle any problem using ML, it's essential to follow a set of steps to attain the desired solution [126]. Having an overview of the system under consideration helps determine which features directly influence the target variable. In cases where the relationship is unknown, conducting a collinearity test assists in gauging the extent of feature dependency on the target variable. Figure 2.3 depicts the generalized workflow for addressing problems using ML.



• Understanding the problem and analyzing the output
DEFINE THE OBJECTIVE OF PROBLEM
DATA ACQUISITION AND DATA PREPARATION
• Collecting the data and preparing in format acceptable to the algorithm
• Analyzing the model parameters significance on output
HYPOTHESIS TESTING
TRAIN AND TEST THE MODEL
• Splitting the data to train, validate and test the model
• Choosing correct performance evaluation
DECIDE EVALUATION MEASURE

Figure 2.3 The generalized workflow for addressing problems using ML.

Step 1: Define the objective of the problem: The crucial initial step in problem-solving is to thoroughly understand the problem and gain insight into the necessary inputs and expected outputs. Subsequently, it's important to ascertain whether the requisite labeled data for training is available. If not, alternative methods must be explored to address the problem using the available data. This approach aids in selecting the most suitable ML technique to achieve optimal outcomes [105].

Step 2: Data acquisition and data preparation: Once the problem has been thoroughly understood, the subsequent step involves acquiring data from reliable sources, such as real-world scenarios or standard repositories [127]. However, in cases where data is unavailable from these sources, synthetic data generation methods can be employed, such as modeling and simulation using standard software tools [97], [108]. The dataset utilized for training ML models may encompass various data types, as depicted in Figure 2.4. For example, electrical fault data may exhibit categorical, time series, or unstructured (images of voltage and current waveforms) characteristics [27].

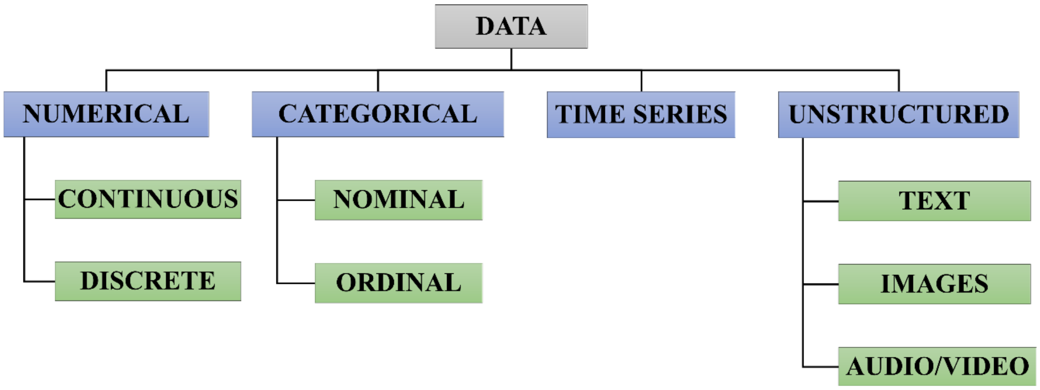


Figure 2.4 Diversity of data available for training various ML models.

For optimal performance, the dataset should be sufficiently large, as a larger size typically yields better results. Raw data typically isn't compatible with ML algorithms, necessitating preprocessing to convert it into a format suitable for input into the ML algorithm during the training phase. Data preprocessing encompasses several steps, including resampling, feature scaling, normalization, or standardization, flattening, feature extraction, dimensionality reduction or data compression, noise filtering, and data splitting [70], [76], [108], [113].

Data sourced from diverse channels may present various inadequacies, as depicted in Figure 2.5. To mitigate these issues, data re-sampling techniques are employed, which involve handling missing values and removing redundant or irrelevant data [27], [108]. Outliers are not removed immediately, as they may represent special cases [128]. Categorical datasets require special treatment, with nominal features represented using sparse matrices and ordinal features ranked accordingly [108].

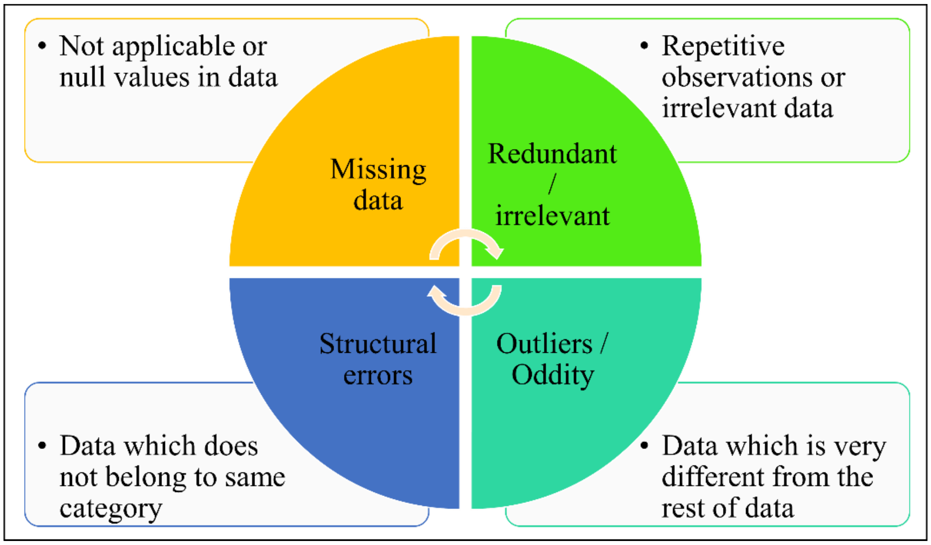

• Not applicable or null values in data
Missing data
Redundant / irrelevant
• Repetitive observations or irrelevant data
Structural errors
Outliers / Oddity
• Data which does not belong to same category
• Data which is very different from the rest of data

Figure 2.5 Various issues associated with data types.

Feature scaling is another crucial preprocessing step aimed at scaling all attributes in the dataset to a uniform scale. This can be achieved through normalization (min-max scaling) or standardization (centered at mean 0 with standard deviation 1). After scaling, correlations between attributes are computed to understand their contributions to the output or target [129]. Dimensionality reduction is essential for handling large datasets, as it reduces the number of features contributing to outputs, thereby lowering model complexity and preventing overfitting. This results in a more generalized model with reduced computation time and storage space [76], [127].

Feature extraction involves transforming the data using techniques such as the Fourier transform (FT) or discrete wavelet transform (DWT) [130]. These transformations provide insights into the system's behavior and are crucial for fault detection and classification [46], [124]. For instance, wavelet transforms preserve both time and frequency components, making them well-suited for fault diagnosis in power systems [131]. The work proposed in [38] used the empirical wavelet transform combined with the Hilbert transform, which enhanced fault classification and location estimation accuracy.

Principal component analysis (PCA) is also utilized for fault analysis, where it identifies unique fault signatures for classification and location estimation [132]. While PCA is commonly used for dimensionality reduction or data compression, in fault analysis, it serves as a tool for recognizing unique fault signatures. Additionally, [48] provides a comprehensive comparison of the advantages and disadvantages of different feature extraction techniques, serving as a valuable reference for selecting the most suitable feature extraction technique.

Step 3: Hypothesis testing: Hypothesis testing serves as a statistical method to assess the statistical significance of an ML model's performance. It begins with formulating hypotheses about the model's performance or the significance of input parameters. Typically, these hypotheses include (a) a null hypothesis assuming that the parametric attributes or input features have a certain relationship with the target variable or that one ML model will perform better than another; (b) an alternate hypothesis that contradicts the null hypothesis [96], [133].

Initially, the null hypothesis is assumed to be true, and a confidence level is set. However, for the significance of a parameter, the alternative hypothesis should be accepted [96], [133]. This determination is based on the p-value obtained from the hypothesis test. For the alternative hypothesis to be accepted, the p-value for the null hypothesis must be lower [134]. Conclusions regarding the validity of the hypotheses are drawn based on the results of the statistical test. If the results are statistically significant (i.e., the p-value is less than the confidence level), the null hypothesis is rejected in favor of the alternative hypothesis. For example, [135] employed the ANOVA test for sensitivity analysis to determine the significance of features in a dataset generated for high impedance fault (HIF) localization in transmission lines. The choice of a suitable statistical test depends on the formulated hypotheses. Some available tests include the Chi-square test [128], ANOVA (analysis of variance), f-test, and correlation coefficients [112], which are popular techniques for conducting hypothesis testing.

Step 4: Train and test the model: To train and test ML models effectively, the dataset is typically divided into three sections: training, validation, and testing sets, often allocated as 70%, 15%, and 15%, respectively. This division can be modified during the model training phase. The training data is utilized to train the ML algorithm, while the validation set is employed to assess the performance of the trained model [136]. If the model's performance meets the desired criteria, it is saved for future use. However, if the performance is unsatisfactory, the algorithm may undergo re-training with tuned hyperparameters, or an alternative algorithm may be selected. For enhanced evaluation of ML models, techniques like k-fold validation or iterated k-fold validation with shuffling can be employed to partition the input data into training, validation, and testing subsets [137]. Once the model demonstrates satisfactory performance on the validation dataset, it is considered ready to generate output on the testing dataset [136].

Step 5: Decide the evaluation measure: Determining the evaluation metric to assess the performance of a model is critical, and it depends on various factors such as the nature of the problem and the characteristics of the dataset's target variable. Understanding the problem type, whether it's classification, regression, or clustering, is essential for selecting the appropriate performance metric. Additionally, the specific objectives of the problem play a crucial role in the selection of metrics. For instance, in cancer detection, the primary goal may be to identify positive cases accurately rather than focusing solely on the overall accuracy of the model. Similarly, in a fraud detection system, false positives can lead to financial losses, while false negatives can undermine trust in the system. In cases where there is a data imbalance issue with classification problems, accuracy alone may not be reliable. Metrics like precision, recall, or F1-score are often more suitable in such scenarios [97], [138]. Therefore, choosing relevant metrics that align with the problem objectives is essential for effective performance evaluation. Commonly used evaluation measures include the confusion matrix, accuracy, precision, recall, F1-score, ROC curve, and AUC for classification problems. For regression tasks, metrics like mean squared error, mean absolute error, and R-squared are commonly employed [129], [138]–[141]. A study by [142] explores the preferred performance metrics for analyzing driving behavior, showcasing them through a bar chart in their study.

2.6 Performance Metrics

Performance metrics are indispensable in evaluating the effectiveness of classifiers and regression models. They provide insights into the accuracy and reliability of model outcomes. Here, we delve into commonly used metrics for both classification and regression models in machine learning.

2.6.1 Classification Metrics

Classification accuracy stands as the primary metric for assessing the performance of classification models, providing crucial insights into the reliability of model outcomes. To gain a deeper understanding of model performance, additional metrics such as the confusion matrix, sensitivity, specificity, precision, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUROC) can be utilized. Below, we delve into the explanations for these metrics.

2.6.1.1 Classification Accuracy

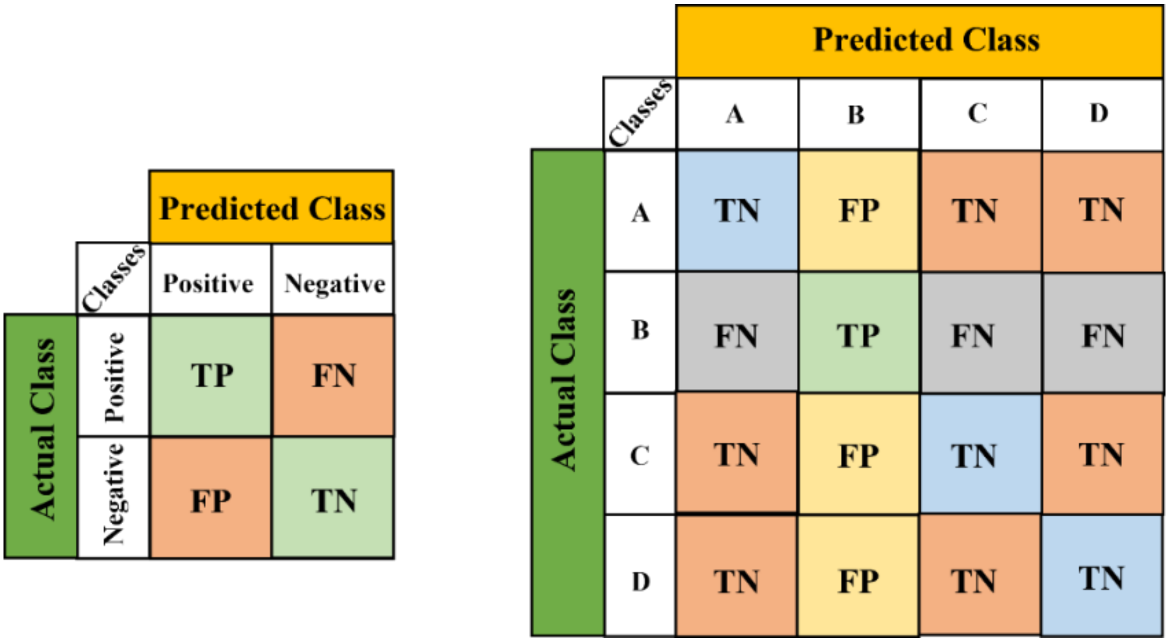
Classification accuracy serves as a widely utilized parameter for evaluating classifier performance, offering a fundamental overview of model effectiveness. However, in scenarios involving multi-class classification, where data imbalances and diverse class characteristics are present, classification accuracy alone may not suffice for comprehensively assessing classifier performance. In such cases, understanding the performance across individual classes becomes crucial. The classification accuracy for a model can be computed as given in equation 2.1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad \ldots(2.1)$$

Here, TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives. Hence, classification accuracy calculates the average accuracy percentage across all classes. For an ideal classifier, classification accuracy should reach 1 or 100%, indicating zero false positives and false negatives [143].

2.6.1.2 Confusion Matrix

The confusion matrix is an n*n matrix, where n represents the number of classes classified by the classifier. In cases of 100% classification accuracy, the confusion matrix takes the form of a diagonal matrix. However, when it deviates from a diagonal structure, each off-diagonal element signifies incorrectly classified samples for that class. The confusion matrix can be binary or multi-class, as depicted in Figure 2.6. Each row of the confusion matrix provides details on the total number of samples belonging to that particular class when the sum of samples in that row is computed [143].



		Predicted Class	
	Classes	Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

		Predicted Class			
	Classes	A	B	C	D
Actual Class	A	TN	FP	TN	TN
	B	FN	TP	FN	FN
	C	TN	FP	TN	TN
	D	TN	FP	TN	TN

Figure 2.6 Confusion matrix for binary and multi-class classification.

2.6.1.3 Sensitivity / Recall

Sensitivity, also known as recall, measures the ability of a classifier to correctly identify all relevant instances, i.e., correctly classify a sample belonging to a positive class out of all instances. It is calculated as the ratio of true positives (TP) to the total number of samples in that class [144]. The calculation of recall for class B in Figure 2.6 is given in equation 2.2.

$$Sensitivity = \frac{TP}{TP+FN} \qquad \ldots(2.2)$$

2.6.1.4 Specificity

Specificity, also known as the true negative rate, is a metric used to evaluate a classifier by assessing its ability to accurately identify negative instances. It represents the model's capacity to correctly distinguish samples not belonging to the evaluated class. Specifically, it is calculated as the ratio of true negatives to all negatively classified samples for that class. The computation of specificity for class B in Figure 2.6 is outlined using equation 2.3.

$$Specificity = \frac{TN}{TN+FN} \qquad \ldots(2.3)$$

In essence, specificity indicates the percentage of negative instances correctly classified as such out of all instances truly belonging to the negative class. A higher specificity value indicates better accuracy in identifying negative instances, while a lower specificity value suggests a potential misclassification of negative instances as positive.

2.6.1.5 Precision

The precision metric assesses the accuracy of positive predictions generated by a classifier, indicating the correctness of classifying a specific class. It quantifies the proportion of true positive predictions out of all positive predictions, regardless of their accuracy. Precision is computed as the ratio of true positives to all positively classified samples for that class [144]. The precision for class B in Figure 2.6 is determined by equation 2.4.

$$Precision = \frac{TP}{TP+FP} \quad \ldots(2.4)$$

A higher precision value signifies that the classifier generates fewer false positive predictions, whereas a lower precision value indicates a higher incidence of false positives among the positive predictions made by the classifier.

2.6.1.6 Area Under Receiver Operating Characteristic (AUROC) Curve

The ROC curve is a plot of true positive rate (TPR) versus false positive rate (FPR), where TPR is a measure of sensitivity, which is discussed above, and FPR is a measure of the ratio of FP for a specific class to the total number of samples not belonging to that class, or 1 minus specificity. It is a performance metric commonly used to evaluate the performance of binary classification models. The ideal value of the AUROC curve must be 1 [145].

$$TPR = \frac{TP}{TP + FN} \quad \ldots(2.5)$$

$$FPR = \frac{FP}{FP + TN} \quad \ldots(2.6)$$

2.6.1.7 F1-score

The F1 score serves as a comprehensive measure of a classifier's accuracy, striking a balance between precision and recall. This balance makes it particularly valuable when dealing with imbalanced class distributions. Calculated as the harmonic mean of precision and recall as given in equation 2.7, a perfect F1-score of 1 indicates flawless precision and recall, signifying accurate positive predictions with no false positives or false negatives. Conversely, a score of 0 suggests poor performance, where either precision or recall (or both) is negligible. Notably, an F1-score of 1 implies 100% precision and recall, concisely summarizing the classifier's performance in terms of both metrics [143].

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad \ldots(2.7)$$

2.6.2 Regression metrics

The commonly used regression metrics are mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), mean absolute percentage error (MAPE), and r-squared (R2-score). One or more of these can be used for comparing the performance of various ML regression models.

2.6.2.1 Mean Absolute Error (MAE)

It is the simplest metric that computes the error, i.e., finding the average of the absolute difference between the actual value and predicted value of the model output for all test samples as given in equation 2.8. The model showing the least MAE is the best-performing model out of all models used on a particular training and testing dataset [146].

$$MAE = \frac{1}{N} \sum |y - \hat{y}| \qquad \ldots(2.8)$$

2.6.2.2 Mean Squared Error (MSE)

In this metric, the square of difference between the actual value and predicted value for all test samples is computed and then the summation of all squared error, and mean is evaluated to get MSE for the trained model as given in equation 2.9. For this metric also the model showing the least MSE is the best-performing model out of all models [146].

$$MSE = \frac{1}{N} \sum (y - \hat{y})^2 \qquad \ldots(2.9)$$

2.6.2.3 Root Mean Squared Error (RMSE)

RMSE provides a single measure of the model's prediction error, with lower values indicating better model performance. It is calculated like MSE; the only difference is that it is the square root of MSE. This makes the error unit in the same range as the target variable, making it easy to relate in terms of the target variable [147].

$$RMSE = \sqrt{\frac{1}{N}\sum(y - \hat{y})^2} \qquad \dots(2.10)$$

2.6.2.4 Mean Absolute Percentage Error (MAPE)

Is a measure of the error between the actual value and the predicted value as a percentage. Thus, it is calculated similarly to MAE, with the only difference being that each error term is divided by that term's actual value to express that error as a percentage. Then averaging all the error percentages to get the MAPE of a model [146].

$$MAPE = \frac{1}{N}\sum\left|\frac{y - \hat{y}}{y}\right| \qquad \dots(2.11)$$

2.6.2.5 R Squared (R^2) Score

It quantifies the degree to which the regression line of the model outperforms the mean line across the entire dataset. The mean line is a horizontal line acting as a baseline. If the regression model fits worse than the baseline, then the R^2-score comes out to be negative. It speaks about the goodness of fit of the model. The value of the R^2-score lies between 0 and 1, the closer the value is to 1, the better the estimation result [60].

$$R^2 Score = 1 - \frac{Squared\ sum\ error\ of\ regression\ line}{Squared\ sum\ error\ of\ mean\ line} \qquad \dots(2.12)$$

$$R^2 Score = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} \qquad \dots(2.13)$$

2.7 Dimensionality Reduction Techniques

Dimensionality reduction techniques are essential methods used to decrease the number of features or variables in a dataset while retaining its vital information. These methods play a critical role in managing high-dimensional data, enhancing computational efficiency, and mitigating challenges such as the curse of dimensionality. Commonly utilized dimensionality reduction techniques include principal component analysis (PCA), kernel PCA, linear discriminant analysis (LDA), autoencoders, and singular value decomposition.

2.7.1 Principal Component Analysis

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of multidimensional data by extracting effective features. However, its linearity assumption may limit its effectiveness for some non-linear processes. Kernel PCA (KPCA) addresses this limitation by projecting the data into a high-dimensional kernel space using nonlinear functions, capturing non-linear relationships. Nevertheless, KPCA can be less effective for large datasets due to computational constraints and may not perform well in real-life systems with numerous samples [148].

In PCA, the parameter n_components determine the number of principal components onto which the data will be projected during dimensionality reduction. Specifying n_components as an integer value retains a fixed number of principal components, controlling the output dimensionality. Alternatively, setting it to 'mle' (maximum likelihood estimation) allows PCA to automatically select the number of components based on likelihood estimation. Additionally, specifying a float between 0 and 1 indicates the proportion of variance that the retained components should capture, offering flexibility in dimensionality reduction. For example, setting n_components to 0.95 will retain principal components until they collectively capture 95% of the variance in the data [149].

PCA has been widely employed for fault detection due to its effectiveness in separating data information into significant and residual subspaces. However, its linear nature may limit its applicability to inherently nonlinear processes, where the relationship between variables is nonlinear. To address this limitation, various nonlinear extensions of PCA have been proposed for fault detection, categorized into three main approaches.

The first category involves kernel methods, with kernel PCA (KPCA) being a prominent choice. KPCA maps data into a high-dimensional nonlinear feature space, enabling linear PCA to be applied effectively. However, KPCA requires eigen decomposition (ED) on the kernel Gram matrix, whose size is the square of the number of data points, which can be computationally intensive for large datasets. Additionally, determining the kernel and associated parameters in advance can be challenging. The second category comprises linear approximation techniques for nonlinear processes. In this approach, local linear models are constructed and integrated using Bayesian inference. While linear approximation is straightforward, it may struggle to handle strong nonlinearities in the process. The third category consists of neural-network-based models, including robust autoencoders (RAE) and auto-associative neural networks. These models train a feedforward neural network to perform identity encoding, with a bottleneck layer for feature extraction. While encoding can address nonlinearities, it does not incorporate the orthogonal constraints used in PCA [150].

2.7.2 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a supervised learning technique, in contrast to PCA, which is unsupervised. In LDA, interclass separation is achieved by replacing the total covariance matrix in PCA with a general separability measure, resulting in the discovery of m eigenvectors of the scatter matrix [151]. In LDA, n_components denote the number of dimensions or components in the reduced feature space. Setting n_components as an integer value indicates the exact number of dimensions the data will be projected onto. For instance, n_components set to 2, project the data onto a two-dimensional space. If n_components is not specified, LDA retains a maximum of min (n_classes - 1, n_features) components, where n_classes represent the number of classes and n_features denote the number of features.

LDA aims to identify linear combinations of features that optimally separate the classes. The maximum number of linear discriminants is limited to the number of classes minus one. Setting n_components greater than the number of classes may lead to overfitting and introduce noise into the transformed data as it attempts to create discriminants unsupported by the underlying class structure.

In Sarlak and Shahrtash work LDA was utilized for feature extraction of high impedance faults (HIF) in the distribution system using a supervised approach, compared with unsupervised PCA [151]. Additionally, Sarwar et al. presented a HIF detection and localization model for the IEEE 13-node bus distribution system, employing PCA for feature extraction and Fisher Discriminant Analysis (FDA) for further dimensionality reduction to enhance fault localization accuracy [50].

2.8 Unsupervised Learning

Unsupervised learning entails training ML models without class labels, aiming to uncover hidden patterns or inherent structures within the data. It proves valuable in extracting insights from datasets comprised solely of input variables, devoid of labeled responses. Consequently, unsupervised learning finds application in scenarios where only input data is available, offering valuable insights into the dataset [95], [96], [133]. This data-driven approach facilitates the exploration of the data, leading to the formation of clusters within the dataset [97], [106]. These clusters delineate distinct patterns and can manifest as either hard clustering or soft clustering. Notable applications of clustering include market research, gene sequence analysis, and object recognition [137], [138], with clusters adept at uncovering uniformities within the data [27].

2.8.1 Hard Clustering

Hard clustering refers to clustering wherein each data point belongs to only one cluster, i.e. there is a clear distinction among different clusters without any overlapping [95], [96]. Some of the available hard clustering techniques are K-means, K-medoids, and hierarchical clustering [95], [96]. Figure 2.7 illustrates the distinctive feature of the mentioned hard clustering technique.

2.8.1.1 K-means

K-means clustering offers rapid clustering for large datasets, ideal when the desired number of clusters is predetermined. It divides sample data into k clusters, with each point's appropriateness in a cluster determined by its distance from the cluster center [95], [96]. Notably, cluster centers do not coincide with data points [95].

In [67], K-means clustering is utilized to predict missing data, enhancing overall accuracy. Similarly, [152] employs synchrophasor data from PMUs to gauge voltage fluctuations, employing K-means clustering to categorize power quality levels. While simple K-means detects network faults, the two-stage K-means optimization classifies power quality levels in [152]. Additionally, [27] leverages K-means clustering (k = 7) to detect and classify localized faults (permanent faults leading to prolonged power outages).

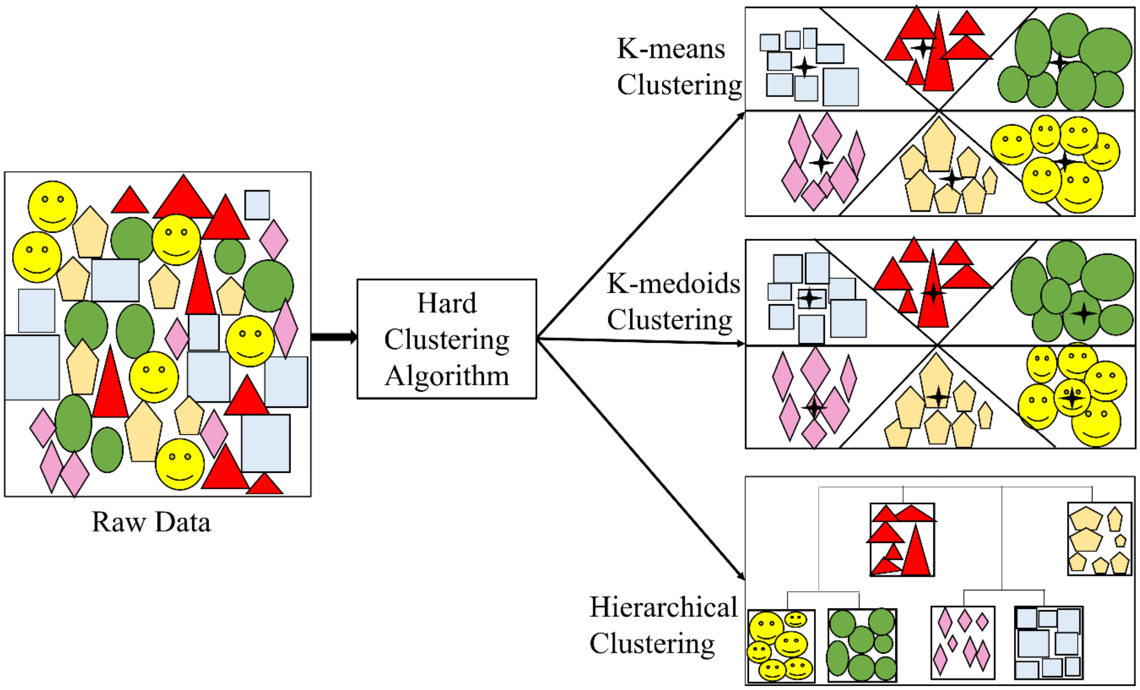


Figure 2.7 Illustration of the hard clustering technique.

2.8.1.2 K-medoids

K-medoid clustering is similar to k-means clustering; the only difference lies in the positioning of cluster centers, as cluster centers coincide with data points [95]. It is preferable for clustering large categorical datasets, particularly ordinal data, and when the cluster count is predetermined [95], [96]. Notably, for non-metric datasets, clusters are delineated by minimizing the sum of distances (MinSOD) [138], [153].

In [154], a k-medoid-based pick-up current selection module for k-setting groups in relays is proposed for the IEEE 13 node test feeder, where pickup current values are sourced from directional overcurrent relays. Employing the partitioning around medoids (PAM) algorithm, the scheme harnesses k-medoids clustering and is more robust to outliers than k-means clustering [154].

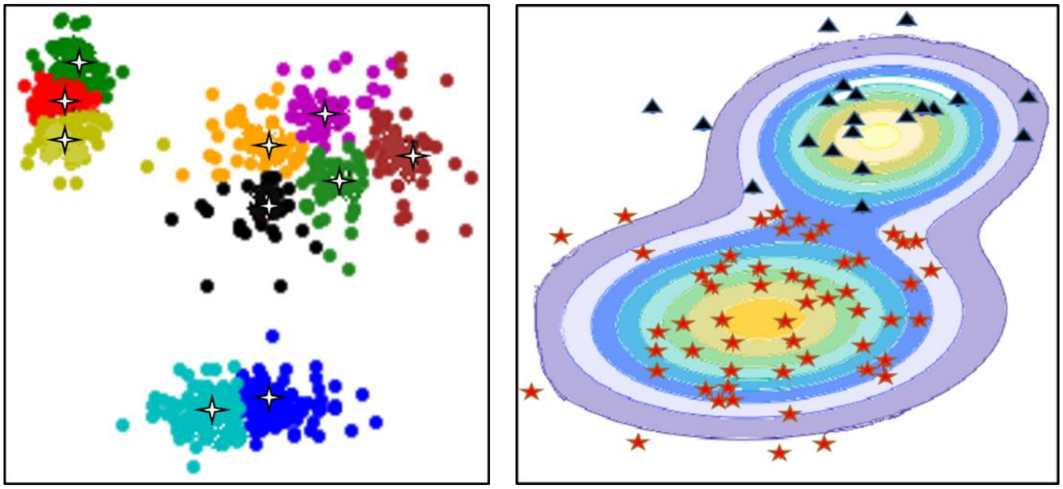
2.8.1.3 Hierarchical clustering

Hierarchical clustering builds a tree-like structure to group similar data points. It's handy when you're not sure how many clusters you need and when the data is complex and needs organizing for better understanding [137].

In a study described in [155], a hierarchical clustering method was used to pinpoint disturbances in the IEEE-39 bus system. This approach was not only computationally simpler but also effective in identifying various sources of voltage fluctuations. By analyzing fault signals and creating a matrix based on Euclidean distances, pattern recognition was employed to extract information. This matrix was then used for hierarchical clustering to estimate fault locations and the LDA for fault classification [155]. Another study, detailed in [156], utilized PCA for initial clustering to categorize fault events in power systems. The severity of these events was determined by evaluating the compactness of the clusters. Subsequently, hierarchical clustering was employed for accurate fault classification. To identify faulty sections, hierarchical clustering of PMU data near the affected area was conducted based on cluster compactness and Euclidean distance estimation [156]. Also, hierarchical clustering combined with incremental learning has been used to classify 11 fault types in [157]. This involved clustering Fischer-Rao-amplitude-phase distance data and fault features were extracted from clusters using Karcher means, preserving signal shape across different fault scenarios.

2.8.2 Soft Clustering

Soft clustering involves clustering techniques where overlapping is allowed, meaning that data points can belong to one or more clusters without clear distinctions. Common soft clustering methods include fuzzy C-means and Gaussian mixture models [96]. Figure 2.8 provides a graphical representation of soft clustering techniques.



(a) (b)

Figure 2.8 Graphical illustration of the soft clustering techniques (a) Fuzzy C-mean and (b) Gaussian Mixture Model.

2.8.2.1 Fuzzy C-mean

Fuzzy C-mean clustering is a technique where data points can belong to more than one cluster, as shown in Figure 2.8(a), differing from k-means clustering in its allowance for cluster overlap and fuzziness in cluster boundaries [95], [96].

In a study by [158], fuzzy C-means clustering was utilized to establish four cluster centers along the faulted line section. Additionally, a K-nearest neighbors (KNN) model was applied to form the cluster center around the faulted section for fault location identification. Another approach for fault detection and classification, outlined in [159], utilized fuzzy adaptive resonance theory (ART) for cluster formation based on fault voltage and current patterns. Subsequently, KNN was employed for fault classification, demonstrating improved performance compared to a simple KNN classifier.

2.8.2.2 Gaussian mixture model

The Gaussian mixture model clustering method forms clusters of data points based on different multivariate normal distributions with associated probabilities [95]. This approach is favored when data points may belong to multiple clusters of varying sizes and correlation structures [95], as depicted in Figure 2.8(b).

In a study detailed in [82], a Gaussian hidden Markov model (HMM) was employed for fault detection, classification, and localization in smart grid systems using features such as frequency and voltage variations. Two HMMs were trained for detecting faulty and normal grid operation based on frequency signal features, while several HMMs were trained for different fault scenarios for fault classification. Fault types such as generator faults, line trips, and load losses were identified using matching pursuit decomposition (MPD) of voltage signal features. The extracted MPD features were then clustered to resemble a fault contour map, aiding in the localization of the faulted part of the smart grid [82]. Another model, detailed in [134], utilized a Gaussian Markov random field (GMRF) for fault detection and locating the faulted subfield. Buses' phasor angles were treated as random variables, and a conditional correlation matrix was computed to analyze their dependency graph, providing insights into fault detection and identifying faulty subfields. The HMM approach proved effective for multivariate samples and different fault scenarios, outperforming other supervised learning classification algorithms due to the dynamic nature of power systems [51]. Hence, the HMM algorithm can perform both clustering and classification tasks within its framework.

2.9 Supervised Learning

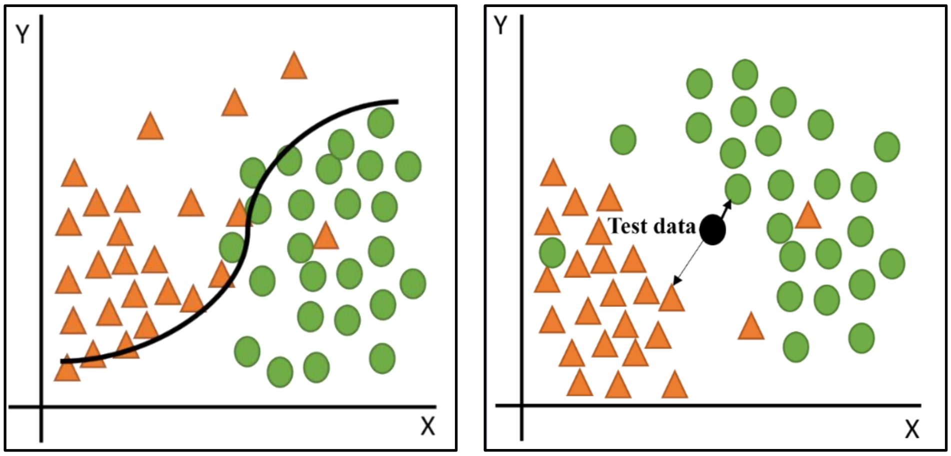
Supervised learning is similar to learning with a teacher, aiming to minimize errors in performance [95], [133]. This method is suitable when both input features and corresponding labels for each data point are provided for model training, enabling the model to predict responses for new data [95], [96], [133]. Supervised learning performs either classification or regression tasks based on specific requirements [97]. Prediction accuracy for unseen data improves with a well-populated dataset containing prototypes for such cases and as the model learns more deeply [160].

2.9.1 Classification Models

Classification models categorize input data into distinct classes or categories, providing a discrete response based on available class labels in the dataset [95], [96]. The following subsections explore different classification algorithms, along with details of their applications in power system fault diagnosis found in the literature.

2.9.1.1 Logistic Regression

Logistic regression is a statistical method used for binary classification tasks. The model evaluates the probability of a given input belonging to a particular category. Despite its name, logistic regression is primarily used for classification, not regression. It estimates the probability that an instance belongs to a particular class using a logistic function, which maps any real-valued input to a value between 0 and 1 [161], [162]. Figure 2.9 (a) graphically represents the logistic regression (LR) model.



(a) (b)

Figure 2.9 Graphical illustration of (a) Logistic Regression and (b) K-Nearest Neighbor model.

In studies [140], [162], NN and LR were employed for fault root cause identification. The correct classification rate (CCR), which measures the classifier's performance by calculating the proportion of correctly classified inputs, has been used as a performance metric in the proposed study [162]. CCR for NN was slightly higher than LR, although the difference was almost negligible [162]. NN exhibited a better true positive rate compared to LR, while LR performed better in terms of true negative rate in most cases. Despite LR's speed, NN's geometric mean outperformed LR, suggesting LR is biased towards majority data and has inferior accuracy for minority classes [162]. In another study [163], LR was utilized for logistic odd fitting, acting as a nonlinear classifier. LR accommodates both categorical and continuous features without needing a specific distribution or variance [163]. However, LR entails higher computational costs due to its use of maximum likelihood estimation with large datasets [163]. LR was favored over discriminant analysis (DA) in [163] due to the prevalence of categorical features in the outage dataset. Both LR and DA exhibited similar performance and good predictability. In a separate work [40], the issue of imbalanced data in ML and data mining was addressed. Since performance metrics like CCR, geometric mean, true positive rate, and true negative rate can be affected by imbalanced data, ROC curves were evaluated to analyze distribution network fault diagnosis performance as it is unaffected by data imbalance [40]. Five ML classification techniques LR, artificial neural network (ANN), SVM, KNN, and artificial immune recognition system (AIRS) were compared. LR and ANN performed well, whereas SVM performed comparably for large threshold values but poorly for small ones. KNN showed similar performance to LR, ANN, and SVM at a fixed threshold value of 0.5, while AIRS exhibited the best performance according to the ROC curve.

2.9.1.2 K-nearest neighbor (KNN)

The KNN is a non-parametric method that classifies samples based on their proximity to neighbors in the feature space [95]. The choice of nearest neighbor is determined by metrics like Euclidean, Minkowski, or Manhattan distance between feature vectors [95], [96]. KNN is employed in scenarios where memory usage and prediction speed are not limiting factors [95]. Its core components include distance calculation, the training dataset, a similarity metric, and the value of k [164]. In a study referenced as [164], the Mahalanobis distance function was utilized for fault detection in HVDC systems. The graphical representation of the KNN model is provided in Figure 2.9(b).

In a comparative analysis conducted in [41], four classifiers DT, radial basis functions neural network (RBFNN), NB, and KNN were evaluated for microgrid fault detection. KNN demonstrated the fastest classification speed for both training and testing data in a noise-free environment, while NB outperformed others for testing data in a noisy environment. Another study [165] introduced a KNN-based fault classifier that, unlike relays, does not require voltage and current values and relies solely on voltage and current waveforms for fault detection in power systems. This approach effectively captures unusual patterns in current waveforms during faults, serving as a reliable diagnostic tool [165]. Additionally, a feature extraction technique combining KNN and cross-correlogram analysis was presented in [42], achieving higher accuracy (99.67%) compared to other methods like discrete wavelet transforms (DWT) + artificial neural network (ANN), wavelet transform, SVM, and multi-resolution S-transform + probabilistic neural network (PNN). The simplicity of the KNN classifier, along with its lower computational time and memory requirements compared to ANN, SVM, and other techniques, was highlighted.

2.9.1.3 Support Vector Machine (SVM)

Support vector machine operates by establishing a decision boundary, known as a hyperplane, which effectively separates samples of one class from others [95], [96]. In cases of misclassification, SVM employs a loss function to penalize errors, offering an effective strategy to mitigate incorrect classifications. Utilizing kernel functions, SVM transforms non-linearly separable samples into higher dimensional spaces where they are more likely to become linearly separable [49], [164]–[168]. The transformation of SVM from lower dimensions to higher dimensions to handle non-linearly separable datasets has been illustrated graphically in Figure 2.10. SVM embraces structural risk minimization and adheres to statistical learning principles based on the Vapnik-Chervonenkis (VC) dimension rule. This approach enables SVM to effectively address high-dimensional pattern recognition and non-linear problems, rendering it suitable for fault diagnosis applications [30], [126], [140], [166], [169], [170].

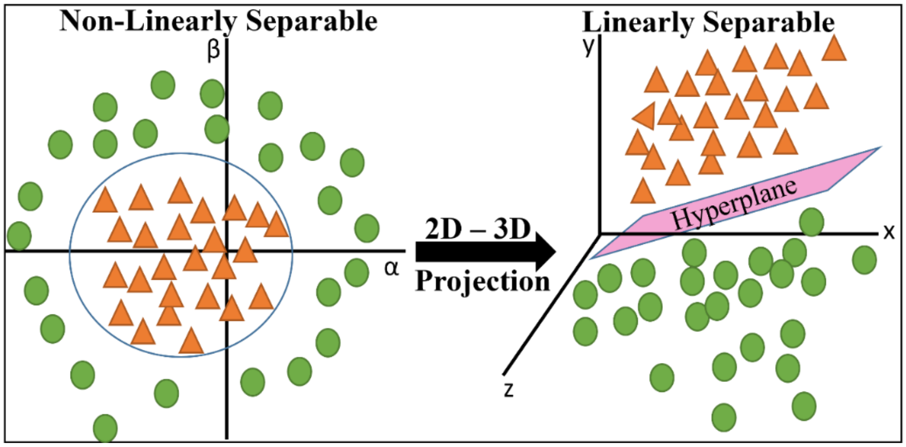


Figure 2.10 Graphical illustration of Support Vector Machine model.

In [167], post-fault voltage and current normalized wavelet energy were utilized as input for the SVM classifier, with DWT to extract fault features from recorded voltage values.

Four SVMs were deployed for individual phase and ground fault detection, employing a Gaussian radial basis function (RBF) kernel for transformation. The authors of [171] utilized Gaussian RBF and polynomial kernel functions for fault classification, noting superior fault classification accuracy with the RBF kernel compared to polynomial kernel SVMs. Similarly, in [172], a Gaussian kernel function-based SVM achieved higher classification accuracy for fault classification. This method utilized DWT with Daubechies (db7) as the mother wavelet to extract normalized wavelet energy from two cycles of transient fault current waveforms, resulting in a robust fault diagnosis accuracy of 98.5% for the presented DWT-SVM approach. Another method proposed in [173] utilized multi-class SVM, using only half a cycle of the post-fault current waveform as input. Preprocessing of post-fault current involved wavelet transformation with Daubechies (db5) as the mother wavelet. Employing one-versus-one and one-versus-rest SVM classification strategies with the RBF kernel, the proposed method achieved an accuracy of 98.8% for fault diagnosis and remained robust against changes in network conditions. A novel quarter sphere SVM (QSSVM) was introduced in [77], incorporating temporal attribute correlation and attribute correlation (TA-QSSVM and A-QSSVM, respectively). TA-QSSVM achieved fault classification accuracy close to 100%, while A-QSSVM, emphasizing attribute correlation, attained 99% classification accuracy for automatic fault classification. Both methods are suitable for online smart metering applications, addressing labeled and unlabeled data, respectively.

SVM stands out among classification techniques due to its solid mathematical foundation [96], [174]. By optimizing structural risk rather than training error, SVM reduces the risk of classification error for future data, making it suitable for sparse training data applications [96], [174]. An extended Kalman filter (EKF) based SVM fault detection scheme was proposed in [49], where EKF estimated changes in magnitude and phase of current signals. This information was fed to a Gaussian kernel based SVM for classifying lines into faulty and non-faulty conditions.

In a comparative study [175], wavelet transform-based SVM (WT-SVM) outperformed wavelet transform-based extreme learning machine (WT-ELM) for fault classification and localization, achieving a classification accuracy of 99.1%. The importance of preprocessing was highlighted, with WT-SVM demonstrating better performance compared to WT-ELM when utilizing DWT. Both techniques proved robust to variations in parameters such as source impedance, fault distance, and pre-fault power level. In [176], a fault feature extraction technique called the determinant function is introduced, offering the advantages of lower memory space and computational complexity. The performance of SVM with this feature extraction method was evaluated for classification. However, the SVM classifier struggled to differentiate between LLL and LLLG faults, treating them similarly. Another approach, detailed in [177], utilizes wavelet-based preprocessing to train SVM for fault detection and classification. This method involves computing the RMS values of pre-fault and post-fault three-phase line current and voltage waveforms.

A novel approach for fault detection, classification, and localization was presented in [178], using a two-stage finite impulse response (FIR) filter with four SVMs for fault detection and classification. The model also incorporated eleven support vector regressors (SVRs) for fault localization on a 50km transmission line. In [179], the SVM classifier has been used for fault detection, classification, and localization, relying on data collected by PMUs utilizing bus voltage and phase angle data as input. The SVM classifier required 100-time steps of data, giving only 88% accuracy due to its limitation in capturing temporal information, while on the same data, the recurrent neural network (RNN) with long-short-term memory (LSTM) improved classification accuracy to around 95%. In a similar vein, the study described in [83] leverages positive and zero sequences voltage data from PMUs for fault detection and classification, employing the RBF kernel SVM. Additionally, a six-phase fault detection and classification approach utilizing DWT is outlined in [131]. Here, post-fault voltage and current signals undergo preprocessing using a low-pass Butterworth filter. Another fault location prediction method is presented in [170], employing hierarchical SVM with DWT (db4) to process zero sequence current in fault scenarios. This hierarchical approach yields higher faulty section identification accuracy compared to simple SVM methods. Furthermore, fault location estimation using traveling wave techniques and faulty section identification using SVM with DWT is proposed in [180], [181]. Similarly, an SVM-based fault identification scheme for overhead lines or underground cables and a faulty section identification scheme for hybrid transmission lines is detailed in [37], where SVM identifies faulty transmission sections while Bewley diagrams give precise fault location estimation.

Other fault detection methods include the Kullback-Leibler Divergence (KLD) method in [54] for incipient fault detection via SVM and Support Vector Data Description (SVDD) in [16] for fault detection in distribution systems with varying penetration levels of distributed energy resources (DERs). The study in [182] proposed an SVM scheme for identifying grid faults or islanding in low voltage distribution grids connected to photovoltaic (PV) modules, ensuring reliable islanding and grid fault detection.

2.9.1.4 Neural Network (NN)

The neural network (NN) is a computational model inspired by the human nervous system, characterized by its massively parallel distributed connections capable of learning from experience and making predictions [165], [183]. Operating in two phases, the NN first learns from training data by adjusting connection weights to match patterns [95]. During recall, the trained network generates responses for testing data based on these learned parameters [162]. However, NN training can be time-consuming, particularly during cross-validation, and requires substantial computational resources for training, validation, and testing [162]. Figure 2.11 illustrates the basic structure of a three-layer NN. The NN depends on error correction for fine-tuning the weights of nodes, which is typically achieved using backpropagation or gradient descent algorithms. Various gradient descent algorithm variants, such as batch gradient descent, mini-batch gradient descent, and stochastic gradient descent, are available for this purpose [129].

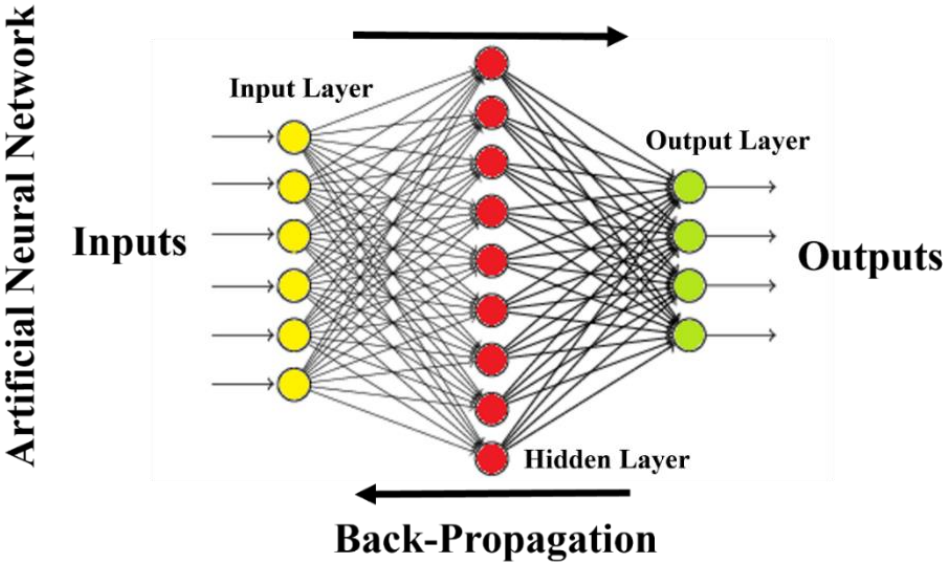


Figure 2.11 Schematic diagram of Neural Network model.

In work [184], the initial step involves utilizing K-means clustering for data preprocessing, resulting in labeled data. Subsequently, a priori association rule is applied to extract features highly correlated to the target fault type. Fault classification prediction is then conducted using the stochastic gradient descent (SGD) algorithm. This three-layered fault classification module yields higher accuracy compared to the simple SGD algorithm. Moreover, the training speed of the SGD algorithm surpasses that of simple NN, SVM, and LR. Another variant of NN, known as adaptive resonance theory NN (ART NN), encompasses both unsupervised and supervised learning capabilities [185]. Each ART NN comprises three sets of neurons: the input processing unit, the cluster unit, and the similarity check unit. Various ART NN models, including ART1, ART2, ART2-A, fuzzy ART, and fuzzy ARTMAP, are employed for high impedance fault (HIF) detection and classification using TT (time-time) transform schemes [185]. Additionally, an NN-based scheme for fault detection and classification in transmission lines using oscillographic data has been proposed for a Brazilian utility company [92].

NN demonstrates satisfactory performance for imbalanced data, as its accuracy remains unaffected by the individual classification of majority and minority classes [40]. An adaptive protection scheme for double-circuit transmission lines using ANN is proposed in [186], focusing on detecting and classifying SLG faults. The model utilizes the hyperbolic activation function and the Lenvenberg–Marquardt (LM) algorithm, which yield better results compared to the back-propagation algorithm [186]. The selection of the number of hidden layers in the NN model plays a crucial role in optimizing its generalization ability, which is often determined heuristically. Moreover, selecting the activation function for hidden layer neurons is essential [129]. Utilizing data collected from PMUs at WAMS enhances fault detection and classification accuracy [3]. The data is trained on SVM and ANN classifiers, with ANN exhibiting superior performance and lower storage requirements.

Another work used an artificial immune recognition system (AIRS), inspired by the human immune system, which serves as a supervised learning algorithm. AIRS outperforms ANN in most cases when the threshold value is suitably selected based on the data and ROC curve [40]. In [187], a fault detection, classification, and localization scheme for overhead transmission lines using the S-transform is presented. S-transform generates S-matrices from faulty current signals, facilitating fault current peak value and faulty phase angle determination for classification and fault location estimation. A probabilistic NN (PNN) training methodology employing the LM algorithm is utilized for fault classification and location estimation. For HVDC systems, fault detection, classification, and localization techniques utilizing traveling wave and pattern recognition-based machine learning concepts are explored in [188].

Extreme Learning Machine: Another emerging NN technique is the extreme learning machine (ELM), which consists of a single hidden layer feed-forward neural network [189]. Unlike other NN techniques, the input weights and biases of the hidden layer nodes in ELM are randomly generated without tuning [29]. This simplicity in implementation, coupled with faster learning speed, superior generalization performance, minimal human intervention, and freedom from issues such as local minima and overfitting (common in gradient descent-based learning), makes ELM advantageous [29], [160], [189]–[191].

ELM-based fault classification in series-compensated transmission lines, utilizing DWT (with db2 as the mother wavelet) for data preprocessing, has been demonstrated in [189]. In

[46], a Hilbert Huang transform (HHT) based feature extraction, fault detection, and classification scheme employed ELM, NB, and SVM classifiers, with ELM exhibiting the best performance and NB slightly outperforming SVM. Moreover, [190] utilizes ELM for fault location estimation in HVDC lines, employing DWT for feature extraction and signal processing. In [191], DWT-ELM is introduced for fault detection, classification, and location estimation in a series-compensated 400 kV-300 km transmission line, outperforming DWT-SVR. Additionally, [29] presents modified versions of ELM: the summation-wavelet extreme learning machine (SW-ELM) and the summation-Gaussian extreme learning machine (SG-ELM). SW-ELM incorporates feature extraction capability, while SG-ELM extends SW-ELM with self-learning ability, eliminating the need for feature extraction. The combination of ELM with wavelet transforms in SW-ELM proves excellent for fault diagnosis tasks, allowing efficient simultaneous fault classification and location estimation [29].

Convolutional Neural Network: In [30], a convolutional neural network (CNN) based deep learning approach utilizes continuous wavelet transform for feature extraction from time-frequency fault signals, transforming them into grayscale images.

Similarly, [192] applies CNN to simulated fault data samples from a 220 kV, 100 km long transmission line. This scheme involves splitting the data into time windows, and then merging them to increase the dataset size and enhance classification accuracy. CNN is popular for its specialization in image recognition applications [30], [192]. Recent applications of CNN include fault detection and classification for distribution networks integrated with DGs [17]. The integration of DGs alters network current flow direction and fault current levels, posing challenges for conventional relaying methods, thus motivating the use of CNN in such scenarios [17]. In [52], an HHT feature extraction-based fault classification model constructs a time-frequency energy matrix for digital fault images, utilized by CNN for fault classification. Additionally, [36] proposes an advanced CNN-based faulty line selection scheme utilizing adaptive CNN (ACNN), based on the TensorFlow framework. Compared to Deep Belief Networks (DBN), ACNN demonstrates superior classification accuracy and shorter training times. Fault localization in this model relies on the two-terminal method, relying on negative sequence voltage and current components at both ends of the faulted line. Another faulty line identification scheme for IEEE 39 and 68 Bus systems proposed in [193] employs a CNN classifier using only bus voltages. The author explores optimal PMU placement to validate model accuracy and further refine localization estimation. In [34], a graph convolutional network (GCN) model is introduced for faulty bus identification based fault localization. Results indicate improved faulty bus identification compared to SVM, RF, and fully connected neural networks (FCNN). Moreover, the proposed GCN model demonstrates robustness to measurement errors and adaptability to distribution system dynamics.

Recurrent Neural Network: In [35], a novel fault localization model employing a bi-directional gated recurrent unit (Bi-GRU) is proposed. This model represents an upgrade over traditional recurrent neural networks (RNNs), addressing issues such as gradient disappearance and explosion, commonly encountered when dealing with long sequence data. Unlike conventional methods, this model eliminates the necessity for feature extraction. Instead, it incorporates an attention mechanism to meticulously monitor time-series sequence data both pre- and post-fault, capturing subtle changes that ultimately contribute to more precise fault location estimation.

Deep Neural Network: In [84], an intelligent fault detection approach tailored for microgrid applications is introduced. This method employs DWT for preprocessing the branch's current data acquired from protective relays. However, the accuracy and speed of fault detection, classification, and localization significantly improve with the implementation of a deep neural network (DNN) for fault diagnosis. Notably, the entire fault detection process can be executed in real-time. Figure 2.12 provides a structural depiction of the DNN.

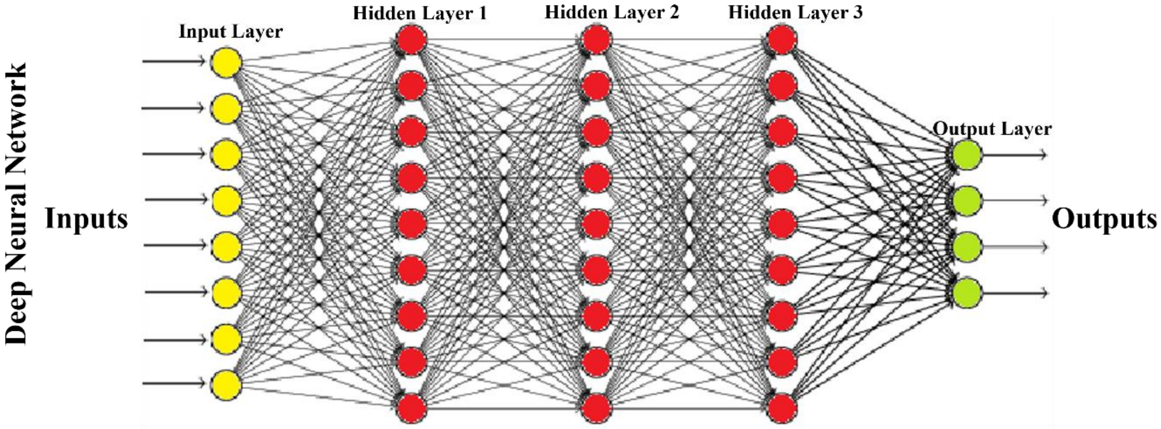


Figure 2.12 Schematic diagram of Deep Neural Network model.

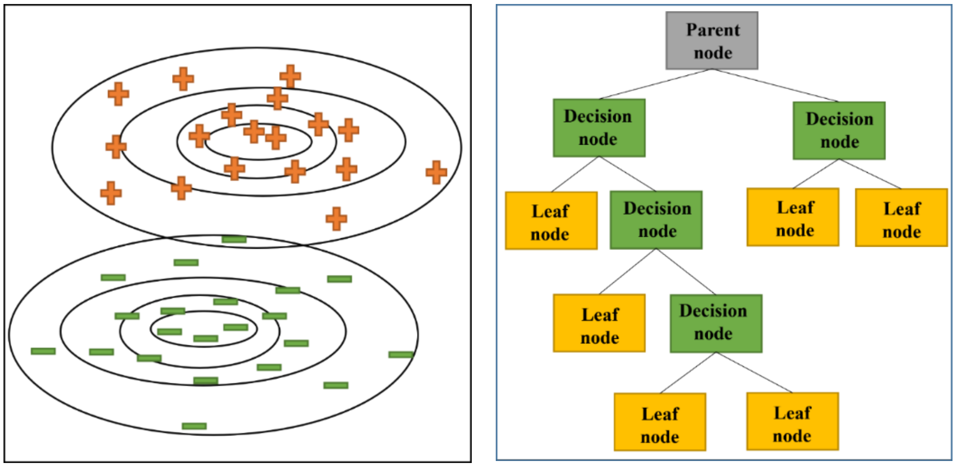
In [194], a sparse self-encoding neural network model employing an unsupervised learning algorithm is introduced to enhance fault detection. Data preprocessing involves wavelet transformation using db3 as the mother wavelet and extracting fault features to train the DNN model. This configuration yields a deep belief network (DBN) structure, exhibiting improved classification accuracy and predictive capabilities. The ANN is renowned for its adeptness in pattern recognition, making it valuable for detecting and classifying faulty lines in power systems [30], [129]. The DNN, an extension of the NN with multiple hidden layers, excels at modeling complex non-linear relationships. Each layer in the DNN extracts a distinct set of features, contributing to its enhanced performance [30], [195]. In [196], a hybrid quantum computing-based deep learning approach is proposed for substation and transmission line fault diagnosis. This scheme demonstrates high computational efficiency, good generalization ability, and fast response time. Additionally, [195] employs a stacked sparse autoencoder (SSAE) based DNN for predicting transient stability status in large power system networks post-fault occurrence. The SSAE-based classifier outperforms simple multi-layer perceptron (MLP) based classifiers in terms of accuracy.

2.9.1.5 Naïve Bayesian (NB)

The Naive Bayes (NB) classifier operates under the assumption that each feature of a class is distinct, classifying new samples based on the highest probability of belonging to a particular class, as determined by the Bayes theorem [41], [95], [183]. It is particularly suitable for small datasets with numerous parameters and is easy to understand [95]. Figure 2.13(a) provides a graphical representation of the NB model.

According to [41], the NB classifier excels in noisy testing environments, exhibiting minimal classification errors within the shortest classification time. The micro-grid fault detection model presented compares four classifiers: RBFNN, KNN, DT, and NB. The DT classifier demonstrates superior classification accuracy in both noisy and noise-free environments compared to KNN, RBFNN, and NB classifiers. Although the NB classifier boasts the fastest classification speed, its testing error is slightly higher than DT and lower than RBFNN and KNN. In [45], a geometric approach-based fault classification is introduced, monitoring the elliptical behavior of power transmission line voltage and current signals. Changes in the elliptical shape, such as radii value and slope, serve as data for the classifier. Various classifiers, including NB, LR, RBFNN, MLP, DT, Decision Table Naive

Bayesian (DTNB), KNN, Adaboost, and bagged DT, are employed, with DTNB and KNN yielding satisfactory results and NB achieving the highest classification accuracy of 98%.



(a) (b)

Figure 2.13 Illustration of (a) Naïve Bayesian and (b) Decision Tree.

[197] asserts that the NB classifier requires minimal time to train despite the size of the dataset, applying it for double-circuit line fault classification. Using DWT for data preprocessing of three-phase current, the NB classifier achieves a 99% classification accuracy for the first zone of protection of the transmission line. For a high voltage 750 kV-600km long transmission line, [61] presents fault detection, classification, and localization models employing NB, RBFNN, bagging, and boosted DT classifiers. Boosting exhibits the lowest accuracy, while RBFNN and bagging show the best fault classification accuracy, and NB and bagging demonstrate the highest localization prediction accuracy. Lastly, [198] proposes a Human-Level Concept Learning (HLCL) based incipient fault detection and classification scheme that involves waveform decomposition and hierarchical probabilistic learning to identify faults based on waveform features.

2.9.1.6 Decision Tree (DT)

The DT classifies samples hierarchically, based on decisions made at tree branches from the root node down to non-leaf and leaf nodes [95]. These branches result from comparing predictor values with trained weights [96]. Despite their relatively low predictive accuracy, DTs are favored for their speed of fitting and minimal memory usage [95]. Figure 2.13(b) provides a graphical representation of the DT model.

According to [41], DTs exhibit the best classification accuracy for both training and testing datasets in noise-free environments, while the NB classifier performs best for testing datasets in noisy environments. In [199], a Classification and Regression Tree (CART) model is introduced for fault detection and classification in a 400 kV transmission line. Utilizing DWT for fault current preprocessing, CART automatically selects prominent features and discards others, handling both numerical and categorical features while remaining robust to outliers.

In [47], a semi-supervised scheme is presented capable of handling both labeled and unlabeled data, achieving high classification accuracy. Co-training the DT and KNN classifiers improves accuracy compared to self-training methods. This co-training enhances pattern recognition, particularly with a small percentage of labeled data (2%), and preprocessing using DWT, offering minimal computational complexity [47]. For single-phase distribution systems, [200] introduces fault detection, open/short circuit fault classification, and locating faulty node work. The work tested three classifiers: KNN, SVM, and DT. DT demonstrates higher classification accuracy, lower computational burden, and faster execution in real-time applications, establishing itself as a robust algorithm. Lastly, [48] proposes a novel feature extraction technique, the mathematical morphology-based DT classifier, enabling fast real-time fault detection and classification. This method outperforms other feature extraction techniques, reducing computational burden while improving classification accuracy.

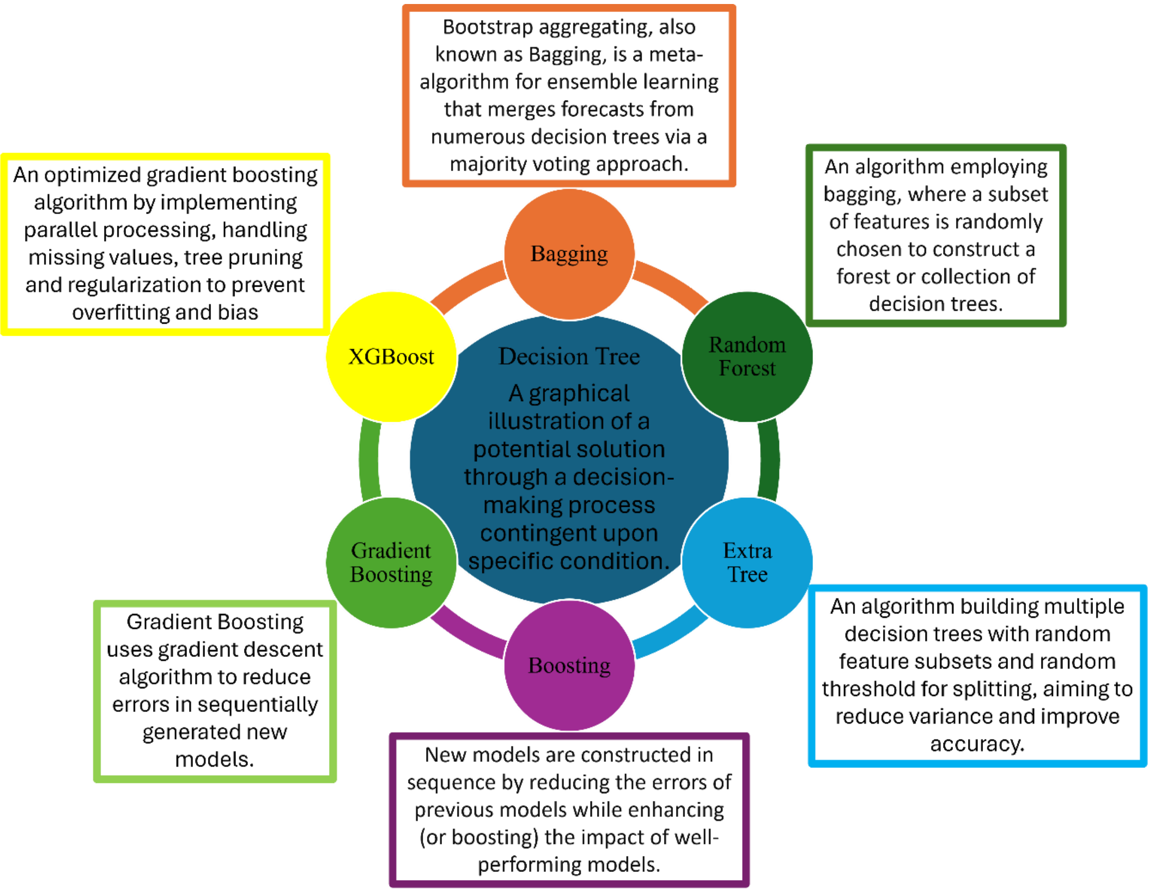

Bootstrap aggregating, also known as Bagging, is a meta-algorithm for ensemble learning that merges forecasts from numerous decision trees via a majority voting approach.
An optimized gradient boosting algorithm by implementing parallel processing, handling missing values, tree pruning and regularization to prevent overfitting and bias
An algorithm employing bagging, where a subset of features is randomly chosen to construct a forest or collection of decision trees.
Bagging
XGBoost
Random Forest
Decision Tree
A graphical illustration of a potential solution through a decision-making process contingent upon specific condition.
Gradient Boosting
Extra Tree
Boosting
Gradient Boosting uses gradient descent algorithm to reduce errors in sequentially generated new models.
An algorithm building multiple decision trees with random feature subsets and random threshold for splitting, aiming to reduce variance and improve accuracy.
New models are constructed in sequence by reducing the errors of previous models while enhancing (or boosting) the impact of well-performing models.

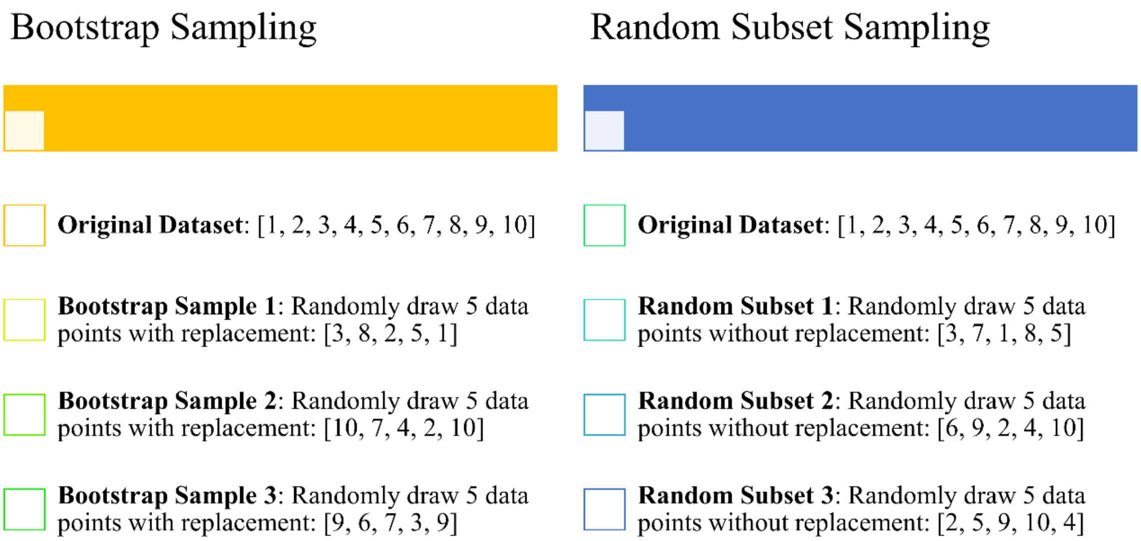
Figure 2.14 Illustration of the key idea behind Ensemble Methods.

2.9.1.7 Ensemble Methods

Ensemble methods, as defined by [95], [201], offer a promising strategy for enhancing model performance by combining multiple classifiers into a unified framework. Instead of relying solely on a single model, ensemble methods harness the diversity of multiple models to achieve more accurate predictions. Their widespread adoption stems from their ability to enhance predictive accuracy, mitigate overfitting, and enhance overall robustness.

Classifiers can be grouped into homogeneous (e.g., bagging, boosting, random forest, rotation forest, and random space) and heterogeneous (e.g., voting and stacking) ensemble methods based on the type of hybridization, as outlined in [141], [142], and [201]. These methods excel, particularly when individual models exhibit distinct strengths and weaknesses, enabling them to complement each other effectively and yield superior outcomes. The hybridization of models can take various forms, including parallel, series, or parallel-series combinations, as elaborated in [201] with a detailed analysis of their applicability in time series forecasting. Among these methods, the bagged and boosted decision tree (BBDT) stands out as it combines both weak and strong learners simultaneously. This ensemble technique leverages the strength of a strong learner for boosting while iteratively incorporating weak learners and adjusting their weights to rectify misclassifications [95]. DT serves as the foundation of all ensemble methods, as depicted in Figure 2.14, summarizing the core concept behind ensemble methodologies.

Furthermore, to understand various ensemble methods comprehensively, it is essential to grasp basic concepts such as bootstrap sampling and random subset sampling. These techniques, elucidated in Figure 2.15, play a crucial role in generating diverse trees within ensemble methods. The primary distinction between the two sampling techniques lies in how they handle the repetition of values within subsets. In bootstrap sampling, values can be repeated within a subset, whereas in random subset sampling, each subset will never contain repeated values.



□ **Original Dataset**: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

□ **Bootstrap Sample 1**: Randomly draw 5 data points with replacement: [3, 8, 2, 5, 1]

□ **Bootstrap Sample 2**: Randomly draw 5 data points with replacement: [10, 7, 4, 2, 10]

□ **Bootstrap Sample 3**: Randomly draw 5 data points with replacement: [9, 6, 7, 3, 9]

□ **Original Dataset**: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

□ **Random Subset 1**: Randomly draw 5 data points without replacement: [3, 7, 1, 8, 5]

□ **Random Subset 2**: Randomly draw 5 data points without replacement: [6, 9, 2, 4, 10]

□ **Random Subset 3**: Randomly draw 5 data points without replacement: [2, 5, 9, 10, 4]

Figure 2.15 Illustration of samples being drawn in bootstrap and random subset sampling.

Bagging: Bagging, which stands for Bootstrap Aggregating, is an ensemble technique that outperforms single classifiers in terms of stability and accuracy by reducing variance and preventing overfitting [202]. In bagging, multiple training sets are randomly created from the dataset by shuffling, and a new decision tree is constructed for each new training dataset. [202]. Each model in bagging carries equal weight, and the final prediction is determined by aggregating the majority vote [203], [204]. The prediction process in bagging involves creating numerous bootstrap samples from the available data, applying a prediction method to each sample, and then combining the results through averaging for regression or simple voting for classification. This combined approach decreases variance, leading to improved performance without sacrificing computation time, making bagging much faster than training a neural network. The subsets of the original training dataset produced by bagging, known as bags, are smaller than the initial dataset and are obtained using bootstrapping-sampling algorithms. These bags provide a comprehensive understanding of the dataset [205]. Figure 2.16 shows a generalized workflow of the bagging ensemble technique.

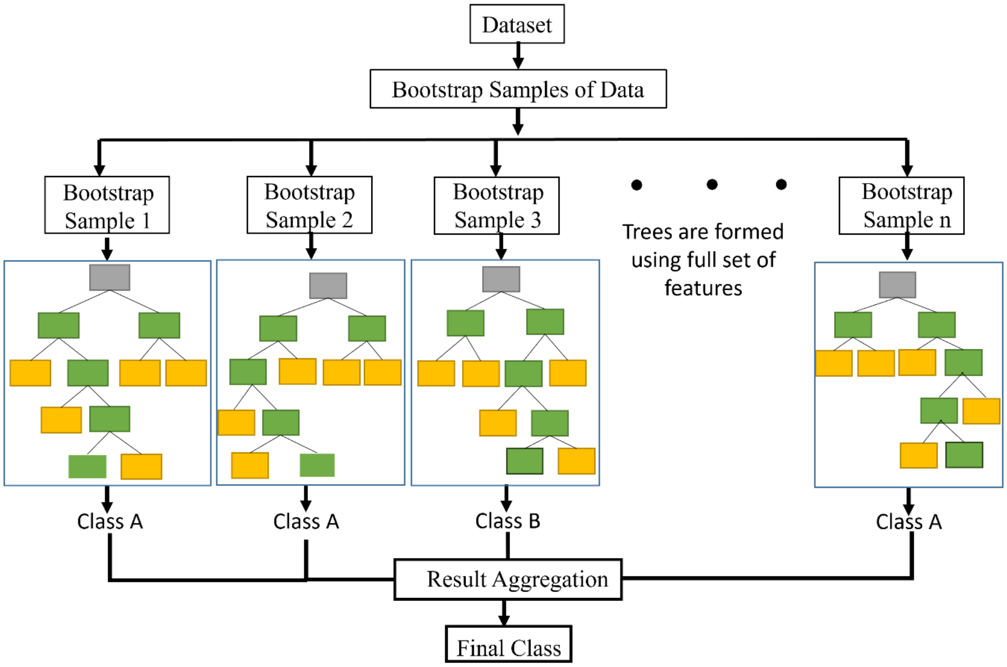


Figure 2.16 Illustration of Bagging model.

Bagging is known to be more resistant to model overfitting compared to boosting. Traditional bagging typically does not involve feature randomization; each base learner, often a DT, is trained on the full set of features at each split [206]. Bagging aims to make models less variable and is often compared to RF in regression problems but may be less effective in classification problems [207]. A bagged tree-based ensemble classifier was implemented for a series compensated transmission line fault classification. This approach utilized the singular value decomposition (SVD) principle and fast discrete orthonormal S-transform (FDOST) for feature extraction from half cycles of post-fault current signals [43]. While bagging has been widely used for power system fault classification in the literature, its application as a regressor for power system fault localization is limited.

Random Forest: Random Forest (RF) is another homogeneous ensemble ML technique that resembles bagging in its formation of multiple decision trees. However, each tree in a RF is decorrelated by selecting a random subset of predictors as split candidates from a full set of available predictors [204]. This process introduces randomness and diversity among the trees, making RF more robust and less prone to overfitting than individual decision trees. In RF, subsets of the data are sampled with replacement, like bagging. Additionally, at each split point of a decision tree, only a random subset of features is considered. For instance, at a split node, RF may use only a random subset of, let's say, 5 out of the 10 available features. This random selection of predictors is repeated at each split of every decision tree in the ensemble [204], [208]. While each tree in a RF may be noisy and unstable, the combination of multiple trees results in low bias, good generalization capability, and reduced variance. This diversity among trees helps prevent them from becoming highly correlated and mitigates the risk of overfitting [208]. Figure 2.17 shows a generalized workflow of the RF ensemble technique.

Determining the optimal number of trees for a RF can be achieved by monitoring the out-of-bag error. This error metric helps identify the number of trees that minimize prediction errors without overfitting the model [203]. For instance, in a study by [208], the preprocessing of harmonics and high impedance fault (HIF) signals involves using an extended Kalman filter to obtain 12 features for both HIF and non-HIF signals. These features, which include amplitude and phase information of HIF current signals, are then utilized to train and test the RF algorithm for fault detection. A fault detection and classification work for a medium voltage distribution network comparing three classifiers: RF, LR, and SVM has been presented in [209]. Among these, RF exhibited the best performance across various performance metrics, including AUC, accuracy, sensitivity, and specificity, followed by SVM and logistic regression, respectively.

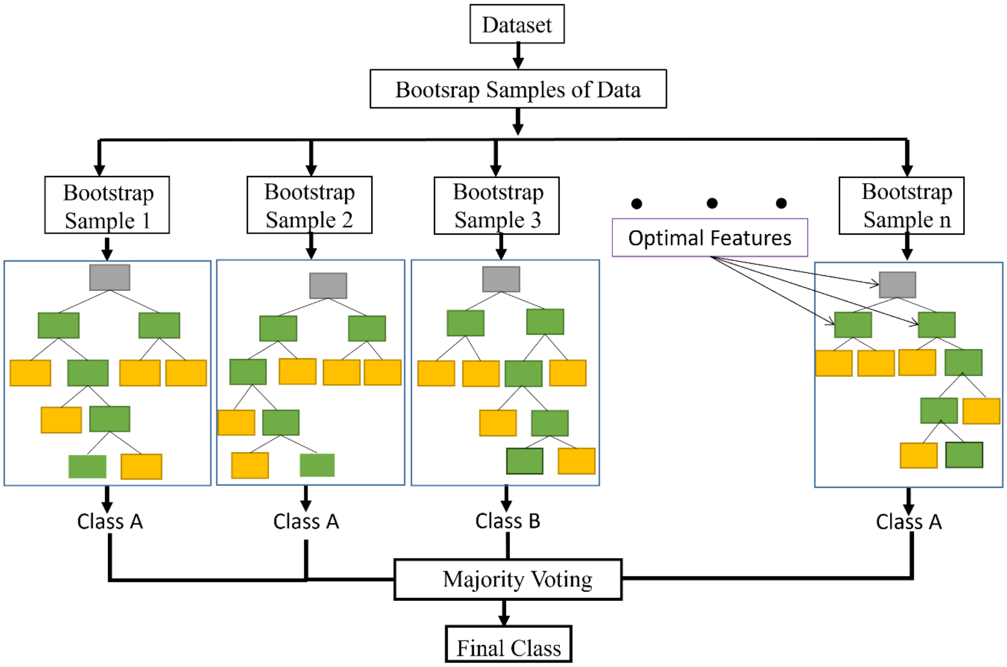


Figure 2.17 Illustration of Random Forest model.

Extra Tree: The Extra-Trees (ET) model follows a traditional top-down structure, comprising an ensemble of unpruned decision or regression trees. What sets it apart from other tree-based ensemble methods is its approach to node splitting: it selects split points entirely at random. Unlike some other ensemble methods, ET builds its trees using the complete learning sample rather than bootstrap replicas [210]. An extremely random tree-growing process is employed, involving attribute randomization from a Random Subspace, where the strength of randomization is controlled by parameter K. The amount of smoothing is determined by the parameter n_{min}, while M sets the number of trees to be formed, consistent with other ensemble methods. Predictions from all trees are aggregated, with the majority voting for classification and arithmetic averaging for regression. Figure 2.18 presents the workflow of the ET algorithm.

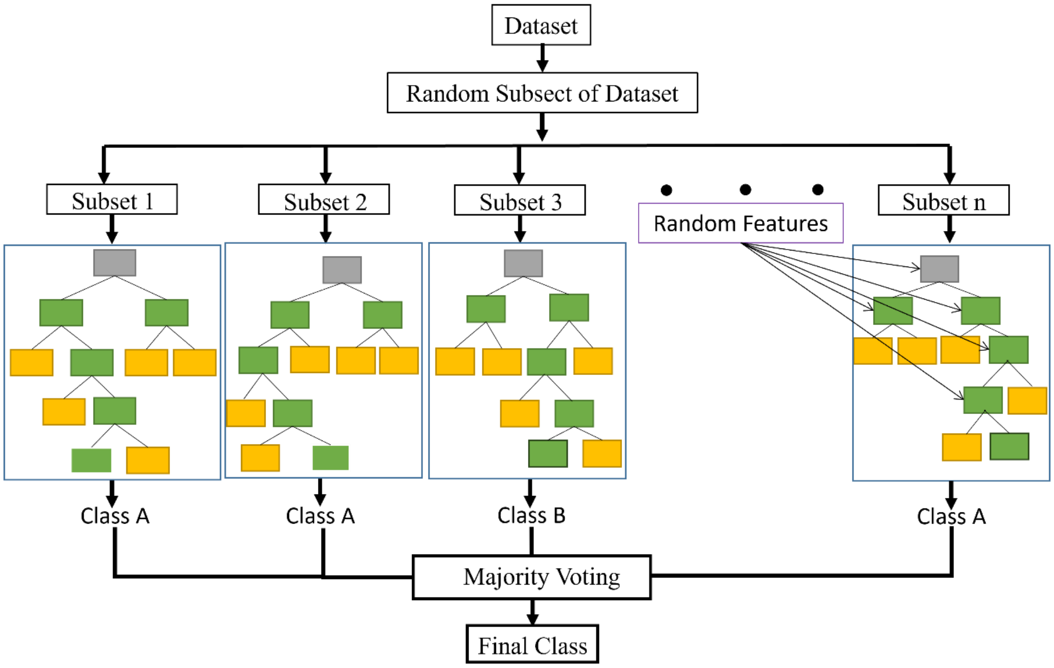


Figure 2.18 Illustration of Extra Tree model.

Increasing the value of K (attribute selection strength) in the ET method shrinks the dimensionality of the input space, particularly along input directions where the output variable varies more strongly. This can mitigate the over-smoothing effect of the dimensionality curse, especially in regression scenarios with irrelevant variables. However, the effect is less pronounced in classification problems. Bias/variance analysis has shown that ET operates by increasing bias while simultaneously reducing variance. When properly tuned, bias increases marginally compared to standard trees, while variance nearly disappears. However, over-randomization can lead to an increase in bias, hindering the algorithm's ability to detect behavior of low importance [210].

In the context of classification problems, increased randomization can be beneficial. Future improvements to randomization techniques should focus on reducing bias, as it is the primary component of ET error. One approach to reducing bias could involve extending trees to include multiple attributes, a concept that has shown some effectiveness in the context of random forests. Additionally, pairing ET with boosting, a technique known to reduce bias, could further mitigate bias. While ET is equally accurate as other ensemble approaches for classification, the Random Subspace approach may occasionally underperform compared to other methods, particularly for regression tasks [210].

Boosting: Boosting involves the sequential accumulation of a series of classifiers within the dataset, essentially transforming weak learners into strong learners [95], [61]. Boosting, often referred to as a "meta-algorithm," operates sequentially, with each subsequent model aiming to rectify the shortcomings of its predecessor. Each model in the sequence relies on the output of the previous one. The primary objective of boosting is to reduce bias within the model. By combining multiple weak learners, boosting creates a strong learner capable of achieving higher accuracy across the entire dataset. While individual models may excel on specific portions of the data, they may not generalize well to the entire dataset. Consequently, the ensemble's performance is significantly improved by the contribution of each individual model. Commonly used boosting algorithms include AdaBoost, Gradient Boosting Machine (GBM), Extreme Gradient Boosting Machine (XGBM), Light GBM, and CatBoost [207].

In a study focusing on fault detection, classification, and localization in a 600 km 750 kV transmission line, four classifiers - NB, RBFNN, bagging, and boosting were employed. The NB, RBFNN, and bagging classifiers demonstrated comparable classification and location prediction accuracy, while the boosting classifier performed poorly in both cases [61]. In another scenario involving an IEEE 13 bus SLG fault test system, a combination of ANN and SVM with bagging and adaptive boosting (AdaBoost) was used [67]. Boosting, demonstrated by AdaBoost, aims to reduce bias by creating a series of classifiers, with each subsequent model attempting to correct the mistakes of its predecessor. While boosting algorithms like AdaBoost are effective at improving overall accuracy, they are susceptible to overfitting [206]. AdaBoost has been applied to earth fault detection using continuous wavelet transformed grayscale image fault data, with a comparison against CNN and SVM [30]. Additionally, AdaBoost was utilized for fault classification on a modified WSCC 3 machine 9 bus system; however, its performance was notably inferior to the Bagging ensemble method [43]. AdaBoost tends to select data that enhances its prediction accuracy. In the work presented in [67], for the faulty phase classification, SVM outperformed other classifiers, while for fault impedance level detection, ANN and bagging showed strong performance. Bagging also excelled in faulty line segment identification. However, overall accuracy with AdaBoost was lower compared to other methods due to its sensitivity to noisy data and outliers. In another application, a faulty feeder detection and classification system utilized AdaBoost in conjunction with CART and SVM algorithms [211]. This model employed DWT for data preprocessing, generating time-frequency matrices of transient fault current signals, and demonstrated high classification accuracy with both algorithms.

Gradient Boosting: Unlike RF, which employs the bagging principle, gradient boosting (GB) is based on the principle of boosting, i.e., combining models with high bias and low variance error to reduce the bias while keeping a low variance. Instead of using deep trees and different training datasets, boosting trees employ shallow trees that are trained on the same dataset but where each tree is specialized in a specific characteristic of the input-output relationship. In particular, successive shallow trees are trained in series, where the nth tree is trained with the goal of reducing the prediction errors of the previous (n-1)th trees [62]. Consequently, GB combines models characterized by high bias and low variance error to mitigate bias while maintaining low variance, whereas RF combines models with low bias and high variance error to reduce variance while maintaining low bias. In RF, the final decision is determined by majority voting, whereas in GB, the final decision is arrived at through a sequential combination of three or more iterated trees (typically fewer than 8). A GB tree based fault diagnosis methodology has been demonstrated for fault detection, classification, and faulty branch identification in [62].

XGBoost: The Gradient Boosting Machine (GBM), known for its simple parallelism and high prediction accuracy, has emerged as a powerful tool in the realm of artificial intelligence. Among its implementations, XGBoost stands out as one of the most effective algorithms for supervised learning. XGBoost is a scalable and efficient implementation of GBM, suitable for both classification and regression tasks. Some benefits of the XGBoost model are: (1) Parallel Computing: XGBoost automatically employs multithreading parallel computing, making it faster than standard ensemble learning methods. This feature is particularly beneficial for handling large datasets commonly encountered in real power grid applications. (2) Regularization: XGBoost incorporates a regularization term that enhances its generalization capability, mitigating the tendency of decision trees to overfit the data. (3) Data Handling: XGBoost is a tree structure model that doesn't require the data gathered by PMU in the power system to be normalized, thus simplifying the preprocessing steps. It can also effectively handle missing values, making it appropriate for PMU-based data applications [212]. Data scientists often prefer XGBoost due to its efficient out-of-core computing performance [213]. For classification, XGBoost optimizes the following objective function:

$$C_{obj} = \sum_{i=1}^{n} \log\left(1 + \exp\left(-2y_i\,\hat{p_i}\right)\right) + reg \times \Omega(f) \qquad \ldots(2.14)$$

Whereas for regression, XGBoost optimizes the following objective function:

$$R_{obj} = \sum_{i=1}^{n} \frac{1}{2}(y_i - \hat{y_i})^2 + reg \times \Omega(f) \qquad \ldots(2.15)$$

Where y_i is the actual label or value of the ith instance (either -1 or 1 for binary classification), $\hat{y}_i$ is the predicted value of the ith instance, and $\hat{p}_i$ is the predicted probability of the ith instance belonging to the positive class. Ω(f) is the regularization to prevent overfitting, and reg is the regularization parameter. T stands for the number of terminal nodes in the tree, γ is responsible for controlling the complexity of the tree, $\boldsymbol{\lambda}$ is the L2 regularization term, and $\|w\|^2$is the squared L2 norm [212].

$$\Omega(f) = \gamma T + \frac{1}{2}\boldsymbol{\lambda}\|w\|^2 \qquad \ldots(2.16)$$

Built on the gradient boosting architecture, XGBoost continuously improves learner effectiveness by iteratively adding new decision trees to minimize the residual error. Unlike traditional gradient boosting, XGBoost approximates the loss function using a Taylor expansion, striking a better balance between bias and variance and often requiring fewer decision trees to achieve high accuracy [213]. Notably, it focuses on reducing computational complexity during the best split determination step, which is the most time-consuming process in decision tree construction algorithms. XGBoost employs regularization techniques, fitting the residual error using first and second derivatives. It also supports column sampling to mitigate overfitting and reduce computation. To accelerate decision tree training without compromising accuracy, XGBoost employs several optimization techniques. For instance, it uses a column-based storage format to pre-sort data, enabling the parallel determination of optimal splits for each attribute. Additionally, XGBoost employs a sparsity-aware approach to efficiently omit missing values from the computation of the loss gain of split candidates. XGBoost has demonstrated remarkable performance in ensemble learning and has been successfully applied to transient stability prediction in power systems. However, its potential for power system fault diagnosis remains largely unexplored.

2.9.2 Regression Models

Regression techniques are employed to estimate or predict values for unseen data by analyzing patterns within existing data [95]. These algorithms are particularly suited for predicting continuous responses, and they require values at each underlying point to make accurate predictions for unseen data [137]. Among the regression techniques available, some commonly used ones include support vector regression (SVR), Gaussian process regression (GPR), and regression trees (RT). Other notable regression methods include linear regression, non-linear regression, and generalized linear models [95]. Linear regression models, known for their simplicity, ease of interpretation, and efficiency in training and fitting, are statistical techniques used to model continuous response variables as linear functions of predictor variables [95]. Linear regression has been used for fault localization in the IEEE 39 Bus system, and the localization performance has been compared with Bi-GRU, BPNN, and other regression models, namely RT, SVR, and RF. Non-linear regression models, on the other hand, are parametric and are utilized to capture non-linear relationships between input variables through non-linear equations.

2.9.2.1 Gaussian Process Regression

The Gaussian Process Regression (GPR) model, being non-parametric, is commonly employed for predicting continuous response variables [95]. It finds extensive application in addressing spatial data interpolation challenges characterized by uncertainty and the optimization of complex designs [57]. In a study presented in [57], a methodology integrates DWT for feature extraction, utilizing NB and SVM classifiers to identify and categorize faults in an 88 kV distribution system. Simultaneously, SVM regression (SVR) and GPR are employed for fault localization. Comparative analysis reveals higher classification accuracy for the SVM classifier compared to NB, while GPR demonstrates superior fault localization accuracy over SVR. Consequently, a hybrid approach combining SVM for classification and GPR for fault location estimation is proposed for distribution system fault diagnosis [57]. Similarly, a model for fault detection, classification, and localization in a 66 kV transmission system is presented in [59]. This approach also leverages DWT for feature extraction, with NN and SVM serving as classifiers. The SVM classifier demonstrates superior classification accuracy compared to NN. GPR is utilized as a fault location estimator, yielding highly accurate estimations. Another scheme presented in [58] employs DWT for feature extraction, NN as a fault detector and classifier, and GPR as a fault location estimator for the 22 kV distribution system. Faults are classified as high impedance fault, capacitor switching, or load switching.

2.9.2.2 Support Vector Regression

The SVR model operates similarly to the SVM classifier but with slight modifications tailored for predicting continuous response variables [95], [96]. Instead of defining a decision boundary (hyperplane) like the SVM classifier, SVR aims to minimize the deviation in the sample data from the parameter value as much as possible. In [178], a scheme utilizing 11 SVRs is introduced for fault location estimation. For a hybrid transmission line network incorporating overhead and underground cables, fault detection, classification, and localization are addressed in [56]. This approach employs the fast discrete orthogonal S-transform (FDOST) for fault current feature extraction, SVM as a classifier, and SVR as the fault location estimator. Notably, this scheme is robust against noise, variations in load angle, fault inception angle, and fault resistance.

In [214], fault detection, classification, and localization in a 132 kV transmission system are discussed. This approach utilizes the stationary wavelet transform (SWT) for fault feature extraction, the relevance vector machine (RVM) for fault detection and classification, and SVR for fault location estimation. The classification accuracy of the proposed SWT-RVM-SVR scheme is compared with DWT-SVM-GPR. Moreover, in [39], SVR's fault location estimation capability is compared with ELM and NN for distribution systems. The proposed ELM employs the Wavelet Transform (WT) for feature extraction, demonstrating immunity to fault resistance, loading variations, inception angle fluctuations, measurement noise, and thermal expansion or contraction of conductors. The ELM-WT scheme exhibits faster training speed and better generalization compared to SVM-WT and NN-WT. An SVR based 400kV-300km transmission line fault localization using only the amplitude of fault voltage has been presented in [215].

2.9.2.3 Regression Tree

The regression tree (RT) shares similarities with the DT used for the classification but is adapted to predict continuous response variables [95]. It finds application in predicting categorical, discrete, or nonlinear sample data [92]. In [35], RT serves as a base estimation algorithm for comparing fault localization results with the proposed Bi-GRU algorithm. Despite yielding lower estimation accuracy compared to the proposed model, RT exhibits significantly faster training time. Consequently, RT stands out for its simplicity and swift performance. RT has also been employed in [55] for fault localization in series compensated double-circuit 400 kV transmission line, and its performance has been compared with least squares SVM and ELM.

2.9.2.4 Bayesian Ridge Regression

The Bayesian ridge regression model has been reported to perform well for data with multicollinearity. Multicollinearity exists in the dataset when the predictor variables are correlated with each other. The inherent collinearity doesn't impact the general fit of the model. Nevertheless, multicollinearity can lead to several outcomes, including elevated standard errors for estimated parameters of collinear variables, increased sampling variability resulting in widely varying estimated parameters across different samples, and inaccurate interpretation of the effects of predictor variables due to the inevitable changes in one variable affecting another [216]. Bayesian ridge regression employs features of the Bayesian method and ridge regression. Initially, the Bayesian method evaluates the information about the dataset distribution, referred to as the prior distribution of parameters, using the likelihood function. Further, the posterior distribution is computed for unknown scenario prediction using the Bayes theorem given in equation 2.17:

$$P(\theta|x) \approx P(x|\theta)P(\theta) \qquad \text{…(2.17)}$$

Where P(θ|x) is the posterior distribution, P(x|θ) is the likelihood function, and P(θ) is the prior distribution. The information about posterior distribution is usually difficult to evaluate, except for simple cases. The Markov Chain Monte Carlo algorithm is used to get posterior distribution for complex problems. The ridge regression part of the model is responsible for dealing with the multicollinearity aspect of the dataset. Let there be another equation 2.18, where M_{jt} refers to the data multicollinearity. For such a dataset, traditional analytical techniques like the ordinary least squares method would lead to incorrect inferences of variables.

$$a_{jt} = M_{jt}\gamma_t \qquad \text{…(2.18)}$$

However, ridge regression will add a small bias, thereby enhancing the precision of regression parameter estimates. The ridge regression method can be seen as a variation of traditional posterior Bayes regression, where there's an exchangeable prior distribution applied to the components of the regression vector [216]. However, to render this assumption believable, an initial standardization process is required. This involves standardizing both the variables Z_{jt} and their corresponding coefficients γ_t using equations 2.19 and 2.20. Thus, now a_{jt} becomes equation 2.21.

$$Z_{jt} = \frac{(M_{jt} - \bar{M}_{jt})}{sd(M_{jt})} \qquad \text{…(2.19)}$$

$$\gamma_t = \frac{b_t}{sd(M_{jt})} \qquad \text{…(2.20)}$$

$$a_{jt} = M_{jt}b_t \qquad \text{…(2.21)}$$

Thus, Bayesian ridge regression is best suited for datasets having multicollinearity and gives better performance for such datasets than other ML models. Its capability to handle hierarchical data structures is praiseworthy. Traditional analysis techniques like ordinary least squares will result in inaccurate variable inference when dealing with multicollinear datasets. Ridge regression will result in a small amount of bias, but it will also produce estimates of the regression parameters that are more accurate. A variation of standard posterior Bayes regression with an exchangeable prior distribution on the components of the regression vector can be seen in the ridge regression approach. Nevertheless, a prior standardization is required to make this assumption realistic [216]. Bayesian ridge regression causes homogeneous shrinkage by converting all the features to have regression coefficients with common variance to produce a Gaussian distribution [217].

2.9.2.5 Ensemble Methods Regression

Ensemble methods, as discussed in the classification section, encompass both classification and regression algorithms. Bagging regression, RF regression (RFR), ET regressor (ETR), and XGBoost regressor (XGBR) exhibit positive attributes similar to those found in classification models. These methods leverage the collective wisdom of diverse models to enhance prediction accuracy and robustness. Bagging entails training multiple instances of the same regression algorithm on different subsets of the training data, typically achieved through bootstrapping. The final prediction often emerges as an average or weighted average of predictions from individual models. RFR, on the other hand, constructs numerous DTs during training and outputs the mean prediction of these trees for regression tasks. Renowned for its simplicity, scalability, and resistance to overfitting, RFR stands as a formidable ensemble method. Among boosting algorithms, XGBR is anticipated to deliver rapid and precise performance, although its exploration remains relatively limited. Ensemble methods in regression offer several advantages, including improved predictive accuracy, enhanced generalization, and greater resilience against overfitting. Consequently, they find widespread application across diverse domains where precise regression predictions are paramount, such as finance, healthcare, and marketing.

2.10 Advantages / Disadvantages of Fault Diagnosis Schemes

So far, several methods, such as conventional methods, WAMs monitoring, and ML-based methods, have been discussed for power system fault diagnosis. Hence, in this section, the advantages and disadvantages of these fault-diagnosis methods have been enlisted. This characteristical comparison helps in the proper selection of fault diagnosis methods as per the availability of resources. Table 2.1 presents the advantages and disadvantages of the conventional fault diagnosis, whereas Table 2.2 presents the advantages and disadvantages of ML-based fault diagnosis.

Table 2.1 Advantages and disadvantages of the conventional fault diagnosis methods.

Methodology	Advantages	Disadvantages
Impedance-based method	• No need for additional sensors or equipment to be installed. • Traditional and well-researched method. • Economical [73].	• Prior knowledge of system parameters is a must. • Prior knowledge of fault type is a must. • Highly dependent on system parameters and with aging location estimation accuracy will be affected. • Domain expertise is needed. • Mathematically extensive and time-consuming. • The calculations carried out are under certain assumptions which causes errors in the fault location estimation. • With the change in system topology remodeling of the system is needed thus, inefficient to adapt to the changes [22], [73],[29], [31].
Traveling waves method	• Higher accuracy • Quick and precise fault location estimation. • Insensitive to system parameters. • No prior knowledge of system parameters is needed. • Suitable for transmission lines [73].	• Installation of costly equipment needed. • Interpreting fault location from high transient signals needs expertise. • Unsuitable for multi-terminal lines, distribution lines, and lines with many tappings [73].

Table 2.2 Advantages and disadvantages of Machine Learning based fault diagnosis.

Techniques	Advantages	Disadvantages
Neural Network	• Gives satisfactory performance for imbalanced data [161]. • NN provides noise immunity, and fault tolerance and is robust [129]. • ELM generates weights and biases randomly without tuning and has only one hidden layer [29]. • ELM is simpler, has fast learning speed, least human interference, and is free from issues like local minima and overfitting [29]. • CNN is famous for fault classification via image recognition [109]. • DNN and CNN have inbuilt feature extraction tendencies. • DNN provides very high accuracy for fault classification [76].	• NN takes a longer time to train a model (Xu et al., 2005). • Selection of the number of hidden layers is a big concern for good generalization ability [76]. • Black box nature. • CNN and DNN require a very large amount of data to train hence, need large memory [76]. • The performance is highly dependent on the model's architecture designing. • Proper hyperparameter tuning is needed.
Decision Tree	• DT is easy to understand and interpret and requires no pre-processing [74], [76]. • Robust and automatic feature selection. • Works best for the categorical dataset hence giving good classification accuracy [95]. • Its union with other techniques is easy e.g. DTNB scheme [45], [76]. • CART gives good results for both classification and regression [199]. • CART itself identifies the most significant variables and eliminates non-significant variables [74].	• DT is prone to overfitting issues [74], [76]. • Becomes complex if too many features are responsible for invoking the target [74], [76]. • Bias towards the majority class.
K-Nearest Neighbor	• Simple and non-parametric [95]. • No model training it learns the whole data. • Works on similarity principle.	• Very large memory space is occupied thus it is computationally expensive [96]. • Need optimal value of k.
Logistic Regression	• Simple and preferred for binary class classification [161], [162].	• Biased towards majority class hence, not suitable for imbalanced data [162].

Ensemble Methods	• BBDT eliminates the drawback of weak learners with the aid of the strong learners thereby, improving the overall accuracy of the classifier [95]. • RF is simple and easy to use giving high accuracy [74]. • EM are relatively robust to noise and outliers as accuracy obtained is either through result aggregation or majority voting [95].	• Computationally complex[74]. • Difficult to interpret.
Naïve Bayesian	• Pretty easy to understand and suitable for a small dataset containing many parameters [95]. • It is based on the probability concept and hence, works well for unseen data classification consequently, the testing accuracy is high [41].	• Not preferred for a very large dataset [95]. • Bias towards the majority class.
Gaussian Process Regression Model	• Famous for regression problems [95]. • Works well for complex designs [95]. • The tuning of hyper-parameters is simple by maximizing marginal likelihood [95].	• Computational time is large and directly depends on the size of the dataset [39]. • Not suitable for high dimensional data.
Bayesian Ridge Regression	• Regularization helps prevent overfitting. • Handles multicollinearity. • Provides estimates of uncertainty.	• Interpretation is complex. • Choice of prior distribution is crucial.

2.11 Status of Research and Research Trend

Considering the vast collection of research dedicated to fault detection, classification, and localization in the literature, it becomes challenging to navigate through the multitude of models and identify those that are underexplored. Therefore, to provide a concise overview of the existing literature in this field, this section presents a tabulated summary of reviewed works. These tables aim to offer a quick reference encompassing the techniques employed, simulation tools utilized, and application systems. Table 2.3 enlists notable research works available for power system fault diagnosis utilizing unsupervised learning.

Table 2.3 Research works on fault diagnosis using unsupervised learning.

Author	Ref	ML Techniques Used	Application Area	Task Performed	Type of fault analyzed
Cordova et al. (2019)	[157]	Hierarchical clustering and Incremental learning, SVM and NN	IEEE-13 and IEEE 37 node test feeder simulated on RTDS	Fault detection and classification	Three phase faults
Li et al. (2019)	[156]	Hierarchical Clustering, PCA	200 - Bus system simulated on PSLF	Fault detection, classification and identify faulted area	Line trip, Generator trip, SLG, LL, LLL, LLLG
Santis et al. (2018)	[153]	K-mean Clustering and Evolutionary optimization	Power Grid managed by ACEA Company in Rome Italy data	Fault detection	Normal/Faulty
Majidi et al. (2015)	[158]	Fuzzy c means and KNN	134 Bus 13.8kV distribution network	Fault location identification	Three phase Fault
Jiang et al. (2014)	[82]	Gaussian Hidden Markov Model, ANN, and SVM	IEEE New England 39 Bus system simulated in PSCAD	Fault detection, classification and identify faulted part	Ground faults in lines and generator
He and Zhang (2011)	[134]	Gaussian Markov Random Field	IEEE 14 Bus System and IEEE 300 bus system simulated on MAT-POWER	Fault detection and identify/ locating faulty field	Line outage
Zhang et al. (2011)	[155]	Hierarchical Clustering, LDA, and Pattern recognition	IEEE 9 Bus and IEEE 39 - bus System simulated on MATLAB	Fault classification and identify faulty section	Three phase Fault

Similarly, Table 2.4 enlists notable research works available for supervised learning-based fault detection and fault cause identification. Moreover, Table 2.5 enlists notable research works available for supervised learning-based power system fault classification. Furthermore, Table 2.6 enlists notable research works available for supervised learning-based faulty line/faulty section identification of transmission lines. Lastly, Table 2.7 enlists notable research works available for supervised learning based exact fault localization.

Table 2.4 Supervised learning-based fault detection/cause identification.

Paper	Ref	ML Technique Used	Application Area	Task Performed	Type of fault analyzed
(Lin et al., 2020)	[16]	iSVDD, HISVDD	IEEE-123 node test feeder as test system on software GridLAB-D	Fault detection for varying DG penetration	SLG Fault
(Sarwar et al., 2020)	[50]	SVM, PCA, and FDA	IEEE-13 node distribution network simulated on MATLAB	Fault detection	HIF Fault
(Chen et al., 2018)	[164]	KNN and SVM	HVDC System simulated on PSCAD 500 kV	Fault detection	HIF
(Qi et al., 2018)	[209]	LR, SVM, and RF	Three Phase Distribution system data from Shanghai China	Fault detection	Undergroun d Three Phase Fault
(Li et al., 2016)	[166]	SVM	Digital fault data recorded from operating transmission line	Fault root-cause identification	SLG from lightning, wildfire, guano, insulator breakdown, and vehicle accident

(Samantaray, 2012)	[208]	RF	138/25kV substation transformer with 100km transmission line on 138kV side and 20km line on 25kV side simulated on PSCAD	Fault detection	HIF Fault
(Chan et al., 2011)	[41]	RBFNN, DT, KNN, and NB	Factory dataset assumed as data from the microgrid	Fault detection	Three Phase fault in Microgrid
(Samantaray and Dash, 2010)	[49]	EKF-SVM, FFT-DT and WT-RBFNN	25kV simulated on MATLAB	Fault detection	HIF Fault
(Cai et al., 2010)	[40]	ANN, LR, SVM, AIRS, KNN	Progress Energy Carolinas Data	Fault cause identification	Tree/Animal Fault
(Cai and Chow, 2009)	[163]	LDA and LR	Progress Energy Carolinas Data of Raleigh city	Fault cause identification	Tree/Animal Fault
(Sarlak and Shahrtash, 2008)	[151]	SVM	IEEE 4-node Test Feeder	Fault detection	HIF Fault
(Xu et al., 2005)	[162]	LR and ANN	Duke Energy Service area data	Fault cause identification	Tree/Animal Fault
(Xu and Chow, 2005)	[140]	LR and NN	Duke Energy Service Region data	Fault cause identification	Animal/Tree fault

Table 2.5 Supervised learning-based fault classification.

Paper	Ref	ML Technique Used	Application Area	Task Performed	Type of fault analyzed
(Pan et al., 2022)	[81]	CNN	IEEE 13 bus microgrid simulated on MATLAB	Fault classification	Three phase fault
(Shah et al., 2022)	[64]	NN and SVM	IEEE 9 Bus system with G3 as wind generator simulated on PSCAD	Fault detection and classification	Three phase fault
(Rai et al., 2021)	[17]	CNN	2 DGs at 25 km each connected to distribution system simulated on MATLAB	Fault detection and classification	Three phase fault
(Baghaee et al., 2020)	[182]	SVM	213.15W PV panel connected to grid at 400V simulated on MATLAB	Fault / islanding detection and classification	Three phase fault / islanding
(Fahim et al. 2020)	[192]	CNN	220kV-100km transmission line simulated on MATLAB	Fault detection and classification	Three phase fault
(Godse and Bhat, 2020)	[48]	DT	400kV-150km transmission line simulated on ATP-EMTP	Fault detection and classification	Three phase fault
(Lala and Karmakar, 2020)	[53]	ANN, SVM, and KNN	NIT Rourkela HV Engineering Lab test setup of 500kV, 500KVA power frequency transformer	Fault detection and classification	HIF
(Ren and Xu, 2020)	[218]	Transfer learning, ELM, and RVFL	New England 10-machine 39-Bus system simulated	Fault detection and classification	Three phase fault

(Wang et al., 2020)	[184]	K-means, Association rules, SGD	IEEE-9 Bus simulated on PSCAD and U.S. Western Power Grid WSCC as test system	Fault classification	Three phase fault
(Freire et al., 2019)	[51]	HMM, ANN, RF, SVM, KNN	Database developed at Laboratory of Signal Processing of Federal University of Para	Fault classification	Three phase fault
(Goswami and Roy, 2019)	[97]	DT, KNN, and SVM	90km line prototype simulated on MATLAB	Fault classification	Three phase fault
(Guo et al., 2019)	[52]	CNN	110kV/10kV 31.5MVA transformer with several feeders simulated on PSCAD	Fault classification	Three phase fault
(Patil et al., 2019)	[204]	RF	9216 fault cases simulated in PSCAD	Fault classification	Three phase fault
(Yu et al., 2019)	[84]	DNN	IEEE-34 Bus System simulated on DigSILENT Power Factory	Fault detection and classification	Three phase fault
(Abdelgayed et al., 2018)	[47]	Co-training of DT and KNN	Transformer-IM model simulated on PSCAD acting as the transmission line	Fault detection and classification	Three phase fault
(Jain et al., 2018)	[83]	SVM	IEEE-14 Bus simulated on PSS/E	Fault detection and classification	Three phase fault
(Guo et al., 2018)	[30]	CNN, Adaboost, and SVM	Distribution System	Fault detection and classification	Earth fault detection in feeder

(Mishra and Rout, 2018)	[46]	ELM, SVM, and NB	IEC Micro-grid model with base power 48 MVA and 4 DGs	Fault detection and classification	Three phase fault - LIF and HIF
(Mishra et al., 2018)	[43]	Bagging, Boosting and KNN	400kV transmission line simulated on PSCAD	Fault detection and classification	Three phase fault
(Tokel et al., 2018)	[3]	ANN	IEEE-13 Bus simulated on OpenDSS	Fault detection and classification	Three phase fault
(Magagula et al.,2017)	[172]	SVM	88kV distribution system simulated on Digisilent Power Factory	Fault detection and classification	Three phase fault
(Shukla et al., 2017)	[131]	SVM	Six Phase Transmission System 138 kV, 68 km Simulated on MATLAB about Springdale-McCalmont line of Allegheny Power System	Fault detection and classification	Six phase fault
(Sreewirote and Ngaopitakkul , 2017)	[168]	SVM	500 kV Transmission System Simulated System on ATP-EMTP program	Fault detection and classification	Three phase fault
(Xi et al., 2017)	[194]	DBN	IEEE-34 Bus simulated system	Fault detection and classification	Three phase fault
(Zeng et al., 2017)	[211]	AdaBoost + CART and SVM	10kV Distribution system simulated on PSCAD	Fault detection and classification	SLG fault
(Gowrishankar, 2016)	[130]	ANN	220kV-430km Transmission Line simulated on MATLAB	Fault detection and classification	Three phase fault

(Dasgupta et al., 2015)	[42]	KNN with Cross-correlation	400kV-300km Transmission line simulated on EMTP-ATP	Fault detection and classification	Three phase fault
(Jamil et al., 2015)	[129]	ANN	400 kV - 300 km Transmission line simulated model on MATLAB	Fault detection and classification	Three phase fault
(Yadav and Swetapadma, 2014)	[139]	KNN	Signals generated from 230 kV 100 km transmission line simulated on PSCAD and processed on MATLAB	Fault detection and classification	Three phase fault
(Gomes et al., 2013)	[45]	DT, NB, DTNB, and KNN	230kV-200km transmission line simulated on PSCAD	Fault detection and classification	Three phase fault
(Nikoofekr et al., 2013)	[185]	ART, Fuzzy ART Fuzzy ARTMAP	Data from Palash feeder in the southwest region of Tehran	Fault detection and classification	HIF fault
(Livani and Evrenosoglu, 2012)	[167]	SVM	230kV Hybrid Transmission line and 6-bus distribution system simulated on ATP	Fault detection and classification	Three phase fault
(Ray et al., 2012)	[189]	ELM	400kV-300km series compensated transmission line simulated on MATLAB	Fault classification	Three phase fault
(Shahid et al., 2012)	[77]	SVM, QSSVM, A-QSSVM, and TA-QSSVM	132kV-300km Transmission Line simulated on MATLAB	Fault detection and classification	Three phase fault

(Singh et al., 2011)	[177]	SVM	400kV-128km transmission line simulated on PSCAD	Fault detection and classification	Three phase fault
(Yusuff et al., 2011)	[176]	SVM	400kV-361.65km	Fault detection and classification	Three phase fault
(Youssef, 2009)	[174]	SVM	300km Transmission line simulated on ATP	Fault detection and classification	Three phase fault
(Bhalja and Maheshwari, 2008)	[171]	SVM	Two terminal transmission line simulated on PSCAD	Fault detection and classification	Three phase fault
(Jain et al., 2008)	[186]	ANN	220kV-100km double circuit line simulated on MATLAB	Fault detection and classification	Six phase fault
(Malathi and Marimuthu, 2008)	[173]	SVM	240kV-200km Transmission line simulated on MATLAB	Fault detection and classification	Three phase fault
(Silva et al., 2006)	[92]	NN	Oscillo graphic data of Brazilian utility company and 230kV-188km transmission line simulated on ATP	Fault detection and classification	Three phase fault
(Zhang and Kezunovic, 2005)	[159]	Fuzzy ART NN and Fuzzy KNN	WECC 9- Bus System simulated on ATP	Fault detection and classification	Three phase fault

Table 2.6 Supervised learning-based faulty line/section identification.

Paper	Ref	ML Technique Used	Application Area	Task Performed	Type of fault analyzed
(Chang et al., 2022)	[219]	SVM	HV 230kV transmission line with MV 22.8kV wind farms	Fault location identification	Three phase fault
(Galvez and Abur, 2022)	[11]	k-means clustering and directed tree model	8 Bus system, IEEE 118 Bus system and IEEE 123 test node feeder	Fault location identification	Three phase fault
(Mrabet et al., 2022)	[44]	RF, SVM, NN, DNN, NB, DT	IEEE 9 Bus system	Fault location detection	Three phase fault
(Teimourzadeh et al., 2021)	[220]	CNN and DQN	Two transmission line of 230kV – 45miles one double circuit and other single circuit	Fault detection and faulty segment identification	SLG fault
(Chen et al., 2020)	[34]	GCN, PCA+SVM, FCNN, and PCA+RF	IEEE-123 bus simulated on OpenDSS software	Faulty bus identification	SLG, DLG, and LL fault
(Liang et al., 2020)	[36]	ACNN	IEEE-33 Bus system simulated on MATLAB	Fault detection, classification, and faulty line identification	Three phase fault
(Sapountzoglou et al., 2020)	[62]	Gradient Boosting	LV system simulated model on MATLAB provided by Efacec company	Fault detection, classification, and faulty phase identification	SLG and LLL faulty

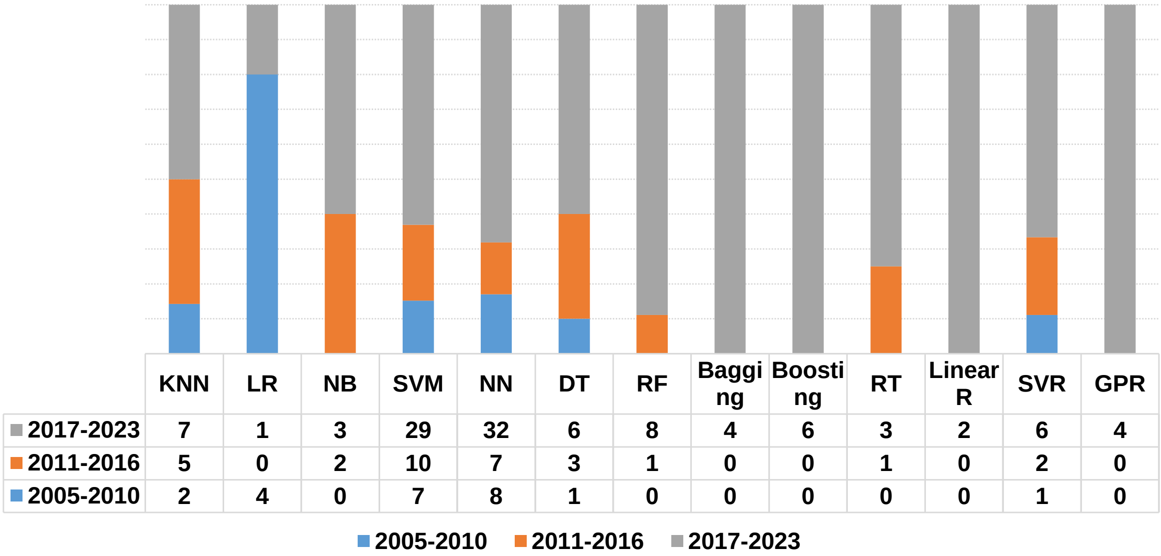
(Gashteroodkhani et al., 2019)	[37]	SVM	230kV 60Hz Hybrid Transmission line of 160km overhead and 32km underground simulated on EMTP software	Fault identification in OHL and UL and faulty section identification	Three phase fault
(Li et al., 2019)	[193]	CNN	IEEE-39 and IEEE-68 bus system simulated on MTALAB	Faulty line identification	Three phase fault
(Maruf et al., 2018)	[67]	ANN, SVM, Bagging, and AdaBoost	IEEE 13 Bus system with two DGs of 1800kVA and 2600kVA simulated on MATLAB	Fault classification, and faulty segment identification	SLG fault
(Bhattacharya and Sinha, 2017)	[179]	SVM and RNN (LSTM)	Simulation Data from Siemens PSS/E software	Fault classification, and faulty line identification	Three phase fault
(Hasan et al., 2017)	[61]	Bagging, Boosting, RBFNN, and NB	750kV-600km transmission line on MATLAB	Fault detection, classification, and faulty section identification	Three phase fault
(Deng et al., 2015)	[170]	SVM	Simulated model on PSCAD	Faulty section identification	SLG fault
(Livani and Evrenosoglu, 2014)	[180]	SVM	230kV 60Hz hybrid transmission line simulated on ATP	Faulty section identification	Three phase fault
(Livani and Evrenosoglu, 2013)	[181]	SVM	230kV, 60Hz three-terminal transmission system simulated on ATP software	Fault detection, classification, and faulty section identification	Three phase fault

Table 2.7 Supervised learning-based exact fault localization.

Paper	Ref	ML Technique Used	Application Area	Task Performed	Type of fault analyzed
(Sahani and Dash, 2020)	[38]	SVM, RT and ELM	Series capacitor compensated double circuit transmission line 400kV, 50Hz simulated on MATLAB	Fault classification, faulty phase identification and localization	Three phase fault
(Zhang et al., 2020)	[35]	Bi-GRU, RT, Linear Regression, RF, SVR and BPNN	New England 10-machine system simulated on PSASP, tested on IEEE-39 bus system	Fault localization	Three phase fault
(Srivastava and Parida, 2019)	[66]	KNN, SVM, Bagging, Linear Regression, RT, SVR, GPR,	200km 5 Bus test system of 11kV MV distribution line with two DGs simulated on Simulink	Fault classification and localization	Three phase fault
(Moloi and Akumu, 2019)	[214]	RVM, SVM, and SVR	132kV two bus system simulated on DigSILENT Power Factory	Fault detection, classification, and localization	Three phase fault
(Moloi and Yusuff, 2019)	[59]	SVM, ANN, and GPR	66kV distribution system simulated on DigSILENT Power Factory	Fault detection, classification, and localization	Three phase fault
(Fei et al., 2018)	[215]	SVR	400kV-300km Transmission line simulated on MATLAB	Fault localization	Three phase fault
(Moloi et al., 2018)	[58]	ANN, GPR	22kV system simulated on Power World Software	Fault detection, classification, and localization	HIF fault
(Patel, 2018)	[56]	SVM	400kV Hybrid transmission line simulated on ATP	Fault detection, classification, and localization	Three phase fault

(Chen et al., 2018b)	[29]	SW- ELM and SG-ELM	Three Separate Transmission Line Model Simulated on MATLAB	Fault detection, classification, and localization	Three phase fault
(Magagula et al., 2017a)	[57]	NB, SVM, SVR, and GPR	88kV Distribution system simulated on DigSILENT Power Factory	Fault detection, classification, and localization	Three phase fault
(Shafiullah et al., 2017)	[39]	ELM, SVR, and ANN	25kV-30km distribution System	Fault detection, classification, and localization	Three phase fault
(Fathabadi, 2016)	[178]	SVM and SVR	230kV-50km Transmission Line simulated on Proteus6/MATLAB	Fault detection, classification, and localization	Three phase fault
(Roy and Bhattacharya, 2015)	[187]	PNN	400kV-300km Transmission line simulated on MATLAB	Fault detection, classification, faulty phase identification, and localization	Three phase fault
(Upendar et al., 2012)	[199]	CART and NN (backpropagation)	400kV transmission line simulated on MATLAB	Fault detection, classification, and localization	Three phase fault
(Malathi et al., 2011)	[191]	ELM and SVR	400kV, 300km, series compensated transmission line simulated on MATLAB	Fault detection, classification, and localization	Three phase fault
(Malathi et al., 2010)	[175]	SVM, SVR and ELM	240kV-225km Transmission line simulated on MATLAB	Fault classification, faulty phase detection and localization	Three phase fault

The provided tables outline the status of research in ML-based power system fault detection, classification, and localization. Additionally, an analysis of available literature has been conducted to identify trends in ML-based fault diagnosis methods, represented graphically in Figure 2.19. This figure illustrates a six-year stacked analysis of works across various models.



	KNN	LR	NB	SVM	NN	DT	RF	Bagging	Boosting	RT	Linear R	SVR	GPR
■ 2017-2023	7	1	3	29	32	6	8	4	6	3	2	6	4
■ 2011-2016	5	0	2	10	7	3	1	0	0	1	0	2	0
■ 2005-2010	2	4	0	7	8	1	0	0	0	0	0	1	0

Figure 2.19 Six-yearly analysis of reviewed works on the model's preferences for ML based fault diagnosis.

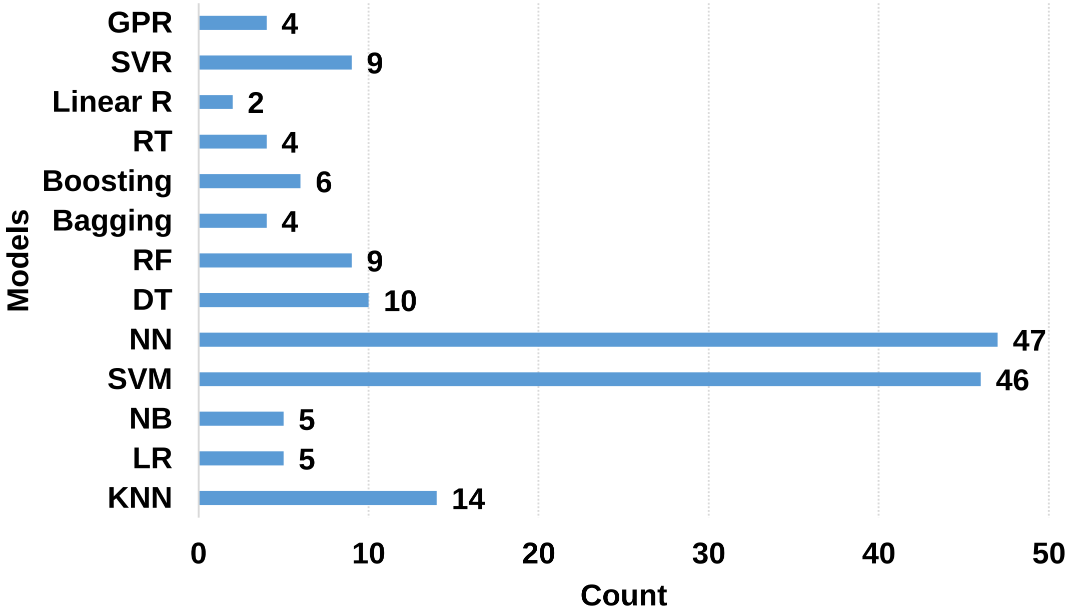
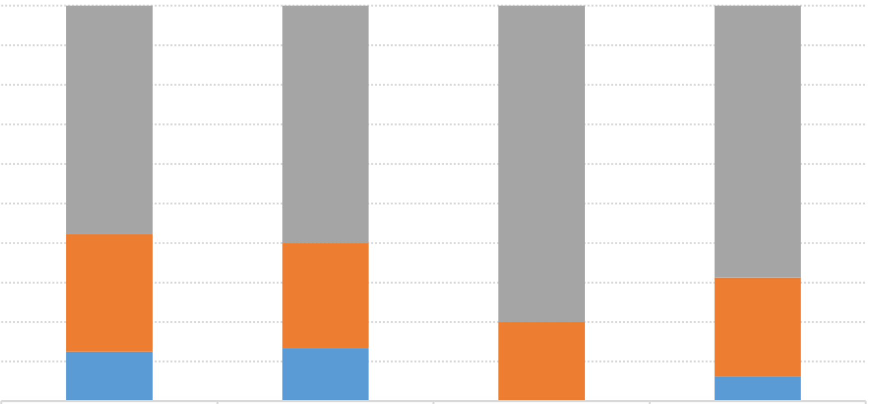


Figure 2.20 Count of model's preferences for ML based power system fault diagnosis literature reviewed.

Furthermore, Figure 2.20 provides an exact count of the literature reviewed in this dissertation, reflecting researchers' preferences for specific models. The analysis indicates that SVM and NN are the most researched models, while ensemble techniques are gaining popularity. However, ensemble regression techniques remain relatively unexplored. In Figure 2.21, a six-year stacked analysis of available works for fault detection, classification, and localization is presented, encompassing both location identification and exact localization approaches.



	Detection	Classification	Location Identification	Exact Localization
■ 2017-2023	37	27	12	11
■ 2011-2016	19	12	3	4
■ 2005-2010	8	6	0	1

Figure 2.21 Six-yearly analysis of reviewed works for ML-based fault detection, classification, location identification, and exact localization.

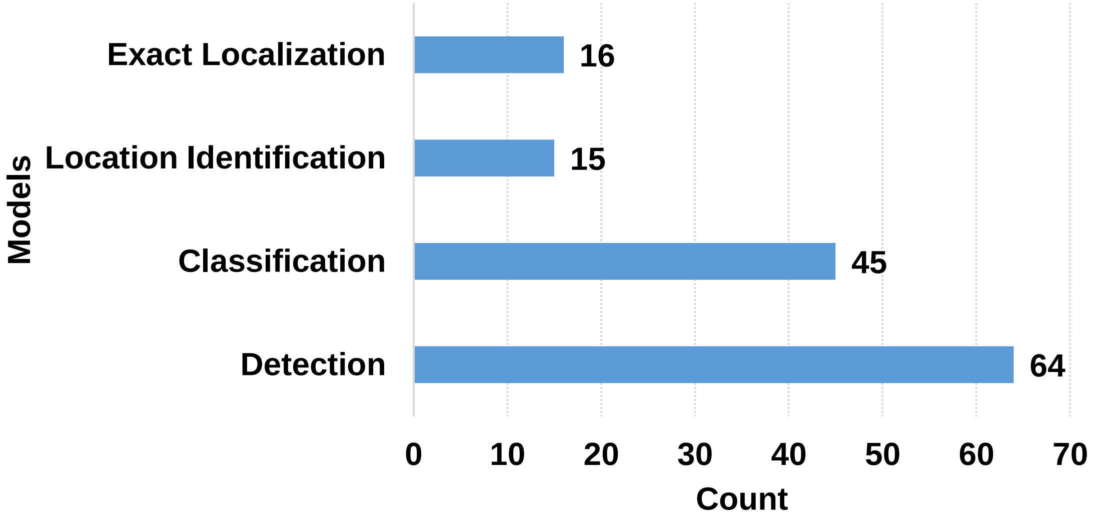


Figure 2.22 Count of works reviewed for fault detection classification and localization.

Finally, Figure 2.22 depicts the exact count of literature reviewed in this dissertation for fault detection, classification, and localization, considering both location identification and exact localization. The analysis reveals that relatively less work focuses on fault localization, with most concentrating on identifying faulty lines or sections. Consequently, research on fault localization using ensemble regression techniques is limited within the literature.

2.12 Research Perspective

Given that ML models are less susceptible to variations in line parameters and these techniques are considered robust and adaptable, it's imperative to investigate their adaptability in RES integrated power systems. Furthermore, a deduction drawn from the research status and trends indicates that ensemble techniques have been rarely explored for exact fault localization. Although bagging has been utilized for power system fault classification in the literature, its application as a regressor for fault localization remains absent. Consequently, this dissertation employs bagging for both fault classification and localization, functioning as a classifier and regressor, respectively. Similarly, the literature lacks performance analyses of ET and XGBoost for power system fault classification and localization. Thus, this dissertation aims to fill this gap by analyzing the performance of ET and XGBoost in both fault classification and localization tasks. Moreover, during the literature survey, it was noted that Bayesian Ridge Regression (BRR) has not been utilized for power system fault localization. Consequently, this study incorporates BRR for fault localization purposes.

2.13 Chapter Summary

This chapter offers a comprehensive review of power system fault detection, classification, and localization using ML techniques. It includes a brief discussion on power system monitoring and various fault diagnostic techniques. The evolution of fault localization from conventional impedance-based and traveling wave-based methods to ML-based approaches is outlined. Additionally, a structured framework for addressing problems using ML paradigms, along with various performance metrics and dimensionality reduction techniques, is provided.

The chapter delves into the extensive literature survey covering unsupervised and supervised learning techniques for fault detection, classification, and localization. Discussions are supported by taxonomical tabulations of research literature, facilitating easy reference retrieval. Moreover, the advantages and disadvantages of fault diagnosis techniques are thoroughly examined. Finally, the chapter highlights the current research status and trends, identifies gaps within the literature, and suggests several unexplored models that can be utilized for power system research areas.

Chapter 3 System Modeling and Fault Database Formation

3.1 Introduction

This chapter provides a detailed description of the chosen system for the proposed study and enumerates the parameters employed in simulating different components of the selected power system. It provides a comprehensive discussion and calculation of various parameters of the selected IEEE 9 Bus system. Additionally, the necessary data for simulating synchronous generators, transmission lines, transformers, and solar photovoltaic (PV) plants is detailed in subsequent sections of the chapter. Furthermore, the chapter extensively discusses the generation and formation of the fault database.

To assess the effectiveness of machine learning models for fault classification and localization within power systems and explore the impact of RES integrations on ML model performance and their adaptability to varying fault data availability, the study employs the IEEE 9 Bus system as a testbed. This standard IEEE system serves as a simplified representation of an electric grid. It facilitates the integration of RES, enabling the evaluation of ML models' fault classification and localization capabilities both with and without RES. The simulations are conducted using the MATLAB 2021b Simulink environment. Solar PV plants with capacities of 10 MW, 20 MW, and 30 MW are also modeled and integrated into the IEEE 9 Bus system for fault database generation. Consequently, a fault database is generated for both the "standard IEEE 9 Bus System" and the "RES integrated IEEE 9 Bus System," encompassing a wide range of fault attributes to incorporate actual field variations, including temperature and irradiance effects on power generation from different sizes of solar PV-based RES integrated into the studied transmission network (IEEE 9 Bus).

While numerous studies in power system fault diagnosis have been conducted on simple networks, such as two-bus transmission lines, the standard IEEE transmission and distribution systems are favored as they closely resemble real power systems. Notable examples of the transmission network are IEEE 9 Bus [44], 14 Bus [83], 39 Bus [152], and 68 Bus [193], and the distribution networks are IEEE 4 Bus [151], 13 Bus [202], 33 Bus [36], and 34 Bus [84]. Additionally, the IEEE 123 Bus is widely referenced for combined transmission and distribution networks [34]. Given its six transmission lines, the IEEE 9 Bus system is selected over larger transmission networks to enable a comprehensive impact analysis of RES integration on ML model performance for all transmission lines of the network.

3.2 Standard IEEE 9 Bus System

This study utilizes the standard IEEE 9 bus system as a reference for conducting the proposed study. Researchers have widely adopted the IEEE 9 bus system as a standardized testing framework for evaluating new methodologies across static and dynamic power system challenges. Test systems are preferred over actual power systems due to their convenience, as real system models often lack comprehensive documentation and are excessively large, making it challenging to identify overarching trends. Additionally, outcomes from real system models are typically less universally applicable than those obtained from test systems. Moreover, the IEEE 9 Bus system has served as a microgrid in a few research articles, given three generator sources, which can be renewable-based sources. Thus, it offers a highly simplified representation of an electric grid [221].

The IEEE 9 bus contains nine buses, three synchronous generating units, six transmission lines, three two-winding transformers, and three loads, as shown in Figure 3.1, illustrating the single line diagram of the standard IEEE 9 Bus system [155], [222].

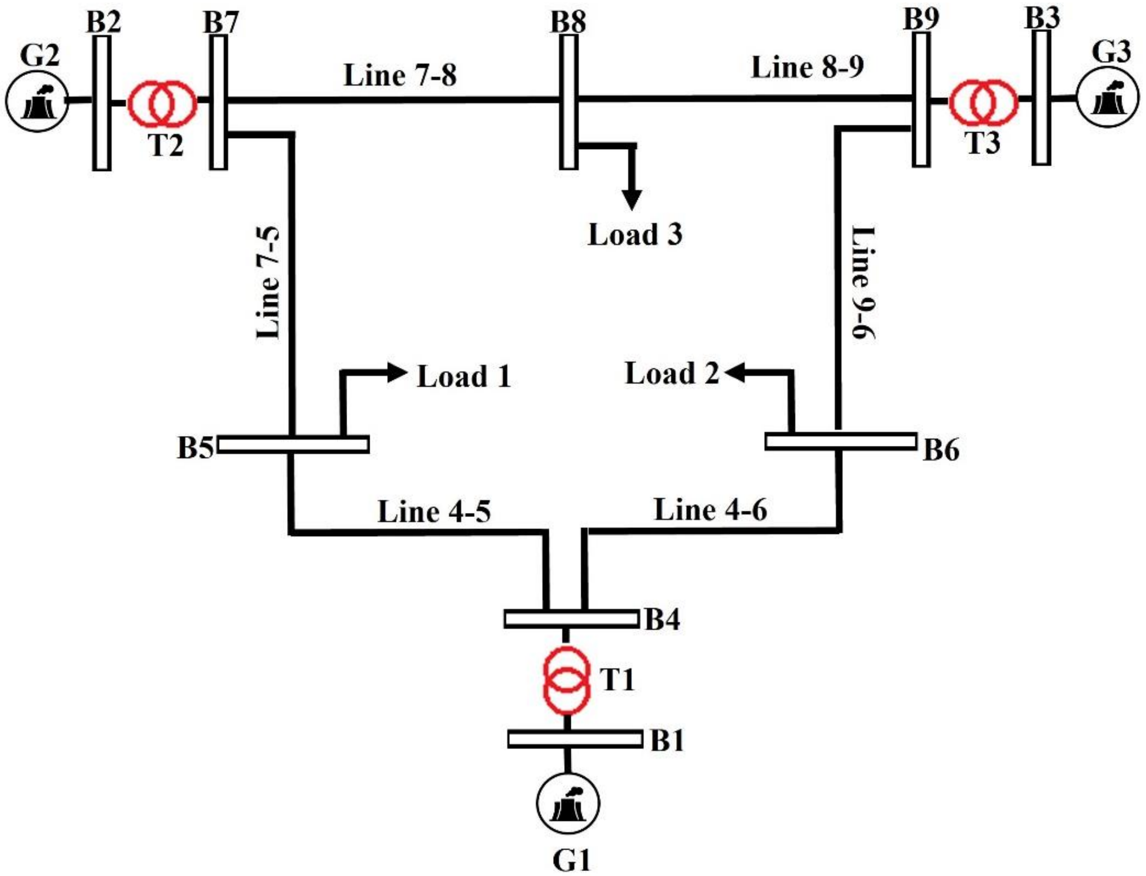


Figure 3.1 Single line diagram of the standard IEEE 9 Bus System.

Access to various network parameters is essential to accurately simulate the IEEE 9 Bus system in MATLAB. These parameters encompass the per-unit values of transmission line resistance, inductance, and susceptance. Moreover, information regarding the bus types of generator buses, along with the active and reactive power ratings of connected loads, voltage ratings of all buses, and power and voltage ratings of generators, is necessary for accurate simulation. These data can typically be obtained from literature sources and have been organized into Tables 3.1 and 3.2 for convenient reference [223].

Table 3.1 Transmission line parameter data.

Transmission Line	**R(pu)**	**X(pu)**	**B(pu)**
1-4	0	0.0576	0
2-7	0	0.0625	0
3-9	0	0.0586	0
4-5	0.010	0.085	0.176
4-6	0.017	0.092	0.158
7-5	0.032	0.161	0.306
7-8	0.0085	0.072	0.149
9-8	0.0119	0.1008	0.209
9-6	0.0390	0.17	0.358

Table 3.2 Bus type and voltage, generator, and load parameters data.

Bus No.	**Bus Type**	**Bus Voltage**	**Voltage (pu)**	**Generator**				**Load**	
				P (MW)	Q (MVAR)	Q_{min}	Q_{max}	P (MW)	Q (MVAR)
1	Slack	16.5 kV	1.04	0	0	0	0	0	0
2	Voltage	18 kV	1.025	163	6.7	inf	inf	0	0
3	Voltage	13.8 kV	1.025	85	-10.9	inf	inf	0	0
4	Load	230 kV	1	0	0	0	0	0	0
5	Load	230 kV	1	0	0	0	0	125	50
6	Load	230 kV	1	0	0	0	0	90	30
7	Load	230 kV	1	0	0	0	0	0	0
8	Load	230 kV	1	0	0	0	0	100	35
9	Load	230 kV	1	0	0	0	0	0	0

The bus and transmission line data presented in Tables 3.1 and 3.2 served as the foundation for simulating the IEEE 9 Bus system for the research conducted in this dissertation. Furthermore, the same system was utilized to integrate renewable energy sources (RES) onto Bus 7 and Bus 5, with capacities ranging from 10, 20, and 30 MW, resulting in variations in power fed to these buses. This research analyzed six distinct RES integration scenarios, leading to adjustments in the power fed values at Bus 7 and Bus 5 based on the specific case under study. However, it is important to highlight that all other values pertaining to bus data, branch impedance, and transmission line data, as listed in Tables 3.1 and 3.2, remained constant throughout the analysis. Therefore, the loads and transmission line parameters remained the same throughout this research.

3.3 IEEE 9 Bus System Modeling

To simulate the IEEE 9 Bus system on the MATLAB Simulink environment, firstly, the length of transmission lines needs to be calculated. To calculate the length of transmission lines, the following relation has been used:

$$Length\ of\ transmission\ line = \left(\frac{\sqrt{X*B}}{2\pi f}\right) * Velocity\ of\ light\ in\ km/sec \quad \ldots (3.1)$$

Where X is the inductance of the line, and B is the susceptance of the line. To calculate the line length, use base voltage V_b as 230 kV and base power S_b as 100 MVA.

Calculating the length of transmission line 4-5:

$$Length = \left(\frac{\sqrt{0.085\ * 0.176}}{2\pi 50}\right) * 3 * 10^5$$

Length of transmission line 4-5 = 116.798 km

Thus, the length of six interconnected transmission lines of the IEEE 9 Bus system for 50Hz frequency as calculated using (3.1) are: Line 4-5 is 116.798km, Line 4-6 is 115.131 km, Line 7-5 is 211.955 km, Line 7-8 is 98.907 km, Line 8-9 is 138.603 km, and Line 9-6 is 235.579

km. Once the length of all the transmission lines has been calculated, the per kilometer values of resistance, inductance, and capacitance need to be calculated using the following equations:

$$R_{actual} = R_{base} * R_{pu} \qquad \ldots(3.2)$$

$$X_{actual} = X_{base} * X_{pu} \qquad \ldots(3.3)$$

$$B_{actual} = B_{base} * B_{pu} \qquad \ldots(3.4)$$

$$R_{base} = X_{base} = \frac{{V_b}^2}{S_b} \qquad \ldots(3.5)$$

$$B_{base} = \frac{S_b}{{V_b}^2} \qquad \ldots(3.6)$$

$$R_{base} = X_{base} = \frac{(230 * 10^3)^2}{100 * 10^6} = 529$$

$$B_{base} = \frac{100 * 10^6}{(230 * 10^3)^2} = 1.89 * 10^{-3}$$

Using the per unit value of transmission line 4-5 resistance from Table 3.1, calculate R_{actual}, positive sequence resistance R_1, and zero sequence resistance R_0 as follows:

$$R_{actual_{4-5}} = 529 * 0.010 = 5.29\, \Omega$$

$$R_{\Omega/km_{4-5}} = \frac{R_{actual_{4-5}}}{Line\ length} = \frac{5.29\, \Omega}{116.798\, km} = 0.04529\, \Omega/km$$

$$R_{1(4-5)} = 0.04529\, \Omega/km$$

$$R_{0(4-5)} = 0.13587\, \Omega/km$$

Similarly, using the per unit value of transmission line 4-5 inductance from Table 3.1, calculate L_{actual}, positive sequence inductance L_1, and zero sequence inductance L_0 as:

$$X_{actual_{4-5}} = 529 * 0.085 = 44.965$$

$$L_{actual_{4-5}} = \frac{44.965}{2*\pi*50} = 0.14312 \text{ H}$$

$$L_{H/km_{4-5}} = \frac{L_{actual_{4-5}}}{Line\ length} = \frac{0.14312\ H}{116.798\ km} = 1.2254\ mH/km$$

$$L_{1(4-5)} = 1.2254\ mH/km$$

$$L_{0(4-5)} = 3.67629\ mH/km$$

Similarly, using the per unit value of transmission line 4-5 susceptance from Table 3.1, calculate C_{actual}, positive sequence capacitance $C_{1,}$ and zero sequence capacitance C_0 as follows:

$$B_{actual_{4-5}} = 1.89 * 10^{-3} * 0.176 = 3.3264 * 10^{-4}$$

$$C_{actual_{4-5}} = \frac{3.3264 * 10^{-4}}{2*\pi*50} = 1.058826\ \mu\text{F}$$

$$C_{F/km_{4-5}} = \frac{C_{actual_{4-5}}}{Line\ length} = \frac{1.058826\ \mu\text{F}}{116.798\ km} = 9.06544\ nF/km$$

$$C_{1(4-5)} = 9.06544\ nF/km$$

$$C_{0(4-5)} = 27.196\ nF/km$$

Likewise, the R_1, R_0, L_1, L_0, C_1, and C_0 values of all transmission lines are calculated. The obtained values and each transmission line length have been tabulated in Table 3.3. The values have been calculated for the 50 Hz frequency system. For different frequencies, the values obtained will change. The transmission line parameters values obtained have been used to simulate the IEEE 9 Bus system on the MATLAB Simulink environment, as shown in Figure 3.2. Each transmission line has been modeled using distributed line parameters by dividing each line into ten equal sections.

Table 3.3 Transmission line length, positive and zero sequence resistance, inductance, and capacitance.

Line	**Length** **km**	$\mathbf{R_1}$ Ω/km	$\mathbf{R_0}$ Ω/km	$\mathbf{L_1}$ mH/km	$\mathbf{L_0}$ mH/km	$\mathbf{C_1}$ nF/km	$\mathbf{C_0}$ nF/km
4-5	116.798	0.04529	0.13587	1.22542	3.67629	9.06544	27.1963
4-6	115.131	0.07811	0.23433	1.34554	4.03664	8.25767	24.7730
7-5	211.955	0.07986	0.23959	1.27904	3.83714	8.68702	26.0610
7-8	98.907	0.04546	0.13638	1.22576	3.67729	9.06462	27.1938
9-8	138.603	0.04541	0.13625	1.22459	3.67377	9.07331	27.2199
9-6	235.579	0.08757	0.26272	1.21511	3.64534	9.14408	27.4322

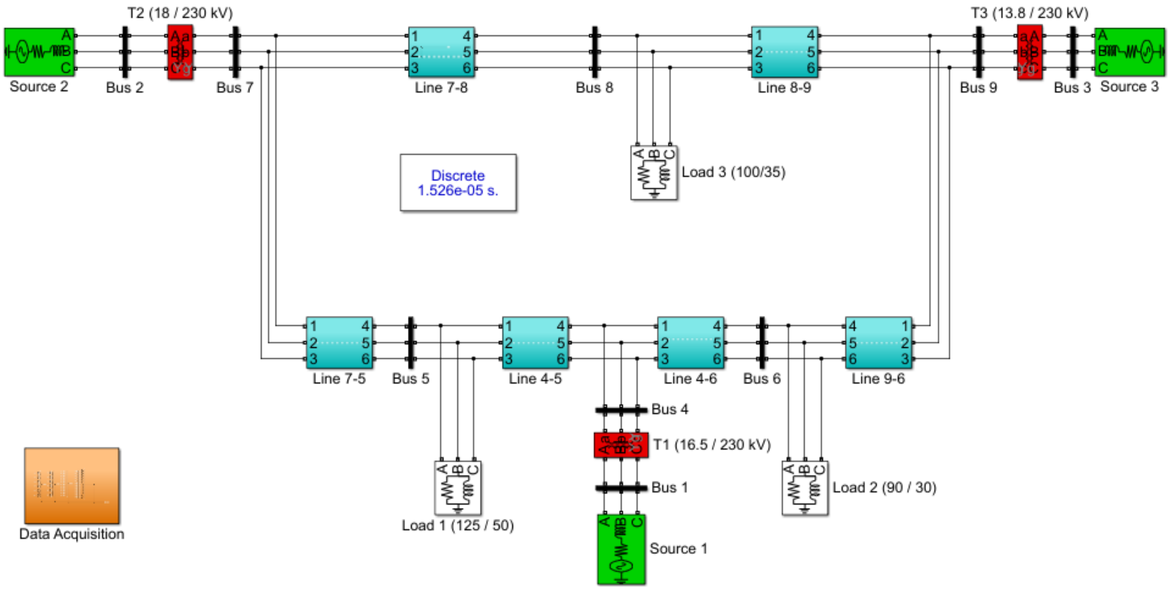


Figure 3.2 Simulated model of the IEEE 9 Bus System on MATLAB Simulink 2021b.

3.4 Renewable Energy Source Integration Considerations

The pursuit of green and clean energy led to the ongoing integration of large-scale RES into the existing transmission and distribution networks across the nations. Among all renewable sources, solar PV power generation capacity alone has reached 849 GW globally, according to the International Renewable Energy Agency (IRENA) 2021 report. Being a significant source of future energy generation, the integration of large-scale solar plants into the power grid is increasing over time, especially in countries like India.

Before integrating RES into the power grid, conducting optimal placement and sizing analysis is imperative. This ensures various objectives such as voltage stability, grid reinforcement, minimization of power loss and on-peak operation cost, and improvement of load factor [9]. Determining the optimal location (bus) and size (generation capacity in MW) of RES integration in a transmission network is essential for efficient operation. Various methodologies are available in the literature for this purpose. In the case of the IEEE 9 Bus system under study, RES integration has been strategically placed after carefully considering optimal placement and optimal maximum sizing analysis [10]. According to Lyapunov exponent estimation, Bus number 7 is the best location for RES integration, while Bus number 5 is the second-best location. The considered maximum RES size at these locations is 30MW [10]. However, the RES integration can be of the maximum allowed size or some lower values. Therefore, three sizes of RES have been integrated into the power system, i.e., 10MW, 20MW, and 30MW. The smaller sizes than the maximum allowable penetration have been considered to incorporate practical issues such as installation time and land availability limits at the point of optimal placement. Given two optimal placement locations with the possibility of three different sizes of RES integration at each, the RES integration into the IEEE 9 bus system can occur in nine different ways. Therefore, instead of opting for a single combination, such as the maximum allocation of 30MW at both buses, the study examines six combinations of different total MW power generations (i.e., 10MW, 20MW, 30MW, 40MW, 50MW, 60MW), treating each as an independent case. Six combinations out of nine are considered to avoid redundancy in the results.

3.5 Solar Photovoltaic Plant Modeling

In this dissertation, the solar PV plant-based RES has been taken under study. The integrated solar-based RES has been modeled, incorporating standard temperature and irradiance variations. The temperature and irradiance values, throughout the day and year, directly affect the power output from solar PV plants. Thus, the amount of power fed into the grid is never constant and keeps fluctuating. Thereby, on the occurrence of a fault, the fault current level may vary for the fault at the same location [13]. Hence, considering temperature and irradiance variations is essential while analyzing solar PV-based RES integration in transmission networks [79].

The photovoltaic solar cells can convert the irradiance energy of the sun obtained at the earth's surface into electrical energy. By using optimization techniques such as Maximum Power Point Tracking (MPPT), the conversion from solar energy to electrical energy can be maximized and fed to the electrical grid, which will be a zero-emission form of power source. The power generated by solar plants can be fed to the grid as per electricity board regulations either at low voltage, i.e., 400V, or at medium voltage, i.e., 18kV, using a step-up transformer. This can be further stepped up to match the voltage level of the point of connection [224].

The design of solar PV power plants requires detailed modeling of (1) a PV array of PV modules grouped in series and parallel strings to obtain the rating of power needed from the plant. (2) A DC-DC boost converter to convert the output voltage of the PV array to the rated voltage of the inverter. (3) A three-phase DC-AC inverter to convert input DC voltage to three-phase AC voltage. (4) A three-phase step-up transformer steps up the low voltage from the inverter to the high voltage to feed the grid at the rated voltage. (5) An MPPT controller to ensure maximum power generation from solar PV plants.

3.5.1 Design of PV array

Numerous solar PV array modules are available on MATLAB Simulink 2021b environment, such as Apollo, Hyundai, Mitsubishi, Sanyo, Sharp, SunPower, Suntech, and many more. In this study, a solar PV plant has been modeled using the SunPower SPR-310E-WHT-D module of a solar PV array. The open circuit voltage V_{oc} and short circuit current I_{sc} of this module are 64.6V and 6.14A, respectively. The maximum power generation P_{MP} from this module can be 315.072W. The voltage at maximum power point V_{MP} is 54.7V, and the current at maximum power point I_{MP} is 5.76A.

The I-V and P-V characteristics of a single PV module have been presented in Figure 3.3. The solar PV plants of the required rating will be formed by connecting several PV modules in series-parallel combinations. Further, IGBT switches, capacitors, and inverters' voltage and current ratings must be calculated. Several PV arrays will be connected in parallel to form the desired rating solar plant to limit the rating of components.

For the exact modeling of the solar power plant, D.L. Popa et al. (2016), work on the design and simulation of a 10 MW solar PV plant can be referred to [224]. Thus, the required power rating of a solar PV plant can be modeled by properly selecting parallel strings and a series of connected modules per string. However, the power obtained at the grid is lower than the maximum power generated by the PV module due to power loss in DC-DC converters and inverters.

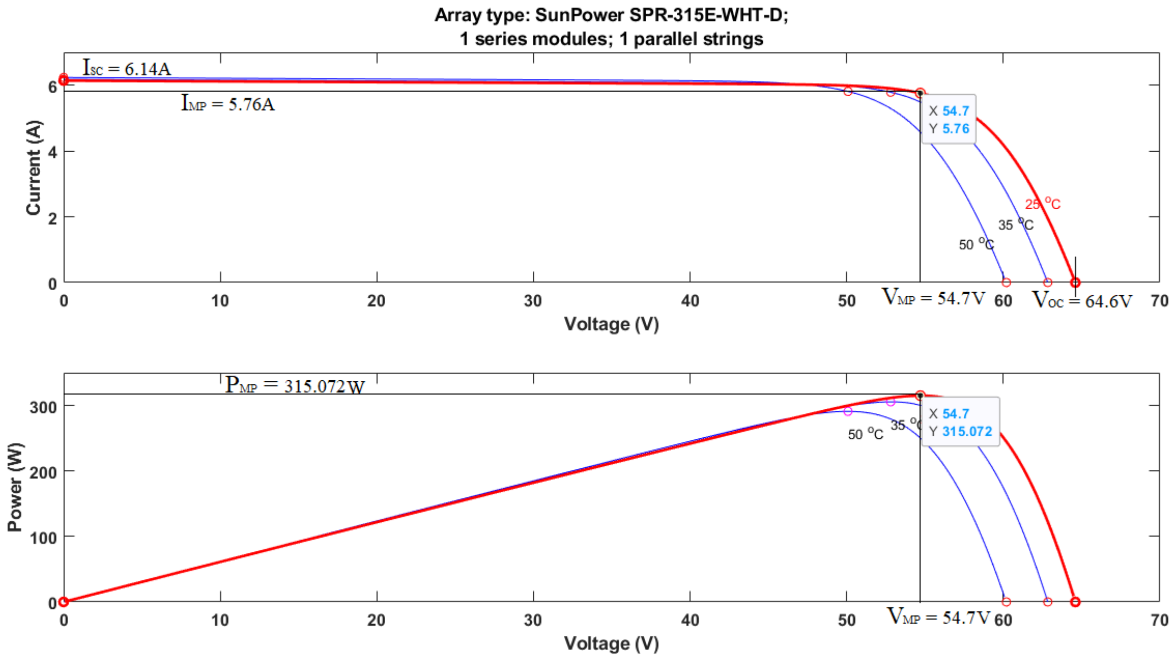


Figure 3.3 I-V and P-V characteristics of a single PV module.

The DC-DC boost converter input voltage value and the available inverter power rating determine the number of necessary series-connected modules per string and parallel strings. The rated DC input voltage for the boost converter is usually chosen as half the output voltage, i.e., DC-Link voltage [224]. Considering that the output voltage of a string of series connected PV modules is the sum of all PV modules, and also considering the minimum DC-Link voltage for the inverter, one can obtain the number of PV modules connected in series (N_{ser}) based on:

$$Nser = \frac{1}{2} * \frac{V_{DC-Link}}{V_{MP}} \qquad \ldots(3.7)$$

Where: N_{SER} = number of necessary PV modules connected in series.

$V_{DC\text{-}Link}$ = the DC link voltage at the inverter input.

V_{MP} = PV module voltage at the maximum power point.

Proper selection of inverter input voltage and current is necessary. A lower input voltage will lead to higher current flowing through the DC-DC boost converter, necessitating higher rating IGBT switches for operation. The minimum DC-Link voltage for the inverter can be calculated as follows:

$$V_{DC-Link} \geq 2\sqrt{2} * V_{PHASE} \qquad \text{…(3.7)}$$

Where: V_{PHASE} = the RMS value of the phase voltage at the inverter's output

Thus, the maximum power from a series-connected PV string will be:

$$P_{STRING} = N_{SER} * P_{MP} \qquad \text{…(3.8)}$$

Once the number of series connected modules is set, the number of parallel strings is computed based on the rated power of the available inverter. Thus, the necessary number of strings in parallel based on the rating of the inverter will be:

$$N_{PAR} = \frac{P_{INV}}{P_{STRING}} \qquad \text{…(3.9)}$$

If the available inverter rating is lower than the required plant rating, multiple inverters can be used and connected in parallel. Thus, to form the required size power plant P_{PLANT} i.e., 10MW, 20MW, and 30MW, the needed number of inverters will be $N_{INV,}$ and PV arrays (N_{ARRAYS}) will be:

$$N_{INV} = N_{ARRAYS} = \frac{P_{PLANT}}{P_{INV}} \qquad \text{…(3.10)}$$

3.5.2 Design of DC-AC Inverter

Inverter sizing is done considering the available inverter rating P_{INV} and PV arrays peak output power for a simple two-level inverter with IGBT switches available in markets. The rated RMS output voltage, power, efficiency, switching frequency, and overload factor must be known for further design. For an 85% efficient inverter, the RMS current through the inverter will be:

$$I_{INV\_RMS} = \frac{P_{INV}}{3 * V_{PHASE}} * \frac{100}{85} \qquad ...(3.11)$$

$$I_{INV\_PEAK} = I_{INV\_RMS} * \sqrt{2} \qquad ...(3.12)$$

To decide the size of the PWM coils inductance value, it is assumed that 5% ripple current of the peak value of injected current flows, thus the peak-to-peak value of current ripple is:

$$I_{INV\_RIPPLE\_PEAK} = 0.05 * I_{INV\_PEAK} \qquad ...(3.13)$$

Thus, the PWM coil inductance value can be calculated as:

$$L_{INV\_PWM} = \frac{\sqrt{3} * V_{DC}}{12 * \delta * f_{SW\_INV} * I_{INV\_RIPPLE\_PEAK}} \qquad ...(3.14)$$

Where: δ = the overload factor

f_{SW_INV} = the switching frequency of the inverter

The overload factor δ can vary between 120% to 180%. The selected value for this dissertation work is 150%. The higher switching frequency offers lower total harmonic distortion (THD) for the injected current and output voltage. The average value of the current is 63.6% of the peak value or 90% of the RMS value:

$$I_{INV\_AVG} = 0.9 * I_{INV\_RMS} \qquad ...(3.15)$$

The peak-to-peak ripple value of the DC-Link voltage is 5%:

$$V_{RIPPLE\_PEAK} = 0.05 * \ V_{DC-LINK} \qquad \ldots(3.16)$$

Thus, the required value of the DC-Link capacitor will be:

$$C_{DC-LINK} \ \geq \frac{I_{INV\_AVG}}{2 * \omega * V_{RIPPLE\_PEAK}} \qquad \ldots(3.17)$$

Where: ω = the angular frequency corresponding to 50Hz grid.

If the calculated value of the $C_{DC\text{-}LINK}$ capacitor is much higher than the one available in the market, then several capacitors can be connected in parallel to achieve the large capacitance needed at the rated voltage value of the DC-Link capacitor. The rated voltage of the DC-Link capacitor i.e., V_{RATED_CDC} is recommended to be at least 20% higher than the operating voltage:

$$V_{RATED\_CDC} \ \geq \ 1.20 * \ V_{DC-LINK} \qquad \ldots(3.18)$$

3.5.3 Design of DC-DC Boost Converter

Based upon the values of DC-DC converter input voltage V_{PV}, DC-Link voltage $V_{DC\text{-}Link,}$ I_{INV_RMS}, which is equal to I_{OUT_RMS} of boost converter, and switching frequency of boost converter f_{SW_BOOST}, the maximum duty cycle d_{max} can be calculated which is generally 0.6. The boost converter generally has 90% efficiency (η). Thus, the maximum allowable duty cycle can be calculated as:

$$d_{max} = \ 1 - \frac{V_{MIN} * \eta}{V_{DC-Link}} \qquad \ldots(3.19)$$

Thus, the maximum current through the IGBT will be:

$$I_{IGBT\_MAX} = 2 * I_{OUT\_RMS} * (d_{max} + 1) \quad \text{…(3.20)}$$

Sizing the L_{BOOST} coil inductance is done considering a 5% ripple current from the maximum current through the IGBT. Hence the peak-to-peak value of the current ripple is:

$$I_{RIPPLE\_PEAK} = 0.05 * \ I_{IGBT\_MAX} \quad \text{…(3.21)}$$

The minimum value of L_{BOOST} coil inductance will be:

$$L_{Boost} \geq \frac{V_{MIN} * (V_{DC-LINK} - V_{MIN})}{I_{RIPPLE\_PEAK} * f_{SW\_BOOST} * V_{DC-LINK}} \quad \text{…(3.22)}$$

Where: V_{MIN} = Minimum input voltage to Boost Converter

The exact value of the boost converter inductor L_{BOOST} and capacitor C_{BOOST} can be calculated using equations 3.23 and 3.24.

$$L_{BOOST} = \frac{d_{max} * V_{PV}}{f_{SW\_BOOST} * I_{RIPPLE\_PEAK}} \quad \text{…(3.23)}$$

$$C_{BOOST} = \frac{d_{max} * I_{OUT\_RMS}}{f_{SW\_BOOST} * V_{RIPPLE\_PEAK}} \quad \text{…(3.24)}$$

Thus, the DC-DC boost converter has been modeled using calculated values of L_{BOOST} and C_{BOOST}, along with the dedicated MPPT control through perturbation and observation control. To design 10MW, 20MW, and 30MW solar PV plants, firstly, several small modules of available P_{INV} ratings were modeled with separate DC-DC converters and inverters. Then, the output from each module was connected in parallel through a transformer to obtain the required rating solar plant. Modeling smaller modules will reduce the ratings of the required boost converters and inverters.

3.6 RES Integration to IEEE 9 Bus System

Optimal placement and sizing analysis are essential before integrating RES into the power grid to ensure voltage stability, grid reinforcement, minimized power loss, on-peak operation cost, and an improved load factor [9]. The optimal placement (bus location) and size (generation capacity in MW) of RES integration into a transmission network can be determined for a given transmission network. Several methodologies are available in the literature for optimal placement and sizing of RES into a transmission network. The RES integrated into the IEEE 9 Bus system under study has been placed after considering optimal placement and maximum allowable sizing analysis [10], [222]. Bus number 7 (Bus 7) is identified as the best location for RES integration, followed by bus number 5 (Bus 5) based on the Lyapunov exponent estimation analysis presented in [10]. The assumed maximum allowable RES size at these locations is 30MW [10]. However, the RES integration may be of the maximum allowed size or some lower value. Therefore, in the presented study, a solar PV plant with penetration ranging from 10MW to 30MW is placed on these buses. The power generation from solar PV plants depends on two significant variables, i.e., temperature and irradiance values, at the point of installation. The variations in weather conditions due to the rotations and revolutions of the sun throughout the day and year cause temperature and irradiance to vary widely. Therefore, the power fed by solar-based RES plants is not constant and varies widely [225]. However, the studies available in the literature for fault diagnosis of RES integrated systems considered RES as a constant source. Hence, in the proposed study, solar-based RES included in the grid incorporates temperature and irradiance effects on power generation from these plants. Thus, the presented study has been done, representing real power systems more closely. Solar PV plants of 10MW, 20MW and

30MW ratings have been modeled, and the power generation from these plants at various temperatures and irradiance values have been recorded. Table 3.4 lists the power fed from the modeled solar PV plant to the grid. Further, the sources of these recorded power were then integrated into the IEEE 9 Bus system to imitate a solar PV plant of varying power generation sources.

Table 3.4 Power fed to the grid from different-sized solar PV plants at varying temperatures and irradiances.

Plant Rating		**10MW**	**20MW**	**30MW**
Temperature	**Irradiance**			
50℃	1000W/m^2	8.67655	17.46925	26.52015
35℃	1000W/m^2	9.2378	18.4908	27.702
25℃	1000W/m^2	9.4962	19.0095	28.47815
25℃	600W/m^2	5.68195	11.3734	17.0392
25℃	300W/m^2	2.80155	5.60785	8.40085
25℃	100W/m^2	0.90155	1.805	2.70465

Thus, the RES integrated IEEE 9 Bus system models were simulated on MATLAB for three different sizes of solar PV plant-based RES at BUS 7 and BUS 5, which have been defined as cases in further study:

Case 1: 10MW at Bus 7

Case 2: 10MW at Bus 7 and 10MW at Bus 5

Case 3: 20MW at Bus 7 and 10MW at Bus 5

Case 4: 20MW at Bus 7 and 20MW at Bus 5

Case 5: 30MW at Bus 7 and 20MW at Bus 5

Case 6: 30MW at Bus 7 and 30MW at Bus 5

The simulated model of the RES integrated IEEE 9 Bus system has been shown in Figure 3.4. The subsystems RES-1 and RES-2 present different-sized solar PV plants, which have been modeled as per the designing procedure discussed in section 3.5 of this chapter. The

RES was integrated into the IEEE 9 Bus system at the rated voltage level of the point of connection in the system, i.e., 230kV, using a step transformer.

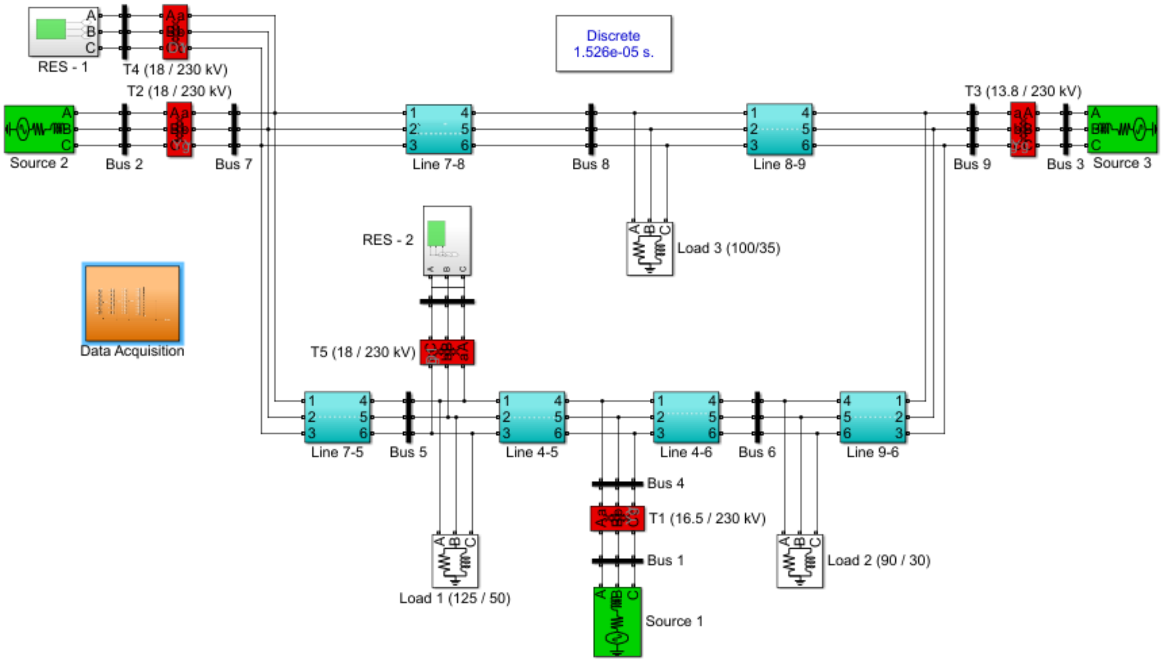


Figure 3.4 Simulated model of RES integrated IEEE 9 Bus System at Bus 7 and 5 on MATLAB Simulink 2021b.

3.7 Fault Attributes

Various attributes may affect the nature of fault current on fault occurrence, such as fault type, faulty phase, fault resistance, fault inception angle, and fault distance from measurement buses. The discussion about these attributes has been detailed in this section to present their range of values and underline how they affect the fault's current nature.

3.7.1 Faulty Phase and Fault Types

Five types of permanent faults may occur in power systems: SLG, DLG, LL, LLLG, and LLL. However, their occurrence may be classified into 11 types, i.e., RG, YG, BG, RYG, YBG, RBG, RY, YB, RB, RYBG, and RYB. Here, R, Y, and B represent the transmission line phases. Hence, this study considers all possible combinations of fault occurrences during fault data generation.

3.7.2 Fault Resistance

The fault resistance means the opposition to the flow of current introduced by the material with which lines come into contact upon the occurrence of the fault. For ground faults, this fault resistance is earth resistivity, while for LL or LLL faults, fault resistance is the material with which two lines or cables are shorted, such as wood or animal. Earth resistivity is an electrical characteristic of the ground and is very important while calculating the zero-sequence impedance of the transmission line. The value of earth resistivity varies significantly with soil type, as peat soil earth resistivity ranges (from 200 to 1200) typically taken as 200Ωm, adobe clay (2 to 200) typically 40Ωm, boggy ground (2 to 50), typically 30Ωm, moist gravel (50 to 3000) typically 1000Ωm, sandy ground (50 to 3000) typically 200Ωm, stony or rocky ground (100 to 8000) typically 2000Ωm and concrete ground (50 to 300) typically 150Ωm. Thus, in this dissertation, we have taken the fault resistance range from 5Ω to 8000Ω for fault data generation to accommodate low- and high-impedance faults [94]. The fault resistance with low or high impedances shows different characteristics. Depending on the fault resistances, the fault current level varies proportionally [17]. Therefore, a wide range of fault resistances has been considered during data generation, as listed in Table 3.5.

3.7.3 Fault Inception Angle

The electrical degree at which a fault occurs is known as the fault inception angle (FIA). It varies from 0° to 360°. The instant at which a fault occurs, i.e., FIA, changes the transient level after fault occurrence. Different faults face maximum transient at different FIAs [226]. Thus, fault data has been generated for various values of fault inception angles in this study, as given in Table 3.5.

3.7.4 Fault Distance

In real-world power systems, a fault may occur at any point; however, to include sufficient diversity in the fault data for ML training, all six lines of the IEEE 9 Bus system have been divided into ten equal sections.

3.8 IEEE 9 Bus System Fault Dataset Generation

The fault current's signature is crucial for fault diagnosis; thus, once the IEEE 9 Bus System has been modeled, the fault dataset has been prepared for a wide range of attribute variations to incorporate numerous scenarios that may vary the fault current after fault occurrence. To create a fault database, fault occurrences are simulated on all six transmission lines of the network, considering various fault attributes that affect the nature of the fault current. Table 3.5 enlists fault attributes for which the IEEE 9 Bus System fault database has been formed.

The fault current depends on factors such as faulty phase, fault types, fault resistance, fault inception angle, and fault distance from buses. On the occurrence of a fault, the three-phase voltages and currents were recorded at six buses. All 36 voltages and currents have been collected at a sampling frequency of 1.6 kHz. The collected fault data comprises three cycles: one pre-fault cycle and two post-fault cycles. Overall, 15840 fault data samples have been generated for the IEEE 9 bus system.

Table 3.5 Fault dataset description for the standard IEEE 9 Bus system.

ATTRIBUTES	Cases
Fault Types	5 (SLG, DLG, LL, LLLG and LLL)
Fault Scenarios	11 (R-G, Y-G, B-G, R-Y-G, Y-B-G, R-B-G, R-Y, Y-B, R-B, R-Y-B-G, R-Y-B)
Faulted Lines	6 (Lines 7-8, 8-9, 7-5, 5-4, 4-6, 6-9)
Fault Locations	Each line is divided into ten equal sections
Fault Inception Angle	0°, 30°, 60°
Fault Resistances	5, 10, 20, 50, 100, 700, 4500, 8000Ω

3.9 RES Integrated IEEE 9 Bus System Fault Dataset Generation

Solar PV plants of 10, 20, and 30MW ratings were simulated on MATLAB individually to get power generation from these plants under varying temperatures and irradiance values. The power generation obtained from these solar plants has been tabulated in Table 3.4. Later, solar PV plants of varying sizes were integrated at Bus 7 and 5. The integrated solar PV plant has been modeled considering standard temperature and irradiance variation throughout the day and year [227]. The key features influencing solar plants' power output are solar irradiance, operating temperature, tilt angle, and load matching for maximum power. Among these, temperature and solar irradiance majorly influence power generation. The power output at lower temperatures is higher, while efficiency drops considerably at very high temperatures; therefore, electrical load adjustment and excess heat removal are needed for optimum power generation. The power output from PV plants has an almost direct linear relationship with solar irradiance [228]. Table 3.6 enlists the range of fault attributes for which the fault dataset for the RES integrated IEEE 9 Bus system has been generated.

Table 3.6 Fault dataset description for the RES integrated IEEE 9 Bus system.

ATTRIBUTES	**Cases**
Fault Types	5 (SLG, DLG, LL, LLLG and LLL)
Fault Scenarios	11 (R-G, Y-G, B-G, R-Y-G, Y-B-G, R-B-G, R-Y, Y-B, R-B, R-Y-B-G, R-Y-B)
Faulted Lines	4 (Bus 7-8, 8-9, 7-5, 4-5)
Fault Locations	Each line is divided into five equal sections
Fault Inception Angle	0°
Fault Resistances	10, 50, 700, 8000Ω
RES Locations	Bus 7 and Bus 5
RES Sizes / Cases	Six combinations Case-1: B7-10MW, Case-2: B7-10MW and B5-10MW, Case-3: B7-20MW and B5-10MW, Case-4: B7-20MW and B5-20MW, Case-5: B7-30MW and B5-20MW, Case-6: B7-30MW and B5-30MW
Irradiance	1000W/m2, 600W/m^2, 300W/m^2, 100W/m^2
Temperature	50°, 35°, 25°C

For fault data generation with RES integration, the RES is firstly placed on Bus 7 and then on Bus 5 in six different combinations, i.e., Bus 7-10MW, Bus 7-10MW and Bus 5-10MW, Bus 7-20MW and Bus 5-10MW, Bus 7-20MW and Bus 5-20MW, Bus 7-30MW, and Bus 5-20MW, and lastly Bus 7-30MW and Bus 5-30MW. Fault data has been generated considering one case or size combination at a time. The selection of buses for RES placement has been made according to the optimal placement study of RES on the IEEE 9 Bus system [10].

In the case of the RES integrated system, fault data has been generated for faults on Lines 4-5, 7-8, 7-5, and 8-9. Among them, lines 4-5, 7-8, and 7-5 are directly connected to the RES integrated bus. Meanwhile, the other three lines, Lines 8-9, 4-6, and 9-6, are not connected to RES. Therefore, we have taken Line 8-9 among the lines not directly connected to the RES integrated bus, as the remaining lines will be affected similarly. Figure 3.5 presents the steps followed in fault data generation for both the standard IEEE 9 Bus system and the solar PV plant-based RES integrated system. It also details the number of fault data samples generated for the study.

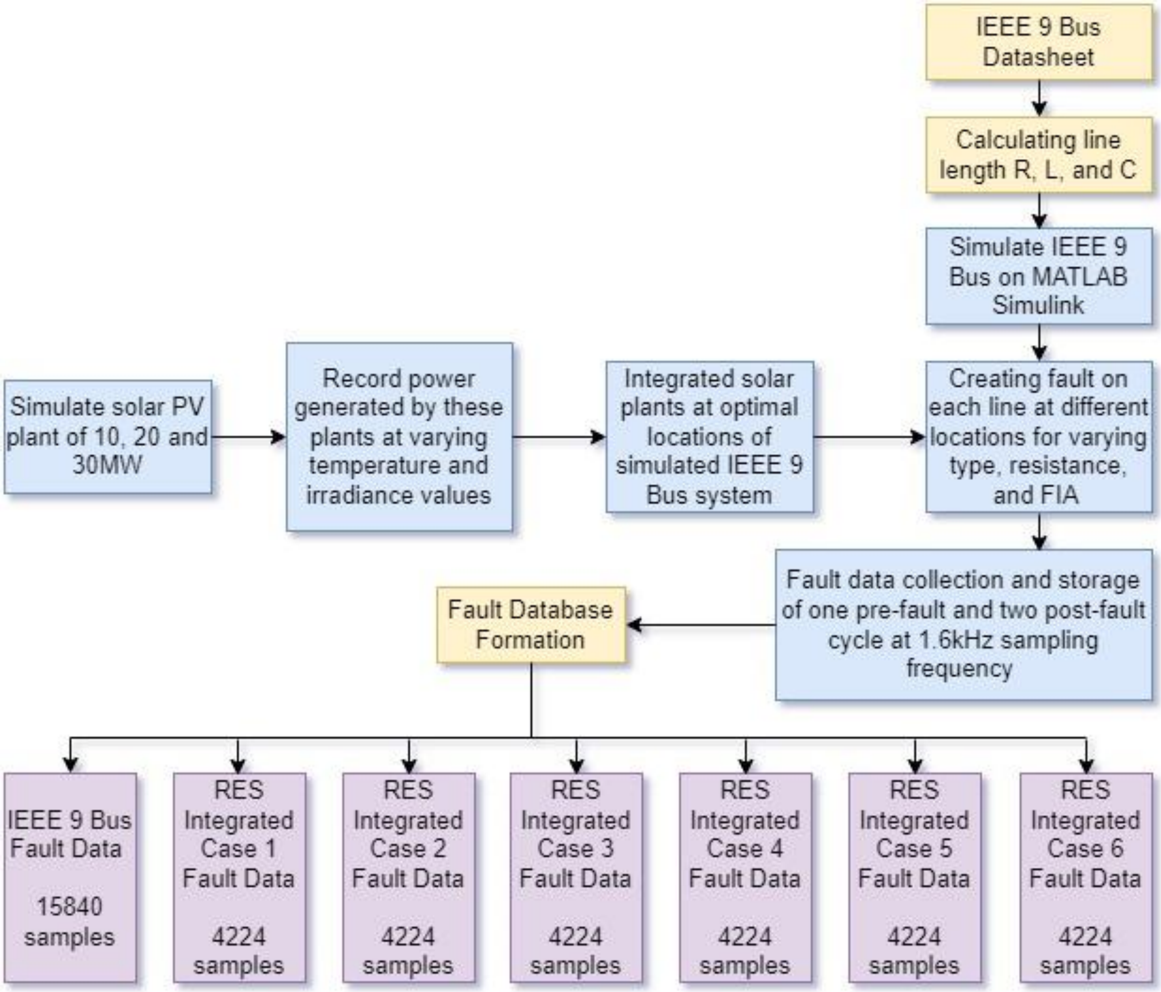

IEEE 9 Bus Datasheet
Calculating line length R, L, and C
Simulate IEEE 9 Bus on MATLAB Simulink
Simulate solar PV plant of 10, 20 and 30MW
Record power generated by these plants at varying temperature and irradiance values
Integrated solar plants at optimal locations of simulated IEEE 9 Bus system
Creating fault on each line at different locations for varying type, resistance, and FIA
Fault data collection and storage of one pre-fault and two post-fault cycle at 1.6kHz sampling frequency
Fault Database Formation
IEEE 9 Bus Fault Data 15840 samples
RES Integrated Case 1 Fault Data 4224 samples
RES Integrated Case 2 Fault Data 4224 samples
RES Integrated Case 3 Fault Data 4224 samples
RES Integrated Case 4 Fault Data 4224 samples
RES Integrated Case 5 Fault Data 4224 samples
RES Integrated Case 6 Fault Data 4224 samples

Figure 3.5 Block diagram outlining the procedural steps for fault data generation and fault database formation.

3.10 Chapter Summary

This chapter provides details about the IEEE 9 Bus system transmission line per unit parameters, loads, generators, and transformers rating datasheet. Various formulas and detailed steps for calculating transmission line per kilometer resistance, inductance, and capacitance have also been summarized. Further, the complete design and modeling of each component of 10MW size solar PV plants have been discussed in detail. The modeled solar PV plant's integration into the IEEE 9 Bus system has been discussed with consideration of its optimal size and placement aspects. Lastly, various fault attributes that affect fault characteristics have been discussed. The steps involved in simulating the IEEE 9 Bus system model on MATLAB have been illustrated using the calculated line parameters. Moreover, fault database formation comprising the fault dataset of the IEEE 9 Bus system and the RES integrated IEEE 9 Bus system has been summarized in this chapter.

Chapter 4 PERFORMANCE ANALYSIS OF MACHINE LEARNING MODELS FOR TRANSMISSION LINE FAULT CLASSIFICATION AND LOCALIZATION

4.1 Introduction

This chapter examines the performance of various machine learning (ML) algorithms in classifying and localizing transmission line faults of the IEEE 9 Bus system. While many of the models examined in this chapter have already been employed for power system fault diagnosis in the literature, this chapter also introduces some previously unexplored models for power system fault diagnosis. Further, a comparative analysis of models' performance with and without dimensionality reduction has also been presented, which demonstrated the model’s ability to deal with high dimensional datasets. The comparative study demonstrated the efficacy of ML models for power system fault classification and localization.

The conventional fault diagnosis techniques rely on power system data obtained at Supervisory Control and Data Acquisition (SCADA) and Wide Area Monitoring Systems (WAMs). They use traditional impedance-based techniques for fault localization. However, these methods are characterized as tedious, complex, time-consuming, and computationally intensive, and they require mathematical modeling and domain proficiency. Another conventional method for fault localization is traveling wave-based techniques, which utilize transient information during fault occurrences. However, this approach is costly as it necessitates costly instruments to capture high-precision transient information for accurate fault localization. While impedance-based and traveling wave-based methods perform well, however, any changes in system configuration can affect their accuracy. Therefore, there is a growing need for intelligent paradigms to enable early fault identification and localization without relying on mathematical modeling and domain expertise.

As a solution, artificial intelligence (AI) and ML techniques are being investigated for diagnosing power system faults, offering potential suitability to real power systems. ML leverages past power system pre- and post-fault data to train models for automated fault identification and localization, facilitating self-healing systems. Trained models can make quick and accurate decisions without human intervention and possess adaptive capabilities to accommodate variations [6]. Rapid fault localization also aids maintenance operators in performing timely maintenance to restore normal power system operation. Successful fault detection and localization not only ensure faster line restoration and power supply continuity but also yield monetary benefits by preventing revenue loss due to electricity supply interruptions and conserving resources typically expended in manual fault location searches [6].

While numerous studies have reported fault classification and detection works with outstanding performances on power system networks, fewer studies on fault localization using ML regression exist in the literature. Power system fault data, consisting of line currents and voltages measured at different buses or network nodes, exhibits some degree of correlation despite variations in fault attributes. Bayesian Ridge regression has been identified as an ideal technique for datasets with multicollinearity, prompting its proposal and comparison with the Extra Tree (ET) regressor and other potential ML regression models for the localization of transmission network faults in this chapter [216]. Although XGBoost is a newer and more advanced ML ensemble technique, Bayesian Ridge regression and ET have been utilized in other fields, yet they remain unexplored for power system fault localization. Additionally, the outstanding performance of Random Forest (RF) in power system fault classification has motivated further exploration into other potential ensemble techniques, such as ET and XGBoost, which have not been extensively applied in power system fault diagnosis. XGBoost, in particular, has been noted for its rapid ensemble technique due to multithreading parallel computing [212]. Compared with RF, ET has been found to perform equally well and has lesser complexity [210].

4.2 Proposed Study

Power system fault diagnosis involves three functional steps, i.e., fault detection, classification, and localization.

Fault Detection: Prompt recognition of fault occurrence.

Fault Classification: Identification of the type of fault out of the five possible fault types, i.e., SLG, DLG, LL, LLL, and LLLG, and the involved faulty transmission line phases.

Fault Localization: Finding the accurate location of the fault on the identified faulty transmission line.

Since relays and circuit breakers are already installed at several points in the transmission network, they quickly detect faults and immediately isolate the faulty line from the remaining network. Various ML-based fault detection works are available in the literature. Thus, fault detection using ML algorithms has not been covered in this dissertation work. Once the fault has occurred and the transmission line is out of operation, it is challenging for the electricity maintenance operators to quickly identify the type of fault and locate the fault for quick maintenance and restoration of the faulty line. Thus, this dissertation mainly focuses on fault classification and localization. Identification of fault types is a must for impedance-based fault localization methods. The schematic diagram of the proposed study has been illustrated using Figure 4.1.

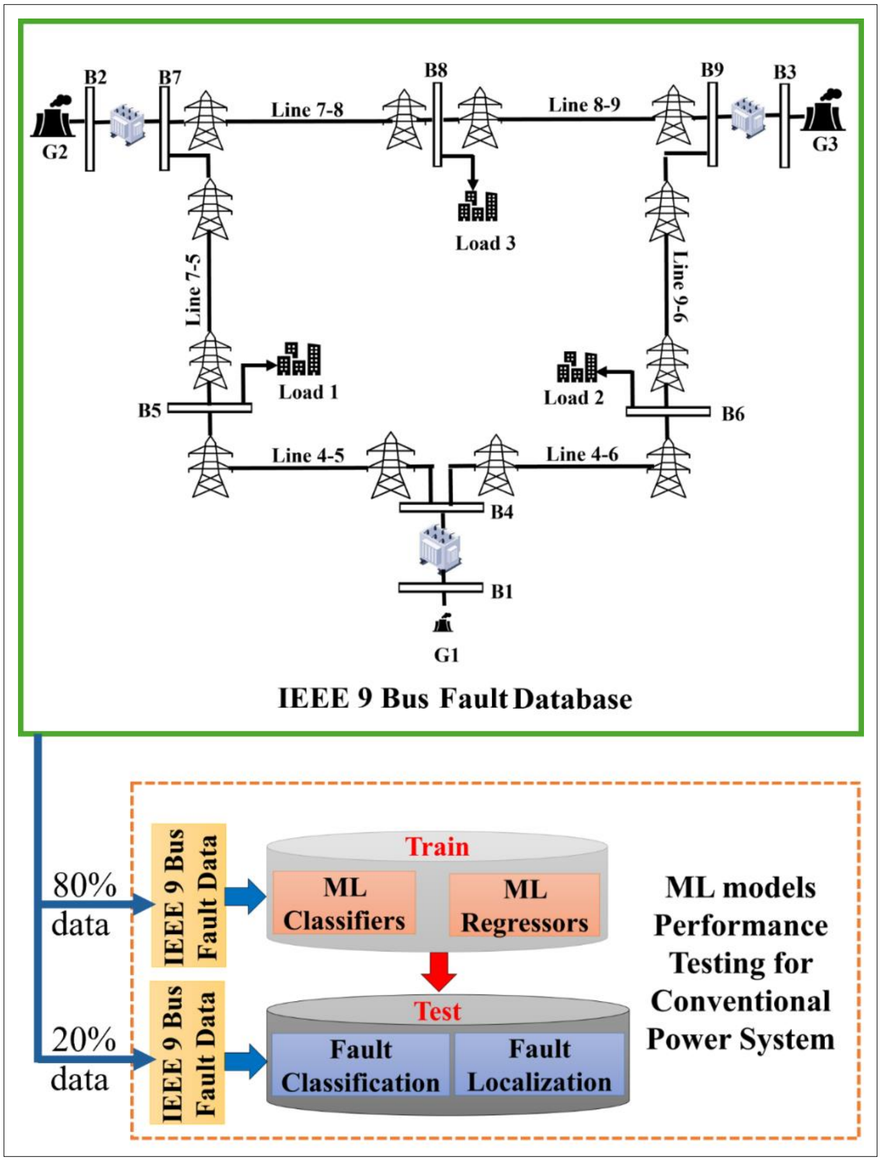


Figure 4.1 Schematic diagram of the proposed study.

The proposed study uses standard IEEE 9 bus system fault data representing conventional power systems. The conventional power system fault data has been used for training and testing several ML classification and regression models for fault classification and localization. Thus, the objective of this chapter is to analyze and compare the fault classification performance of XGBoost and Extra Tree with the potential ML classification models used in the literature for conventional power system fault classification. Also, to analyze and compare the localization performance of Extra Tree and Bayesian Ridge regression with potential ML regression models for location estimation of faults on transmission lines of conventional power systems.

4.3 Dataset Description

The IEEE 9 Bus system has been simulated on MATLAB Simulink environment, and various faults were simulated to generate fault data. The schematic single-line diagram of the IEEE 9 Bus system has been presented in Figure 4.2. A total of 15840 fault data samples were generated for the IEEE 9 Bus system, which has been referred to as a conventional power system. Among all datasets, SLG, DLG, and LL each have 4320 samples, while LLLG and LLL have 1440 samples each. Similarly, the fault data for each line is 2640 samples. Thus, during fault classification, the fault data has been split into 80% for training, i.e., 12672 samples, and the remaining 20% for testing, i.e., 3168 samples. However, during the fault localization process, we split each line's fault data into 80% training and 20% testing, with 2112 training samples and 528 testing samples. Thus, the training and testing of the ML classification models for fault classification and ML regression models for fault localization have been done using the above-mentioned data split. The fault data incorporates a wide range of fault resistance; thus, models are trained and tested for both high and low-impedance faults. Therefore, the models exhibiting good performance demonstrate their efficacy in handling high-impedance faults.

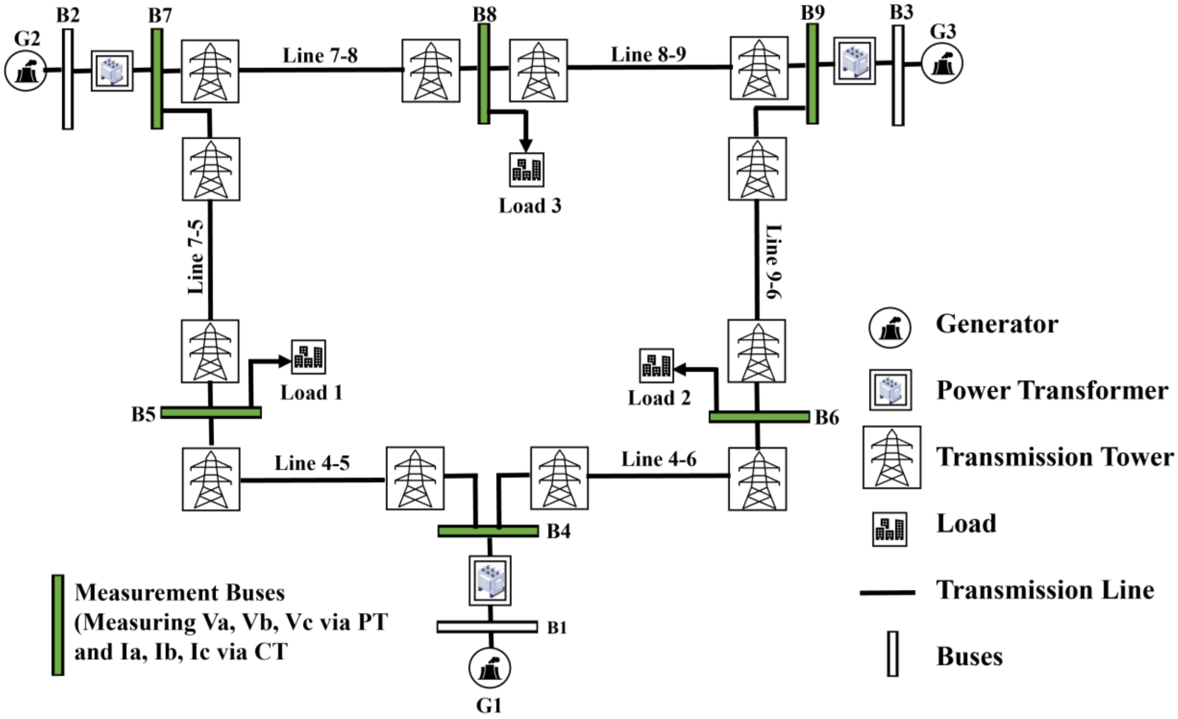


Figure 4.2 Schematic diagram of the IEEE 9 Bus System.

4.4 Conventional Power System Fault Classification

The standard IEEE 9 bus fault data represents a conventional power system. Since the conventional power system has been operational for a significant period, ample fault data representing all fault-type signatures are available for training. Several machine learning classifiers have been tested for their fault classification performance and have been reported in this section. The considered ML classifiers are trained and tested using the conventional power system / IEEE 9 Bus system fault database for their comparative fault classification performances. The steps followed for fault classification have been illustrated using Figure 4.3. Also, the effect of dimensionality reduction using principal component analysis, kernel principal component analysis, and linear discriminant analysis has been shown in this section.

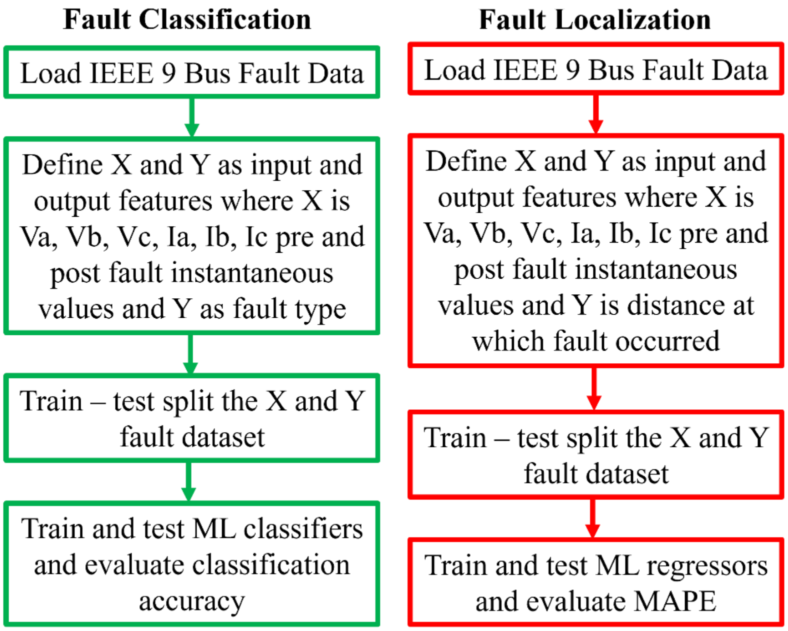

Fault Classification
Load IEEE 9 Bus Fault Data
Define X and Y as input and output features where X is Va, Vb, Vc, Ia, Ib, Ic pre and post fault instantaneous values and Y as fault type
Train – test split the X and Y fault dataset
Train and test ML classifiers and evaluate classification accuracy
Fault Localization
Load IEEE 9 Bus Fault Data
Define X and Y as input and output features where X is Va, Vb, Vc, Ia, Ib, Ic pre and post fault instantaneous values and Y is distance at which fault occurred
Train – test split the X and Y fault dataset
Train and test ML regressors and evaluate MAPE

Figure 4.3 Block diagram of the steps followed for fault classification and localization.

4.4.1 Fault classification

The ML models such as SVM, DT, RF, AdaBoost, ET, LR, Ridge, Gaussian Naïve Bayesian (GNB), KNN, Bagging, XGBoost, and multilayer perceptron (MLP) are trained for transmission line fault classification using voltage and current instantaneous values of one pre-fault and two post fault data. The values of hyperparameters taken while training them have been listed in Table 4.1. The ML models were trained and tested on the Jupyter Notebook platform, employing the Sklearn library [229]. The models used in the present study were tuned to give the best performance, for which a wide range of hyperparameter values were systematically investigated to achieve optimal outcomes across all models presented in the chapter. The nomenclature of model parameters shown in Table 4.1 is explained in the documentation of the scikit-learn application programming interface (API) [229].

Table 4.1 Parameters used for training the ML classifiers.

	ML Models	Parameters

Classification Models	**SVM**	kernel = rbf
		decision_function_shape = ovr
		gamma = scale
		C = 20

	DT	criterion = gini
		max_features = n_features

	RF	n_estimators = 110,
		criterion = gini
		max_features = sqrt

	AdaBoost	n_estimator = 70
		learning_rate = 1.0
		algorithm = SAMME.R

	ET	n_estimators = 120
		criterion = gini
		max_features = sqrt

	LR	penalty = l2
		C = 3.0
		solver = lbfgs
		max_iter = 300

	Ridge	alpha = 5.0
		solver = auto

	Gaussian NB	priors = none

	KNN	n_neighbors = 5
		metric = minkowski

	Bagging	n_estimators = 120

	XGBoost	base_score = 0.5
		booster = gbtree
		learning rate = 0.3

	MLP	hidden_layer_sizes = 100
		max_iter = 1000
		solver = adam
		activation = relu
		learning rate = constant

The testing accuracy and F1-score obtained for various fault types have been presented in Table 4.2. Additionally, Figure 4.4 presents the percentage testing accuracy for fault classification using bar plots. Ensemble methods such as ET, Bagging, RF, and XGBoost performed very well, except for AdaBoost. The classification accuracy of the RF, ET, Bagging, and XG algorithms is 100% for conventional power systems, as sufficient fault data is available. DT demonstrated strong performance overall, except for showing misclassification for LLL faults. Similarly, the KNN model exhibited satisfactory performance. However, SVM struggled to classify non-ground faults in comparison to ground faults.

Table 4.2 Performance of ML classifiers for fault classification.

Models	**Accuracy**	**F1-score SLG**	**F1-score DLG**	**F1-score LL**	**F1-score LLLG**	**F1-score LLL**
SVM	81.47	0.80	0.95	0.66	0.93	0.65
DT	99.87	1	1	1	1	0.99
RF	100	1	1	1	1	1
AdaBoost	86.74	1	0.88	0.83	1	0
ET	100	1	1	1	1	1
LR	94.03	0.92	0.99	0.91	1	0.98
Ridge	96.02	0.95	0.97	0.96	0.98	0.93
Gaussian NB	79.07	0.84	0.86	0.58	1	0.78
KNN	98.86	1	0.99	0.99	0.96	0.96
Bagging	100	1	1	1	1	1
XGBoost	100	1	1	1	1	1
MLP	85.95	0.61	0.81	0.68	1	0.74

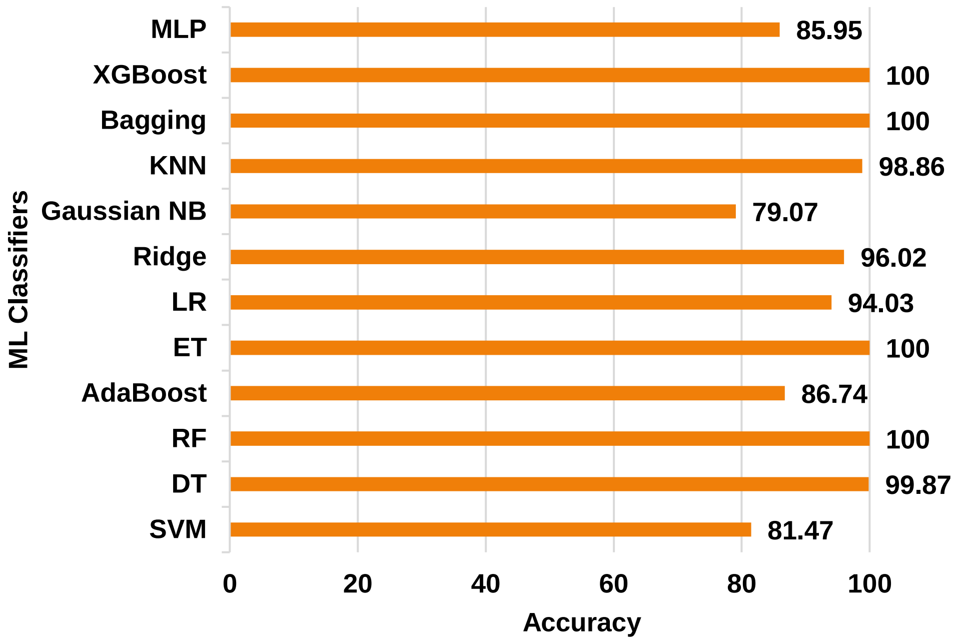


Figure 4.4 Testing accuracy (%) of ML models for fault classification.

4.4.2 Fault classification with dimensionality reduction

In the analysis, various ML models, including SVM, DT, RF, AdaBoost, ET, LR, Ridge, GNB, KNN, Bagging, XGBoost, and MLP, were examined for classifying transmission line faults with dimensionality reduction techniques. Dimensionality reduction methods such as Principal Component Analysis (PCA), Kernel PCA, and Linear Discriminant Analysis (LDA) were applied. The number of components for PCA and kernel PCA were taken as 18, while in the case of LDA it was 4. It was observed that the accuracy of tree-based models, except AdaBoost, remained largely consistent, with some decrease observed in certain techniques. AdaBoost's accuracy notably decreased. Conversely, the accuracy of KNN, SVM, and MLP increased when employing dimensionality reduction techniques, as listed in Table 4.3 and shown in Figure 4.5. Notably, with LDA, SVM achieved an accuracy of 99.40%, while MLP reached 100% accuracy with kernel PCA.

Table 4.3 Fault classification accuracy of ML models with and without dimensionality reduction techniques.

Models	Without DR	PCA	kPCA	LDA
SVM	81.47	89.84	89.99	99.40
DT	99.87	99.53	99.75	99.15
ET	100.00	99.97	99.97	99.62
RF	100.00	100.00	100.00	99.49
AdaBoost	86.74	59.22	47.73	65.03
LR	94.03	56.57	74.08	99.27
Ridge	96.02	61.77	62.85	97.66
GNB	79.07	73.55	73.58	98.99
KNN	98.86	99.46	99.46	99.72
Bagging	100.00	99.75	99.81	99.21
XGBoost	100.00	100.00	100.00	99.40
MLP	85.95	96.05	100.00	99.43

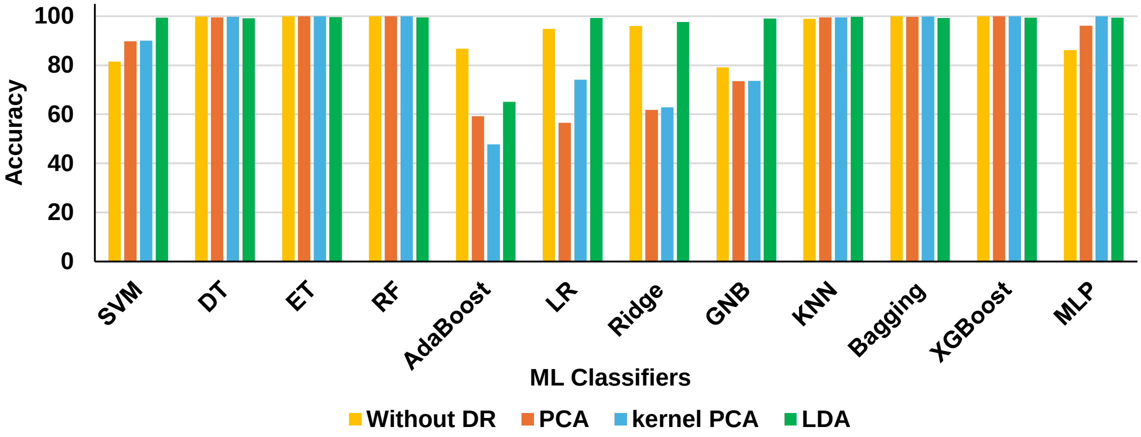


Figure 4.5 Testing accuracy (%) comparison between ML models for fault classification with and without dimensionality reduction techniques.

This suggests that the classification accuracy of tree-based models is unaffected by dimensionality reduction techniques, whereas SVM, KNN, and MLP benefit from such techniques, highlighting their reliance on feature engineering.

4.5 Conventional Power System Fault Localization

Several ML regressor models have been tested for their fault localization performance and have been reported in this section. The considered ML regressors are trained and tested using conventional power system / IEEE 9 Bus system fault database for their comparative fault localization performances. All six transmission lines of the IEEE 9 Bus system have been analyzed for fault localization using regression models. Each line fault data is split into 80% training and 20% testing; thus, training samples for each line localization analysis are 2112, and testing samples are 528. The steps followed for fault localization have been illustrated using Figure 4.3. Also, the effect of dimensionality reduction using principal component analysis has been shown in this section. The values of hyperparameters used for training the ML regressor models are listed in Table 4.4. The description of model parameters given in Table 4.4 is explained in the documentation of the scikit-learn application programming interface (API) [229]. Thus, details regarding each model's hyperparameters can be found on the API documentation webpage.

Table 4.4 Parameters used for training the ML regressors.

	ML Models	Parameters
Regression Models	**Bagging**	n_estimators = 120

	RF	n_estimators = 110
		criterion = squared error
		max_features = sqrt

	RT	criterion = squared error
		max_features = n_features

	KNR	n_neighbors =5
		metric = minkowski

	ETR	n_estimators = 120
		criterion = squared error

	BRR	max_iter = 300

	SVR	kernel = rbf
		decision_function_shape = ovr
		gamma = scale
		C = 300

4.5.1 Fault localization

Similarly, several ML regression models have been evaluated for the fault localization performance of all six transmission lines. In this evaluation, 80% of the data for each line fault was allocated for training, with the remaining 20% used for testing. The obtained location estimation MAPE of the studied models is listed in Table 4.5 and has been presented in Figure 4.6. Among the studied models, Bayesian Ridge regression (BRR) demonstrated the best performance, while RF, Bagging, and RT performed optimally. However, KNR and SVR exhibited inferior performance compared to the other models. Thus, the KNR regression model is not suitable for power system fault localization.

Table 4.5 Regression models localization MAPE for the conventional power system.

Models	Bagging	RF	RT	KNR	ET	BRR	SVR
Line 4-5	2.145	1.442	2.316	9.996	2.610	0.015	8.236
Line 4-6	1.845	1.555	1.691	10.483	1.800	0.015	7.976
Line 7-5	1.233	1.001	0.928	6.695	1.304	0.030	8.327
Line 7-8	2.462	2.056	2.525	10.806	1.955	0.008	8.702
Line 8-9	2.228	1.662	2.420	10.784	2.570	0.015	9.160
Line 9-6	1.409	1.035	1.286	7.069	0.974	0.040	7.648

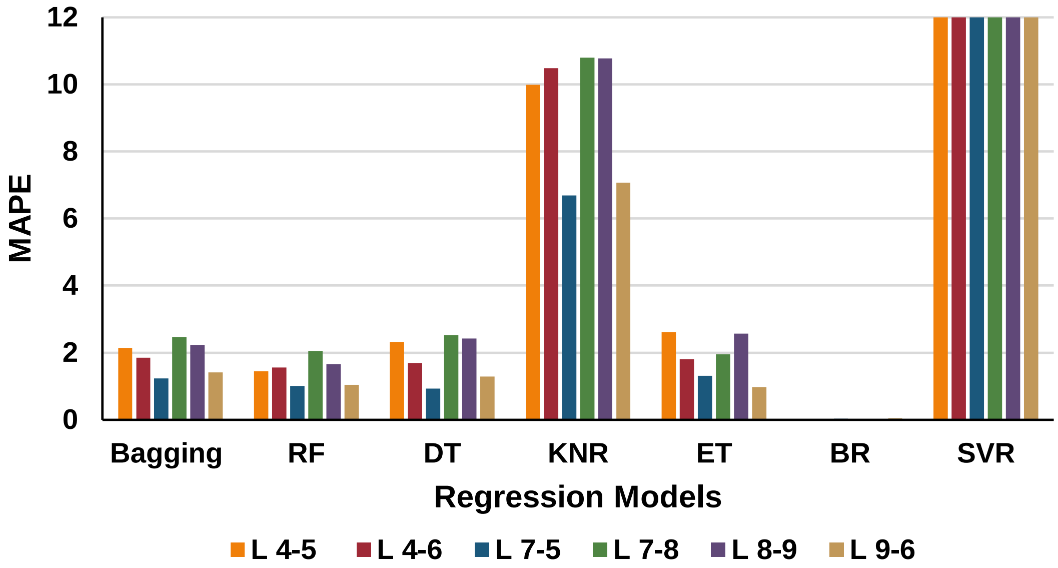


Figure 4.6 Regression models localization MAPE for the conventional power system.

The SVM performance was worse for both classification and localization, as its performance is highly dependent on feature selection and transformation techniques. Further multi-class classification is a complex process with SVM [230]. The present study utilizes voltage and current values without any feature engineering. Thus, SVM performance is poor, particularly for LL and LLL faults, which are non-ground faults. However, the performance of other models is notably good, even without feature engineering. This indicates that RF, bagging, ET, and BRR performance are not dependent on feature engineering techniques.

Thus, it can be deduced that the BRR model is the most suitable model for power system fault localization with sufficient data availability for training. The reason for BRR's superior performance is its ability to deal with multicollinearity. Multicollinearity is a statistical phenomenon that occurs when two or more independent variables in a regression model are highly correlated with each other. In other words, multicollinearity indicates a strong linear relationship among the predictor variables. The power system voltage and current data have a certain degree of relationship between them. To examine whether a certain correlation exists between bus voltages and currents of the power system. The correlation coefficient is determined by dividing the covariance by the product of the two variables' standard deviations.

$$Correlation = \rho = \frac{covariance(X,Y)}{\sigma_X \sigma_Y} \qquad \ldots(4.1)$$

The correlation coefficient has been computed for Bus 4 of the IEEE 9 Bus System when the power system was under normal operation. The obtained correlation coefficient is tabulated in Table 4.6. Any value other than 0 in the matrix shows that those variables are correlated. Values greater than 0 represent a positive correlation, while values lower than 0 represent a negative correlation. The diagonal element will always remain 1, as the variable itself will have a 100% correlation. It can be interpreted from the table that, for normal operation of the power system, the bus phase voltage is 100% correlated with its phase current. Also, all phase voltages and phase currents are correlated with each other. Also, for an increase in phase voltage A, phase voltages B and C decrease in the same ratio, such that the sum of phase voltage coefficients remains zero. A similar relationship can be seen for phase currents. Thus, significant information from phase voltage and current correlation coefficients can be drawn.

Table 4.6 Correlation between power system bus voltage and current for normal operation.

	Va4	Vb4	Vc4	Ia4	Ib4	Ic4
Va4	1	-0.5	-0.5	1	-0.473	-0.527
Vb4	-0.5	1	-0.5	-0.527	1	-0.473
Vc4	-0.5	-0.5	1	-0.473	-0.527	1
Ia4	1	-0.527	-0.473	1	-0.5	-0.5
Ib4	-0.473	1	-0.527	-0.5	1	-0.5
Ic4	-0.527	-0.473	1	-0.5	-0.5	1

When a fault occurs in the power system, the faulty phase voltage drops and the current increases, causing the three-phase voltages and currents to disbalance. Thus, the correlation coefficient between the bus voltage and current has changed, which has been presented in Table 4.7 for the phase A SLG fault. The bus 4 correlation coefficient between phase A voltage and current has changed from 1 to 0.381, while between phase B voltage and current has changed from 1 to 0.993; similarly, phase C voltage and current have changed from 1 to 0.997. Likewise, the correlation coefficient between other phase voltages and currents has also changed. However, due to the occurrence of the SLG fault, the system becomes unbalanced, and thus the sum of the correlation coefficients of phase voltages and currents is not 0. Although the correlation coefficient has changed significantly, there is still a certain level of correlation that exists between phase voltages and currents post-fault. Thus, multicollinearity exists in power system data; therefore, BRR can be used for power system fault localization. The correlation coefficient between the voltages and currents has been computed using Pearson's correlation test.

Table 4.7 Correlation between power system bus voltage and current for phase A SLG fault at Line 4-5.

	Va4	Vb4	Vc4	Ia4	Ib4	Ic4
Va4	1	-0.89	0.404	0.381	-0.91	0.223
Vb4	-0.89	1	-0.61	-0.19	0.993	-0.45
Vc4	0.404	-0.61	1	-0.65	-0.58	0.977
Ia4	0.381	-0.19	-0.65	1	-0.24	-0.79
Ib4	-0.91	0.993	-0.58	-0.24	1	-0.41
Ic4	0.223	-0.45	0.977	-0.79	-0.41	1

4.5.2 Fault localization with dimensionality reduction

In our investigation into the efficacy of various ML regression models such as Bagging, RF, RT, ET, KNR, SVR, and BRR for identifying faults in transmission lines, we observed a notable decline in localization performance when employing dimensionality reduction techniques such as Principal Component Analysis (PCA), Kernel PCA, and Linear Discriminant Analysis (LDA).

Across all six transmission lines tested, the mean absolute percentage error (MAPE) for fault localization exhibited a significant increase with the application of dimensionality reduction techniques. For instance, the MAPE for bagging without dimensionality reduction (referred to as NODR) ranged from 1.2 to 2.5. However, with PCA (36 components), the MAPE ranged between 3 and 33, and with PCA (108 components), it ranged from 4 to 31.6. Similar trends were observed for other regression models such as RF, RT, ET, KNR, SVR, and BRR. For example, RF exhibited a MAPE ranging from 1 to 2.1 without dimensionality reduction, whereas with PCA (36 components), it ranged from 3.6 to 33, indicating a substantial increase in error rates. This pattern persisted across different dimensionality reduction techniques and the various numbers of components tested. Even the top-performing model, BRR, experienced a significant degradation in performance with dimensionality reduction. While BRR achieved MAPE values ranging from 0.008 to 0.04 without dimensionality reduction, these values increased drastically with PCA and kernel PCA, reaching up to 39.4 for PCA (108 components).

Despite testing a wide range of component numbers for PCA and kernel PCA, the observed degradation in performance remained consistent. Therefore, we only present results for PCA (36 and 108 components) and kernel PCA (36 and 108 components) in this study. In summary, our findings suggest that the utilization of dimensionality reduction techniques adversely impacts the fault localization performance of machine learning regression models. Detailed line-wise results of fault localization performance with and without dimensionality reduction are tabulated and presented in this subsection.

Table 4.8 presents the fault localization performance for Line 4-5 both without dimensionality reduction (NODR) and with dimensionality reduction techniques including PCA36, PCA108, kPCA36, and kPCA108. The corresponding visual representation of these results can be found in Figure 4.7, utilizing a column chart format. The findings from the table and corresponding figure indicate a consistent trend: the mean absolute percentage error (MAPE) is minimal when the models operate without dimensionality reduction (NODR). However, upon employing dimensionality reduction techniques, there is a significant increase in MAPE across all models.

Table 4.8 Line 4-5 fault localization performance of ML models with and without dimensionality reduction techniques.

Regressor	NODR	KPCA 36	KPCA 108	PCA 36	PCA 108
Bagging	2.145	30.71	29.38	29.40	28.84
RF	1.442	30.21	29.58	28.86	28.64
RT	2.316	31.39	31.03	30.49	32.35
KNR	9.996	28.67	28.99	26.42	26.24
ET	2.610	35.66	32.59	32.92	32.31
BRR	0.0154	32.58	43.15	27.98	38.32
SVR	8.236	25.20	25.19	25.42	25.42

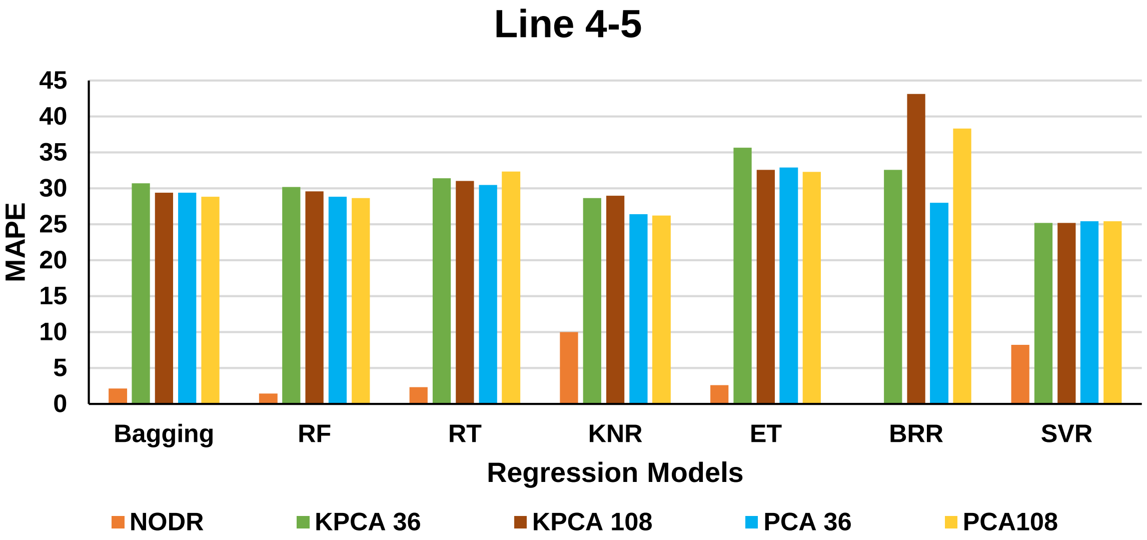


Figure 4.7 Line 4-5 fault localization performance comparison of ML models with and without dimensionality reduction techniques.

Even BRR for line 4-5, which initially demonstrated superior performance in fault localization with a notably low MAPE of 0.015, experienced a substantial rise in error rates when utilizing dimensionality reduction. For instance, MAPE surged to 32.5 for kPCA with 36 components, 32.58 with 108 components, 28 for PCA with 36 components, and 38.3 with 108 components. This highlights the detrimental impact of dimensionality reduction techniques on fault localization performance, even for the most proficient model.

Table 4.9 Line 4-6 fault localization performance of ML models with and without dimensionality reduction techniques.

Regressor	NODR	KPCA 36	KPCA 108	PCA 36	PCA 108
Bagging	1.845	32.16	30.77	28.41	28.09
RF	1.555	32.09	31.07	28.85	28.08
RT	1.691	34.49	34.51	31.52	29.05
KNR	10.483	28.44	28.26	26.74	26.86
ET	1.800	33.40	33.65	31.51	33.27
BRR	0.015	29.47	34.72	29.56	31.12
SVR	7.976	25.07	25.07	24.86	24.86

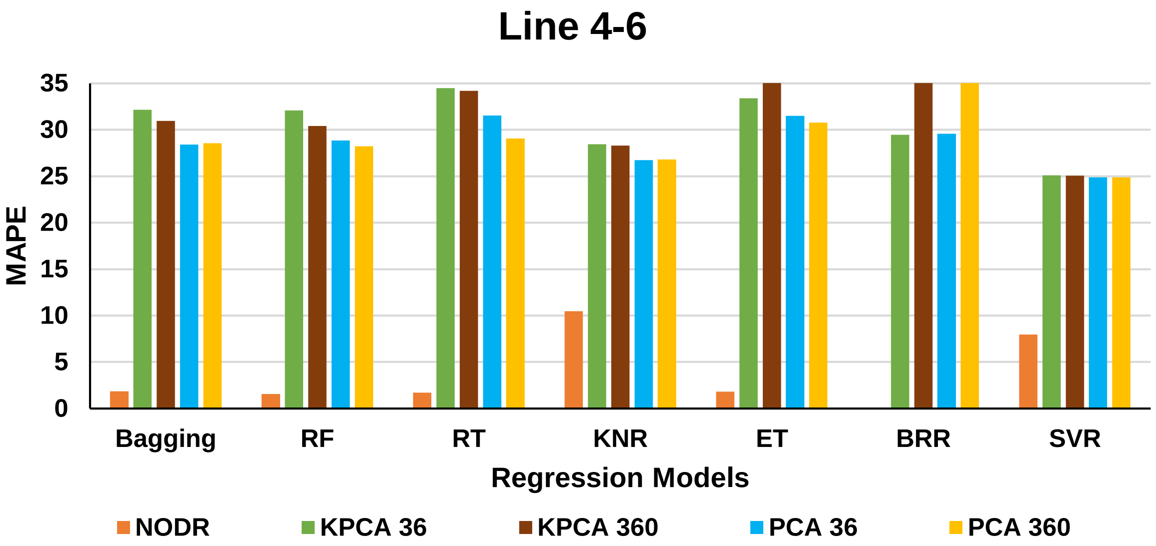


Figure 4.8 Line 4-6 fault localization performance comparison of ML models with and without dimensionality reduction techniques.

Similarly, for Line 4-6, BRR demonstrated superior performance for NODR with a notably low MAPE of 0.015; however, when utilizing dimensionality reduction techniques, MAPE surged to 29.4 for kPCA with 36 components, 34.7 with 108 components, 29.5 for PCA with 36 components, and 31.12 with 108 components, as tabulated in Table 4.9 and represented in Figure 4.8. A similar pattern can be seen for the other four lines, i.e., Line 7-

5, Line 7-8, Line 8-9, and Line 9-6, as tabulated in Tables 4.10, 4.11, 4.12, and 4.13 and represented in Figures 4.9, 4.10, 4.11, and 4.12, respectively.

Table 4.10 Line 7-5 fault localization performance of ML models with and without dimensionality reduction techniques.

Regressor	NODR	KPCA 36	KPCA 108	PCA 36	PCA 108
Bagging	1.233	13.06	14.17	13.40	14.12
RF	1.001	13.12	13.80	13.28	14.08
RT	0.928	15.29	17.39	15.41	17.15
KNR	6.695	23.75	23.56	24.57	24.46
ET	1.304	19.05	19.52	16.44	20.77
BRR	0.030	19.08	34.29	16.25	30.23
SVR	8.327	25.02	25.04	25.10	25.11

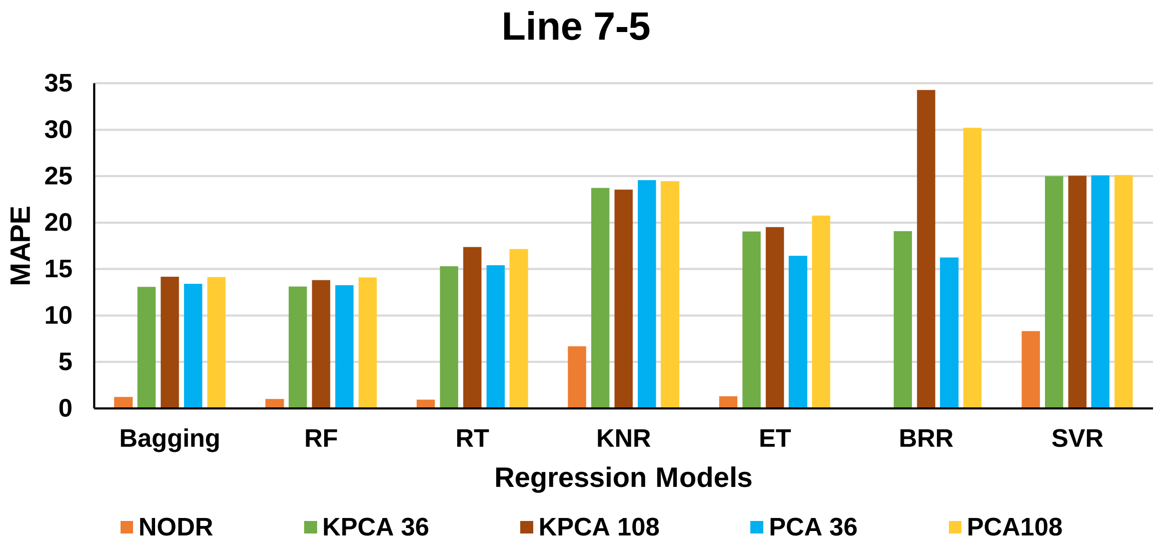


Figure 4.9 Line 7-5 fault localization performance comparison of ML models with and without dimensionality reduction techniques.

Table 4.11 Line 7-8 fault localization performance of ML models with and without dimensionality reduction techniques.

Regressor	NODR	KPCA 36	KPCA 108	PCA 36	PCA 108
Bagging	2.462	32.20	31.40	30.77	30.69
RF	2.056	32.27	30.83	30.35	29.60
RT	2.525	34.66	33.50	34.42	32.11
KNR	10.806	28.49	28.44	29.06	28.95
ET	1.955	35.03	32.32	32.50	33.62
BRR	0.008	32.96	37.43	36.61	39.41
SVR	8.702	24.68	24.67	24.74	24.74

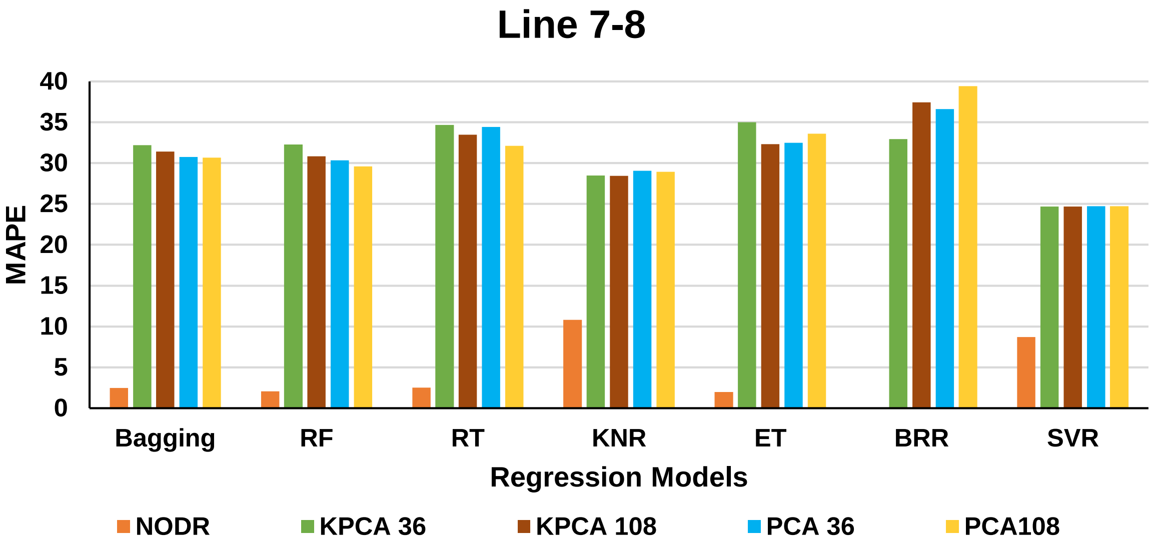


Figure 4.10 Line 7-8 fault localization performance comparison of ML models with and without dimensionality reduction technique.

Table 4.12 Line 8-9 fault localization performance of ML models with and without dimensionality reduction techniques.

Regressor	NODR	KPCA 36	KPCA 108	PCA 36	PCA 108
Bagging	2.229	32.03	31.52	32.34	31.61
RF	1.662	31.32	30.68	32.37	32.03
RT	2.420	35.12	34.33	33.80	32.87
KNR	10.784	30.34	30.44	28.81	29.06
ET	2.570	35.29	35.66	34.44	34.35
BRR	0.015	31.95	46.71	37.06	34.99
SVR	9.16	25.92	25.91	25.99	25.99

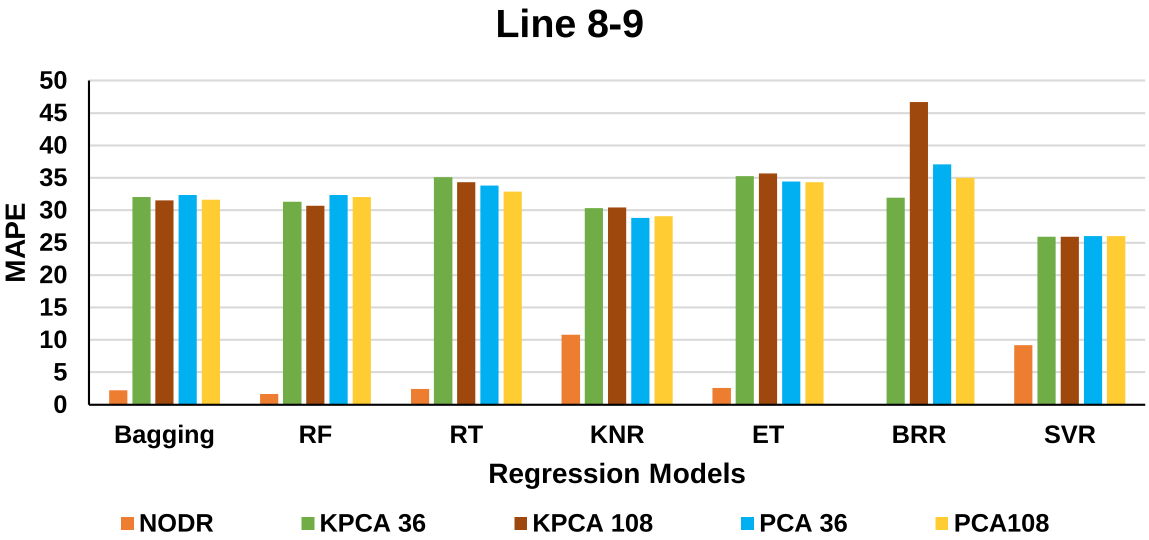


Figure 4.11 Line 8-9 fault localization performance comparison of ML models with and without dimensionality reduction techniques.

Table 4.13 Line 9-6 fault localization performance of ML models with and without dimensionality reduction techniques.

Regressor	NODR	KPCA 36	KPCA 108	PCA 36	PCA 108
Bagging	1.409	14.21	13.76	13.83	13.96
RF	1.035	13.76	13.37	13.63	13.39
RT	1.286	15.05	15.18	14.90	14.11
KNR	7.069	17.03	17.02	17.04	17.00
ET	0.974	14.81	15.12	14.32	15.64
BRR	0.041	13.93	10.50	11.80	18.98
SVR	7.648	22.84	22.84	23.05	19.05

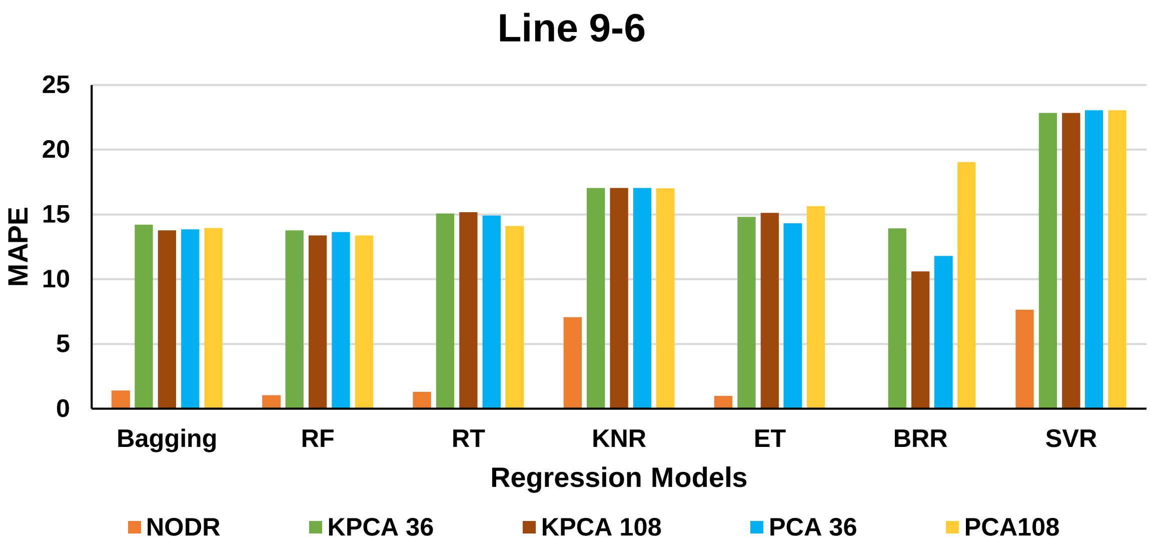


Figure 4.12 Line 9-6 fault localization performance comparison of ML models with and without dimensionality reduction techniques.

4.6 Result Analysis

The results and outcomes of this chapter demonstrate that ML models perform effectively for power system fault classification and localization when sufficient data is available. Therefore, this chapter establishes that ML-based automatic fault diagnosis of transmission networks can be achieved using voltage and current measurements of transmission line buses. The test conducted for conventional power system fault classification without dimensionality reduction revealed that the RF, ET, Bagging, and XGBoost models achieved 100% accuracy. The KNN classifier demonstrated satisfactory performance. Additionally, BRR emerged as a superior model for fault localization compared to Bagging, RF, RT, ET, and SVR regression models. However, KNN computes the average of target values in regression tasks. Consequently, its performance suffers in localization tasks. The performance of SVM was unsatisfactory in terms of both classification and localization. The inadequate performance of SVM in both classification and localization can be mainly ascribed to its extensive dependence on feature selection and transformation techniques, while multi-class classification introduces further challenges.

Upon further exploration of fault classification with dimensionality reduction, it was observed that SVM performance exhibited significant improvement. KNN performance also showed a slight enhancement with dimensionality reduction. However, the accuracy of DT and ensemble methods such as RF, ET, XGBoost, and Bagging remained almost the same with dimensionality reduction, maintaining a consistent 100% accuracy rate. Conversely, AdaBoost accuracy experienced a notable decline with dimensionality reduction. Moreover, all models' performance for fault localization severely deteriorated with dimensionality reduction. The substantial increase in mean absolute percentage error (MAPE) suggests that dimensionality reduction techniques like PCA and kPCA are unsuitable for transmission line fault localization.

4.7 Chapter Summary

This chapter extensively tested various baseline and potential models for classifying and localizing conventional transmission network faults. The findings suggest that machine learning (ML) models excel in power system fault classification and localization, given sufficient data availability. Consequently, this study affirms the feasibility of ML-based automatic fault diagnosis for transmission networks using voltage and current measurements at transmission line buses.

The tests demonstrated nearly perfect accuracy, approximately 100%, for conventional power system fault classification utilizing DT, RF, ET, Bagging, and XGBoost models. Notably, XGBoost showcased superior performance attributed to its faster execution with multithreading parallel computing capabilities, rendering it suitable for transmission line fault diagnosis without requiring data normalization for data from phasor measurement units (PMUs) or fault data loggers. Moreover, Bayesian Ridge Regression (BRR) emerged as the superior model for fault localization compared to bagging, RF, RT, ET, and SVR models. These results underscore the potential of ML techniques for enhancing fault diagnosis efficiency in power systems, paving the way for more reliable and automated fault detection and localization methodologies. Further, the models have also been explored for classification and localization performance with dimensionality reduction. The investigation reveals that SVM, MLP, and KNN performances improved with dimensionality reduction. However, ensemble methods are capable of dealing with high dimensional data effectively without requiring dimensionality reduction. Moreover, fault localization performance degrades significantly with dimensionality reduction techniques.

Chapter 5 PERFORMANCE AND ADAPTABILITY ANALYSIS OF MACHINE LEARNING MODELS FOR TRANSMISSION NETWORK FAULT CLASSIFICATION AND LOCALIZATION WITH RES INTEGRATIONS

5.1 Introduction

The integration of solar and wind energy-based electrical power plants into transmission and distribution networks has been driven by the scarcity of conventional energy resources and their adverse environmental impacts [231]. These renewable energy plants/units can vary widely in capacity, ranging from small-scale units in kilowatts to large-scale installations in megawatts. Typically, large-scale renewable energy generating units are integrated into the transmission level, while smaller-scale units are connected to distribution networks. Over the past decades, numerous large-scale renewable energy sources (RES) have been integrated into grids worldwide [9], with many more expected to follow. This integration reduces reliance on conventional coal-based power plants, thereby reducing greenhouse gas emissions [232]–[234]. While RES-based power plants offer environmental benefits, their integration presents challenges to grid operations. The complexity of power systems increases with RES integration, heightening system vulnerability [234]. Ensuring stability and effective power management in RES integrated power networks, balancing supply and demand, and setting protection devices appropriately becomes particularly challenging. Although power flow in transmission lines is typically unidirectional, integrating RES at different locations introduces tapping points and enables bi-directional power flow in the lines [21]. Additionally, RES can feed fault currents during faults, which results in increased fault currents in lines [18]. Consequently, the signature of a fault occurring at a location will change with the inclusion of a new RES unit, even if the fault attributes remain the same.

Therefore, this chapter aims to analyze a power system in which the system topology has changed due to the integration of RES of varying sizes, and fault data for the altered system is unavailable for ML-based transmission line fault classification and localization. Thus, an investigation has been conducted to test whether ML models trained with conventional power system fault data will continue to accurately identify fault types and estimate fault locations after RES integrations. Further, an adaptability analysis of the considered ML models for power system fault diagnosis post RES integrations has also been conducted. In adaptability analysis ML models are evaluated for their learning ability by incrementally training them with a growing amount of fault data from RES integrated systems over time, in conjunction with conventional power system fault data. This analysis will assist in selecting appropriate ML models based on their learning capabilities for the changed system topology under the practical condition of minimal fault data availability over time. The main objectives and contributions of the research work presented in this chapter are as follows:

1) To analyze and compare the impact of new RES integration on transmission line fault classification and localization performance of various ML models.
2) Analyzing and comparing the adaptability performance of the ML models for classification and localization of faults after RES integration, considering real-world power system scenarios of fault data availability over time.
3) To propose ML models capable of rapid learning with minimal samples of new fault data post RES integrations for diagnosing faults in transmission lines by analyzing the learning trends of the studied models.

To achieve these objectives, fault data has been generated using the standard IEEE 9 Bus transmission network for different power system topologies, i.e., conventional power systems and RES integrated systems of various sizes as six different integration cases. Following on, the proposed analyses have been performed under two practical power system scenarios. Firstly, the ML-based fault classification and localization performance have been analyzed for new RES integration when fault data for RES integrated system is unavailable for training, and the performance has been compared with that obtained for conventional power systems presented in the previous chapter. The analysis has been termed impact analysis of new RES integration on ML models. Secondly, the adaptability analysis of ML models to the gradual availability of fault data over time post RES integration has been analyzed.

5.2 Background

Limited work has been reported in the literature for ML-based power system fault diagnosis with RES integration. A wind energy-based RES integrated IEEE 9 Bus system fault detection and classification using NN and SVM is presented in [64]. Researchers have presented Support Vector Data Description-based faulty region identification for distributed energy resources (DER) integrated network [65] and SLG fault detection for varying levels of DER penetrations in distribution networks [16]. A fault classification study has also been conducted for distribution networks with two distributed generation (DG) units using a convolutional neural network (CNN) [17]. A faulty line identification approach for a RES integrated system utilizing a deep learning framework with CNN layers for feature extraction from voltage and current waveforms has been proposed in [63]. A fault classification and localization in a five bus test system of 11kV medium voltage distribution line with two DGs using linear SVM, KNN, and Bagging for fault classification and GPR, RT, SVR, and linear regression for fault localization have been presented in [66]. A study on the identification of a SLG fault phase and faulty segment using ANN, SVM, Bagging, and AdaBoost utilizing the IEEE 13 Bus test system, incorporating two DGs of 1800kVA and 2600kVA has been presented in [67].

Based on the literature survey presented in this dissertation, it can be deduced that ML-based fault classifiers perform satisfactorily for conventional power system networks [6]. However, the performance of these classification schemes has not been sufficiently tested with RES integrated power systems. Furthermore, studies have yet to explore how the integration of a new RES into a conventional power network impacts the performance of these ML classifiers. Therefore, given the growing trend of RES integration into transmission networks [231], there is a pressing need for performance analysis of potential ML models post RES integration before their implementation in actual power systems. Additionally, it is noted that most fault localization schemes are aimed at identifying faulty lines or sections of transmission and distribution networks using classification approaches [44], [62]. However, pinpointing the exact fault location on a line using ML regressors holds greater value for expedited maintenance of transmission networks. Notably, there is no existing literature on identifying the precise fault location on transmission lines in RES integrated transmission systems using either ML-based classification or regression techniques. Moreover, the wide variability in power generation from RES plants due to weather conditions, particularly temperature and irradiance fluctuations throughout the day and year, directly impacts the power output from solar PV-based RES [78]. Consequently, the power fed into the grid fluctuates, leading to variations in fault current levels [13]. Therefore, it is imperative to consider temperature and irradiance variations while analyzing solar PV-based RES integrated transmission networks [79]. Nevertheless, existing literature on ML-based power system fault diagnosis overlooks these critical issues.

The integration of RES into an existing transmission network can significantly change the power system topology and fault characteristics, depending on the size of the added RES unit [21]. Large fluctuations in power generation from RES can result in substantial deviations of fault currents from their normal values [13], potentially leading to higher misclassification rates of ML models and significant errors in fault location estimation. However, to date, no reported study in the literature examines the impact of new RES integration on the performance of ML models for transmission line fault diagnosis. Therefore, further research is needed to investigate how ML models utilized for fault diagnosis in power systems are affected by the integration of new RES of diverse sizes. Additionally, the literature review suggests that only limited studies are available on ML-based fault diagnosis of RES integrated power systems. Furthermore, these studies assumed that RES had been integrated long ago and that sufficient fault data representing diverse fault variations was available for training ML models. However, this is not true when a new RES is integrated into a real-world power network. Transmission line faults statistically occur infrequently in real-world power systems [77]. Gathering diverse fault data, including various fault locations, types, and attributes, typically requires several years. Therefore, there is a very low probability of different types of faults occurring with significant variations in fault attributes shortly after RES integration. Hence, it is crucial to analyze the performance of ML models while considering these practical issues to ensure uninterrupted power transfer through lines, meeting current and future needs of power system protection and maintenance [21].

5.3 Research Methodology

To assess the fault classification and localization performance of potential ML models on new RES integration, this chapter proposes to analyze two practical scenarios arising from new RES integration: 1) when no-fault data is available for the changed system, and 2) when the changed system's fault data is available over time. The proposed impact testing of new RES integration on ML models and adaptability testing of potential ML models to fault data availability post RES integration have been carried out by optimally integrating RES of different sizes into the 'IEEE 9-Bus System'. The schematic diagram of the RES integrated IEEE 9 Bus system at optimal bus locations, i.e., Bus-7 and Bus-5, has been shown in Figure 5.1.

The main objective of the impact analysis is to evaluate the performance of previously trained ML models post RES integration. This analysis investigates whether ML models trained with conventional power system fault data will continue to accurately identify fault types and estimate fault locations after RES integrations. Solar PV plants of three different MW ratings have been integrated into the IEEE 9 Bus system at two optimal bus locations. Similar to conventional power systems, fault data for all six RES integration cases has been separately generated using MATLAB Simulink, considering temperature and irradiance effects. The performance of previously trained ML models is then assessed on fault samples from RES integrated systems. To present the impact of RES integration on ML models' performance, the results of different sizes of RES integration have been shown in comparison to the results obtained in the previous chapter for conventional power system.

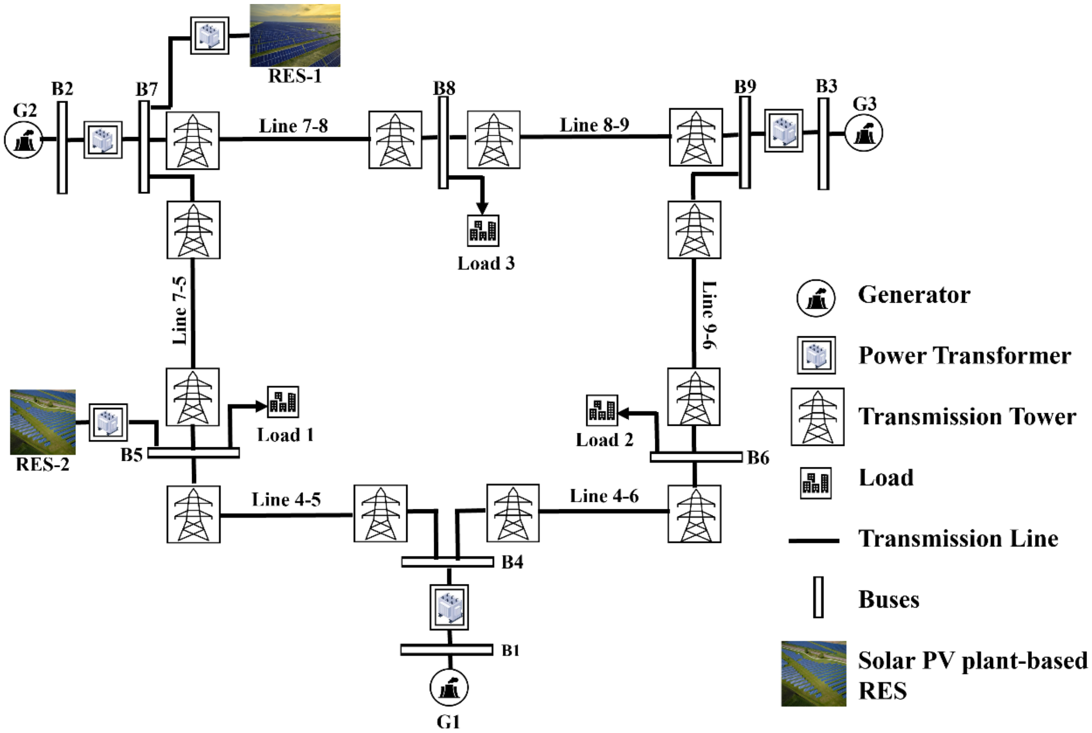


Figure 5.1 RES integrated IEEE 9 Bus system schematic diagram.

In adaptability testing, ML models are evaluated for their learning ability by incrementally training them with a growing amount of fault data from RES integrated systems over time, in conjunction with conventional power system fault data. The primary objective of adaptability analysis is to understand the learning trend of the studied ML models. Thus, in this analysis, training each ML model commences by incorporating 0.25% of new fault data samples from the RES integrated fault database, with this percentage subsequently increasing to 0.5%, 1%, 2%, and so forth until the model's performance matches that observed for the conventional power system.

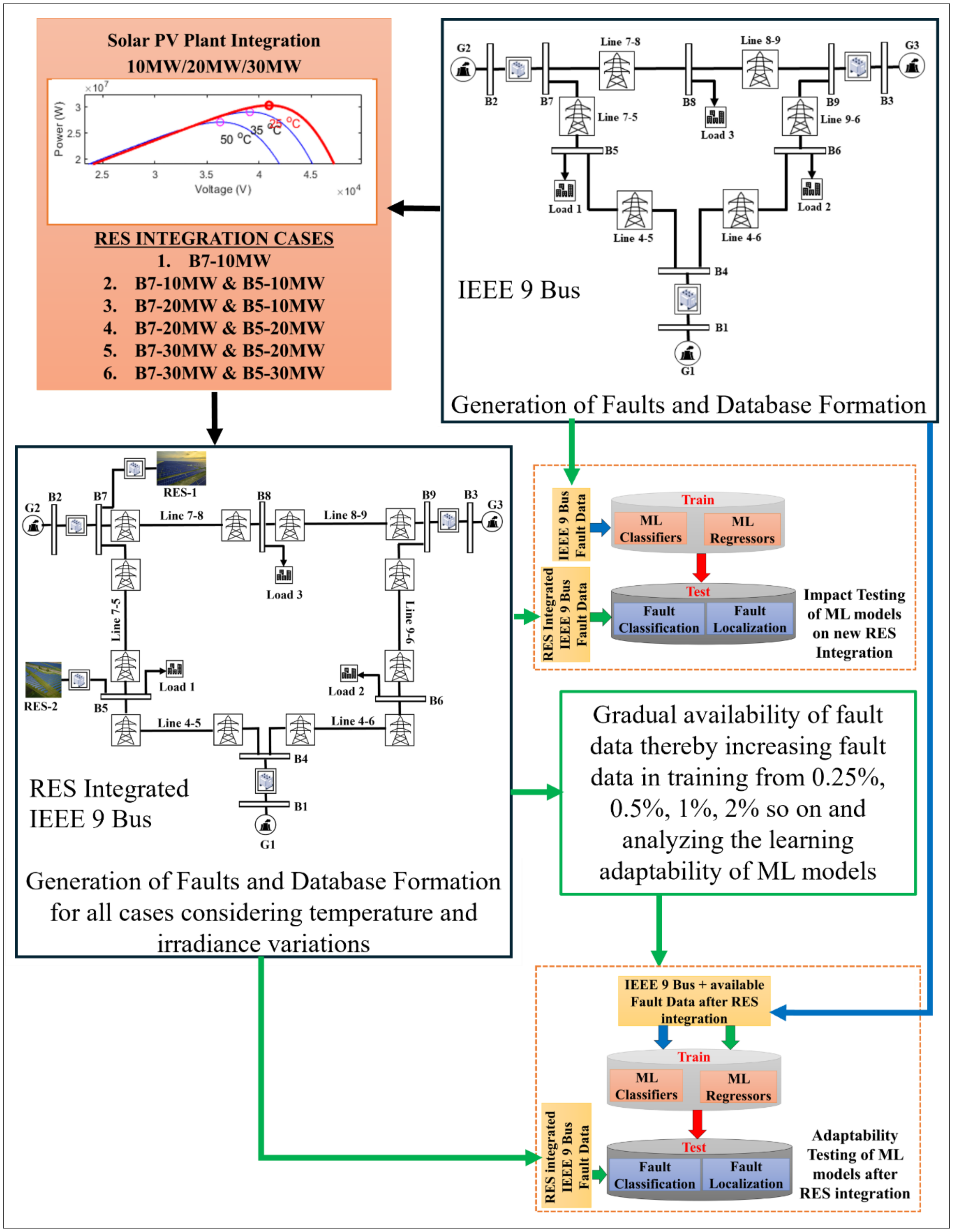

Solar PV Plant Integration
10MW/20MW/30MW
Power (W)
×10^7
3
2.5
2
2.5
3
3.5
4
4.5
Voltage (V)
×10^4
50 °C
35 °C
25 °C
RES INTEGRATION CASES
1. B7-10MW
2. B7-10MW & B5-10MW
3. B7-20MW & B5-10MW
4. B7-20MW & B5-20MW
5. B7-30MW & B5-20MW
6. B7-30MW & B5-30MW
G2
B2
B7
Line 7-8
B8
Line 8-9
B9
B3
G3
Line 7-5
Load 3
Line 9-6
B5
B6
Load 1
Load 2
Line 4-5
Line 4-6
B4
B1
G1
IEEE 9 Bus
Generation of Faults and Database Formation
RES-1
RES-2
RES Integrated
IEEE 9 Bus
Generation of Faults and Database Formation for all cases considering temperature and irradiance variations
IEEE 9 Bus Fault Data
Train
ML Classifiers
ML Regressors
RES Integrated IEEE 9 Bus Fault Data
Test
Fault Classification
Fault Localization
Impact Testing of ML models on new RES Integration
Gradual availability of fault data thereby increasing fault data in training from 0.25%, 0.5%, 1%, 2% so on and analyzing the learning adaptability of ML models
IEEE 9 Bus + available Fault Data after RES integration
Train
ML Classifiers
ML Regressors
RES integrated IEEE 9 Bus Fault Data
Test
Fault Classification
Fault Localization
Adaptability Testing of ML models after RES integration

Figure 5.2 Schematic diagram of the methodology adopted for the proposed study.

This analysis can reveal the fastest learning classification and localization models suitable for dynamically changing transmission networks due to ongoing RES integrations. The model exhibiting the highest classification accuracy with the minimum requirement of new training data is deemed the most adaptable classifier. Similarly, the regression model demonstrating the lowest mean absolute percentage error (MAPE) with the least new training data is regarded as the most adaptable regression model for fault localization. Figure 5.2 gives a schematic diagram of the methodology adopted for the proposed impact and adaptability analysis of ML based fault classification and localization on new RES integration.

5.4 Dataset Description

To investigate the impact of RES integration on ML models' performance and investigate the adaptability of ML models to fault data availability post RES integration for power system fault classification and localization, the RES integration for six different combinations has been considered. Table 5.1 describes the training and testing dataset for impact and adaptability analysis.

The RES integration of six different sizes, i.e., 10, 20, 30, 40, 50, and 60 MW, has been considered in the presented study. The placement of these RES in the IEEE 9 Bus system is as follows: Bus 7-10MW; Bus 7-10MW and Bus 5-10MW; Bus 7-20MW and Bus 5-10MW; Bus 7-20MW and Bus 5-20MW; Bus 7-30MW and Bus 5-20MW; and lastly, Bus 7-30MW and Bus 5-30MW. Fault data has been generated considering one case or size combination at a time.

Table 5.1 Data description for training and testing of ML models.

Analysis	Train and Test Dataset (Power system fault data)
Impact Analysis on New RES Integration	***Training*** *– Fault data without RES integration* **Testing** *–Fault data with RES Integration* of various sizes at: ***Case 1 Testing*** *– 10MW at Bus 7* ***Case 2 Testing*** *– 10MW at Bus 7 and 10MW at Bus 5* ***Case 3 Testing*** *– 20MW at Bus 7 and 10MW at Bus 5* ***Case 4 Testing*** *– 20MW at Bus 7 and 20MW at Bus 5* ***Case 5 Testing*** *– 30MW at Bus 7 and 20MW at Bus 5* ***Case 6 Testing*** *– 30MW at Bus 7 and 30MW at Bus 5*
Adaptability Analysis Post RES Integration	**Case 1:** ***Training*** *–Fault data without RES integration + Available fault data after 10MW RES integrated on Bus 7* ***Testing*** *–Fault data after RES integration (10MW on Bus 7)* **Case 2:** ***Training*** *– Fault data without RES integration + Available fault data after integrating RES (10MW on Bus 7 & 10MW on Bus 5)* ***Testing*** *–Fault data after RES integration* (*10MW on Bus 7 and 10MW on Bus 5)* **Case 3:** ***Training*** *– Fault data without RES integration + Available fault data after integrating RES (20MW on Bus 7 & 10MW on Bus 5)* ***Testing*** *–Fault data after RES integration (20MW on Bus 7 and 10MW on Bus 5)* **Case 4:** ***Training*** *– Fault data without RES integration + Available fault data after integrating RES (20MW on Bus 7 & 20MW on Bus 5)* ***Testing*** *–Fault data after RES integration (20MW on Bus 7 and 20MW on Bus 5)* **Case 5:** ***Training*** *– Fault data without RES integration + Available fault data after integrating RES (30MW on Bus 7 & 20MW on Bus 5)* ***Testing*** *–Fault data after RES integration (30MW on Bus 7 and 20MW on Bus 5)* **Case 6:** ***Training*** *– Fault data without RES integration + Available fault data after integrating RES (30MW on Bus 7 & 30MW on Bus 5)* ***Testing*** *–Fault data after RES integration (30MW on Bus 7 and 30MW on Bus 5)*

The values of fault attributes at which fault data has been generated for conventional power system and RES integrated systems have been discussed in detail in Chapter 3. For the proposed impact and adaptability analysis the detailed train test split of fault data is listed in Table 5.1 giving a clear depiction of the difference between both the analysis performed.

5.5 Impact Analysis of RES Integration on ML-Based Fault Diagnosis

When the ML algorithms are analyzed for fault classification and localization subsequent to the integration of RES into conventional power system. However, the available fault data pertains solely to conventional power systems, rendering the training of ML models with RES integrated power system fault data unfeasible. This limitation is common in real-world power systems undergoing the integration of new RES worldwide. Therefore, investigating this scenario is pivotal to identifying a model capable of effective fault diagnosis post RES integration. As only conventional power system fault data is available, the ML models were trained exclusively using the fault database of the standard IEEE 9 bus system. RES integration can occur in transmission networks of varying sizes, typically lower than the maximum allowable integration level. Previous studies on optimal RES placement on the IEEE 9 Bus System [10] considered 30 MW RES integration at an individual bus, with a total maximum allowed capacity of 60 MW. Hence, in this study, RES integration commenced at 10 MW for one bus and extended up to 30 MW for two buses. A detailed description of the training and testing fault datasets is provided in Table 5.1. The Python environment has been used for training and testing of ML models. The steps followed while conducting the impact analysis for fault classification and localization have been demonstrated using Figure 5.3.

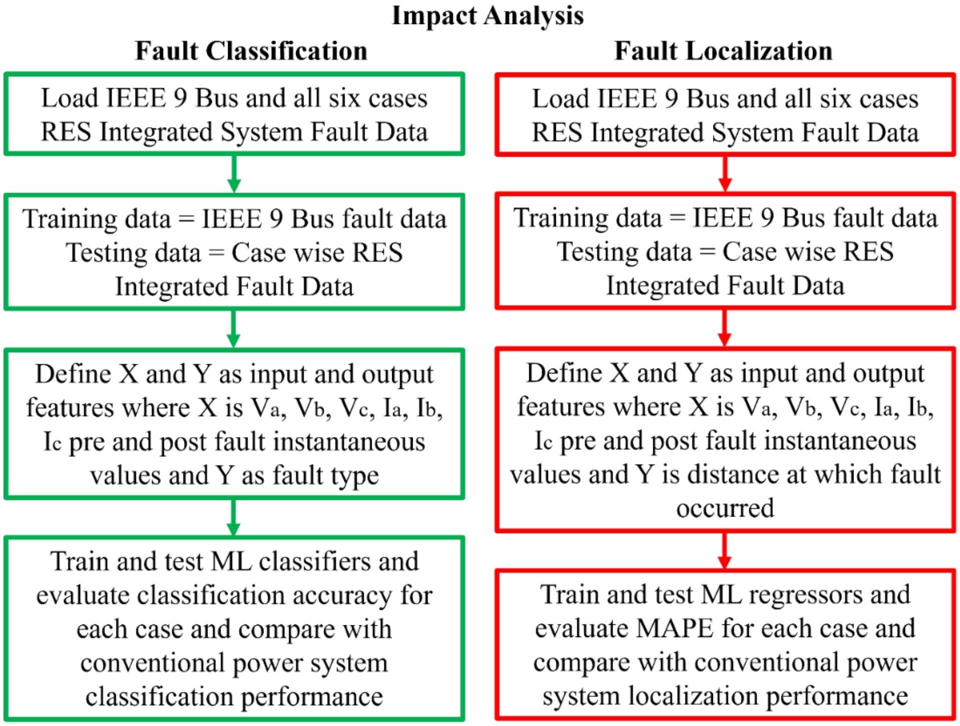

Impact Analysis
Fault Classification
Fault Localization
Load IEEE 9 Bus and all six cases RES Integrated System Fault Data
Training data = IEEE 9 Bus fault data Testing data = Case wise RES Integrated Fault Data
Define X and Y as input and output features where X is Va, Vb, Vc, Ia, Ib, Ic pre and post fault instantaneous values and Y as fault type
Train and test ML classifiers and evaluate classification accuracy for each case and compare with conventional power system classification performance
Load IEEE 9 Bus and all six cases RES Integrated System Fault Data
Training data = IEEE 9 Bus fault data Testing data = Case wise RES Integrated Fault Data
Define X and Y as input and output features where X is Va, Vb, Vc, Ia, Ib, Ic pre and post fault instantaneous values and Y is distance at which fault occurred
Train and test ML regressors and evaluate MAPE for each case and compare with conventional power system localization performance

Figure 5.3 Impact analysis procedural workflow for fault classification and localization.

The model's hyperparameters remain the same as in the previous chapter, as we are testing pre-trained models for impact and adaptability analysis. To illustrate the impact of any size RES integration on ML models' fault diagnosis performance, the results have been presented in comparison to ML models' performance for conventional power system fault classification and localization obtained in the previous chapter.

5.5.1 Fault Classification

When the ML classifiers were investigated for the impact of the new RES integration on their fault classification performance, it was found that performance degraded severely. The testing accuracies obtained for fault classification are listed in Table 5.2 and illustrated in Figure 5.4.

Table 5.2 Impact analysis of the new RES integration on the performance of ML classifiers.

ML Classifiers	CPS	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
SVM	81.47	74.12	66.88	66.33	64.06	63.78	63.63
DT	99.87	50.66	43.72	31.1	29.26	29.11	27.86
ET	100.00	51.39	48.01	46.3	46.59	44.17	42.66
RF	100.00	56.84	54.28	54.26	54.11	53.98	53.57
AdaBoost	86.74	36.36	36.31	36.17	36.22	36.29	36.23
LR	94.03	40.88	40.88	40.91	40.94	40.91	40.93
Ridge	96.02	27.27	27.29	27.24	27.21	27.19	27.17
GNB	79.07	9.09	9.19	9.14	9.18	9.1	9.09
KNN	98.86	90.67	80.87	80.68	80.68	80.53	80.39
Bagging	100.00	56.01	50.33	48.15	46.59	45.45	44.88
XGBoost	100.00	41.26	37.11	36.8	36.79	36.36	36.36
MLP	85.95	72.98	54.26	37.5	37.21	37.07	36.93

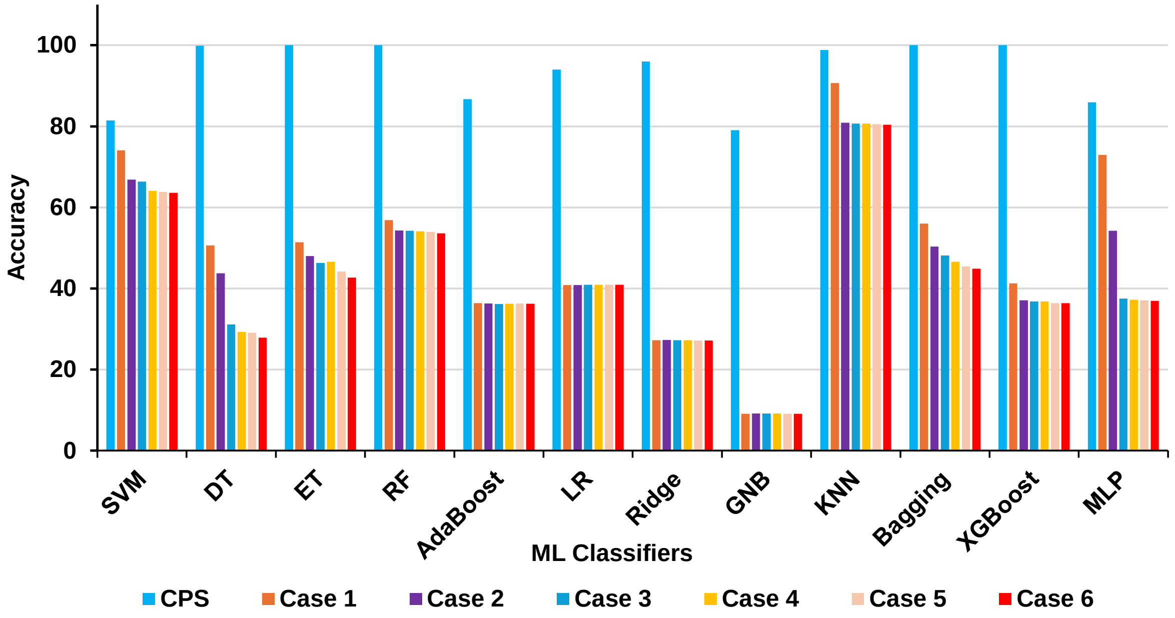


Figure 5.4 Fault classification performance of ML models on new res integration.

It was observed that when ML algorithms are trained on conventional power system fault data and subsequently tested for RES integrated power system faults, the classification accuracy degrades significantly compared to conventional power system (CPS) performance. There are two possible ways to integrate a new RES into a transmission network. First, the incoming RES is added at a location (BUS) where one or more RES units are already connected. Second, the incoming RES is added at a new location (BUS) with no previously connected RES units. If the incoming RES is added at a new location (BUS), the total number of RES-integrated buses will increase. Therefore, the total number of transmission lines directly connected to the RES integrating BUS will also increase, resulting in degraded fault diagnosis performance of the pre-trained ML models. As depicted in Figure 5.4, the classification accuracy of all ML models exhibits a decreasing pattern as the level of RES integration increases. Further, there is a substantial decrease in accuracy from Case 1 (10MW at Bus 7) to Case 2 (10MW each at Bus 7 and Bus 5) as the number of lines directly connected to the RES integrating bus has increased. However, with further growth in the level of RES integrations, i.e., from Case 2 onwards, the number of lines directly connected to RES integration points does not change, and the degradation in accuracy becomes negligible for specific models like KNN, XGBoost, and RF. In contrast, it is minimal for others, such as ET, Bagging, and SVM. Therefore, the classification accuracy for Case 1 (only one RES) is higher than that of Case 2 (two RES) and the subsequent cases. The fault current level and distribution change significantly with each new RES integration. The F1-score for all fault types was poor; hence, it has not been listed.

The KNN classifiers performed better than all the other classifiers tested in the study. Consequently, the KNN model is deemed suitable for unknown scenarios as it can discern fault patterns and extend the acquired knowledge of fault patterns from conventional power systems to untrained scenarios, such as integrating RES of unknown size. The KNN model operates based on the similarity index, thus exhibiting better performance for the altered system than other investigated models [235]. The testing accuracy of models such as DT, ET, and Bagging exhibited a decreasing trend as the fault current level increased with the size of the RES. Conversely, AdaBoost, Logistic Regression (LR), Ridge, and Gaussian NB (GNB) consistently displayed poor classification accuracy for RES integration cases, rendering them unsuitable for ML-based fault diagnosis following new RES integrations. The classification accuracy of RF, KNN, and XGBoost declines with each new RES integration at a different bus. SVM outperformed other classifiers except for KNN in this scenario, indicating its strong generalization ability. However, its performance was inconsistent for conventional power system, suggesting that it should be used with feature engineering for power system fault diagnosis.

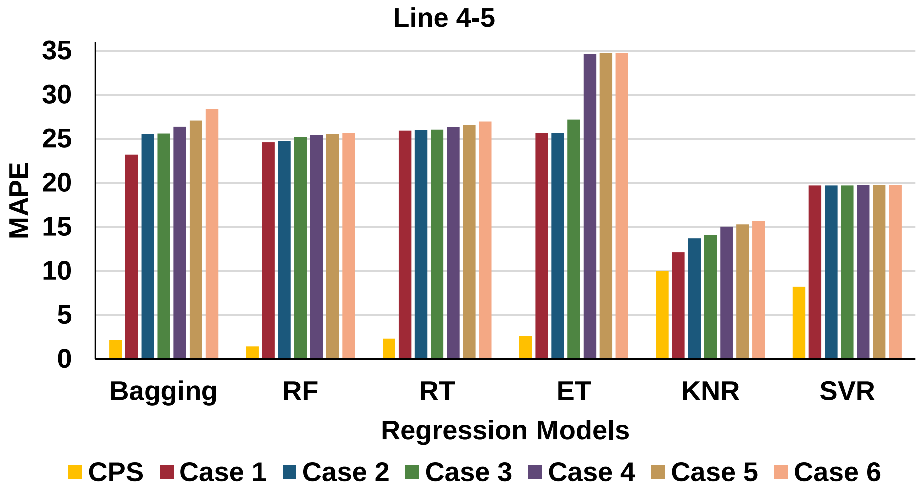
5.5.2 Fault Localization

While analyzing the impact of RES integration on ML regressor models for fault localization performance, four out of six transmission lines were assessed, excluding those not directly connected to RES. Consequently, lines directly linked to RES integrated buses, namely Line 4-5, Line 7-5, and Line 7-8, were chosen for localization analysis. Assuming a similar effect on the other three lines not directly connected to RES, only Line 8-9 was included in the study. Before localization, identification of the faulty line was conducted to determine the line where the fault occurred. Table 5.3 enlists the MAPE of the four lines under study. The results have been presented in comparison to conventional power system fault localization performances. The plots depicted in Figure 5.5 present the column plot of MAPE of ML regression models for fault location estimation on the analyzed lines, illustrating the impact of RES integration.

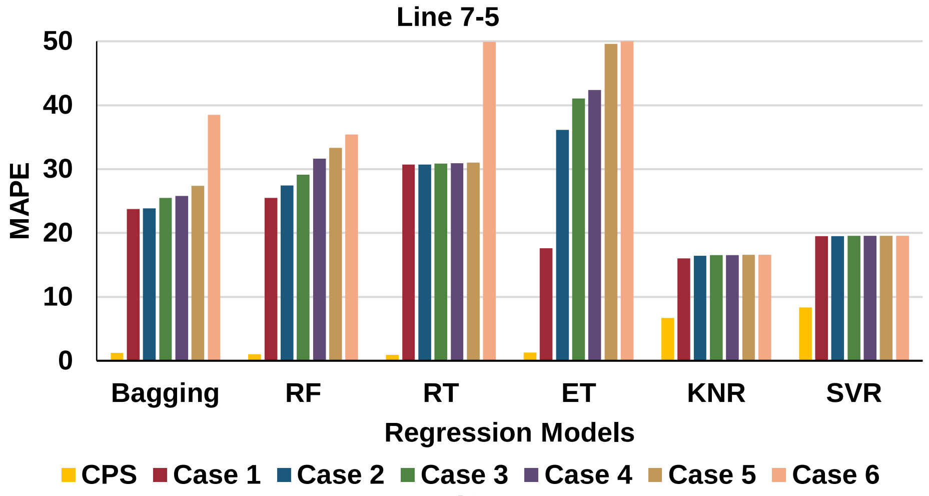
Table 5.3 Location estimation MAPE of regression models for impact analysis.

Line	Models	Bagging	RF	RT	ET	KNR	SVR
Line 4-5	**CPS**	2.145	1.442	2.316	2.610	9.996	8.236
	Case 1	23.20	24.61	25.96	25.68	12.13	19.71
	Case 2	25.59	24.78	26.02	25.68	13.71	19.73
	Case 3	25.63	25.23	26.07	27.21	14.10	19.73
	Case 4	26.39	25.43	26.36	34.65	15.04	19.74
	Case 5	27.07	25.54	26.59	34.75	15.29	19.75
	Case 6	28.37	25.68	26.98	34.77	15.65	19.75
Line 7-5	**CPS**	1.233	1.001	0.928	1.304	6.695	8.327
	Case 1	23.75	25.47	30.73	17.61	16.04	19.52
	Case 2	23.84	27.41	30.73	36.15	16.44	19.52
	Case 3	25.51	29.12	30.85	41.07	16.53	19.53
	Case 4	25.78	31.65	30.89	42.38	16.55	19.55
	Case 5	27.36	33.32	31.02	49.60	16.56	19.56
	Case 6	38.47	35.41	49.88	50.00	16.59	19.56
Line 7-8	**CPS**	2.462	2.056	2.525	1.955	10.806	8.702
	Case 1	15.22	16.87	21.87	24.94	13.69	19.67
	Case 2	28.40	26.03	26.19	37.15	17.50	19.68
	Case 3	30.14	27.03	27.15	42.67	18.07	19.68
	Case 4	31.48	27.41	27.32	44.37	19.19	19.69
	Case 5	36.95	28.15	27.38	44.77	19.82	19.69
	Case 6	39.42	32.90	28.06	45.73	20.17	19.70

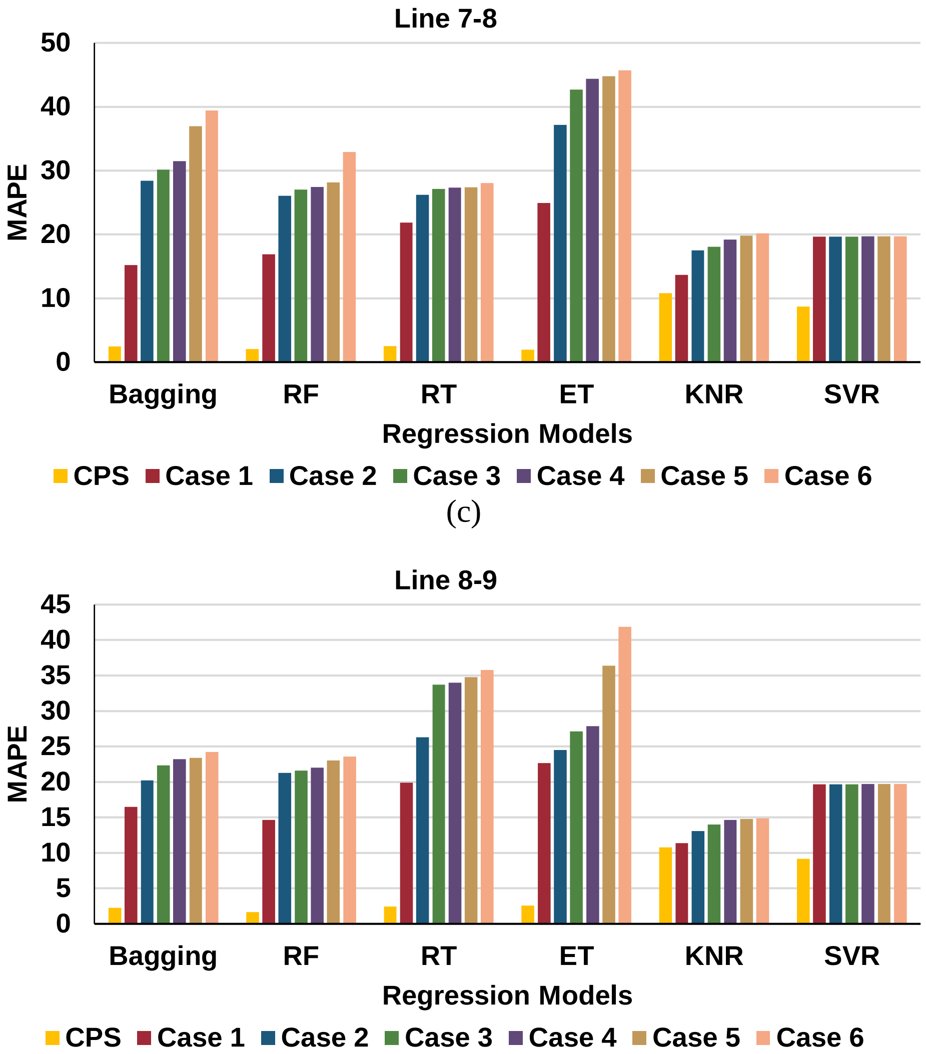
Line 8-9	**CPS**	2.228	1.662	2.420	0.974	10.784	9.160
	Case 1	16.46	14.63	19.88	22.67	11.36	19.66
	Case 2	20.20	21.27	26.30	24.48	13.09	19.67
	Case 3	22.31	21.61	33.69	27.10	14.01	19.68
	Case 4	23.19	22.01	33.97	27.84	14.63	19.69
	Case 5	23.38	23.03	34.77	36.36	14.78	19.69
	Case 6	24.23	23.56	35.79	41.87	14.85	19.70



(a)



(b)



(c)

(d)

Figure 5.5 Fault localization performance degradation of ML models on new RES integration on transmission lines: (a) Line 4-5, (b) Line 7-5, (c) Line 7-8, and (d) Line 8-9.

For most models, MAPE exhibited an increasing trend with the growth in RES integration size, as observed with Bagging, RF, and ET. Similar to classification, KNR outperformed other models in localization, displaying the lowest MAPE across all lines and cases of varying RES sizes. Additionally, SVR demonstrated superior performance compared to other regression models, with its location estimation remaining unaffected by changes in the RES integration size. However, the error percentage for both the KNR and

SVR models is still high enough, showing their inefficacy in fault localization performance. Moreover, the BRR model gave extremely poor fault location prediction, even more than the length of the line; thus, its performance has not been reported. Consequently, it cannot be relied upon in scenarios involving network topology changes, such as RES integration, or in the absence of fault data for the changed system.

The integration of RES plants at any bus alters the fault current level of the network, resulting in inferior localization performance compared to the performance obtained for the conventional power system. Further, power generation from RES, a flexible power source, will bring more fluctuating changes in fault current level due to its weather dependent power generation characteristics. However, a lesser impact is expected if another generation source of fixed power is added. Further, the learning of ML models for such cases will be better compared to fluctuating power sources.

5.6 Adaptability Analysis of ML Models Post RES Integration

The performance of ML algorithms for fault classification and localization, when fault data for the RES integrated system was unavailable was discussed in the previous section. Therefore, in this section, the aim is to analyze the performance of ML models using minimal fault data collected from RES integrated systems of varying sizes. ML algorithms necessitate sufficient data covering all attribute variabilities to learn effectively. Nevertheless, the availability of post RES integration changed system data for training purposes is limited. Consequently, training was conducted using both conventional power system fault data and available fault samples from the RES integrated system.

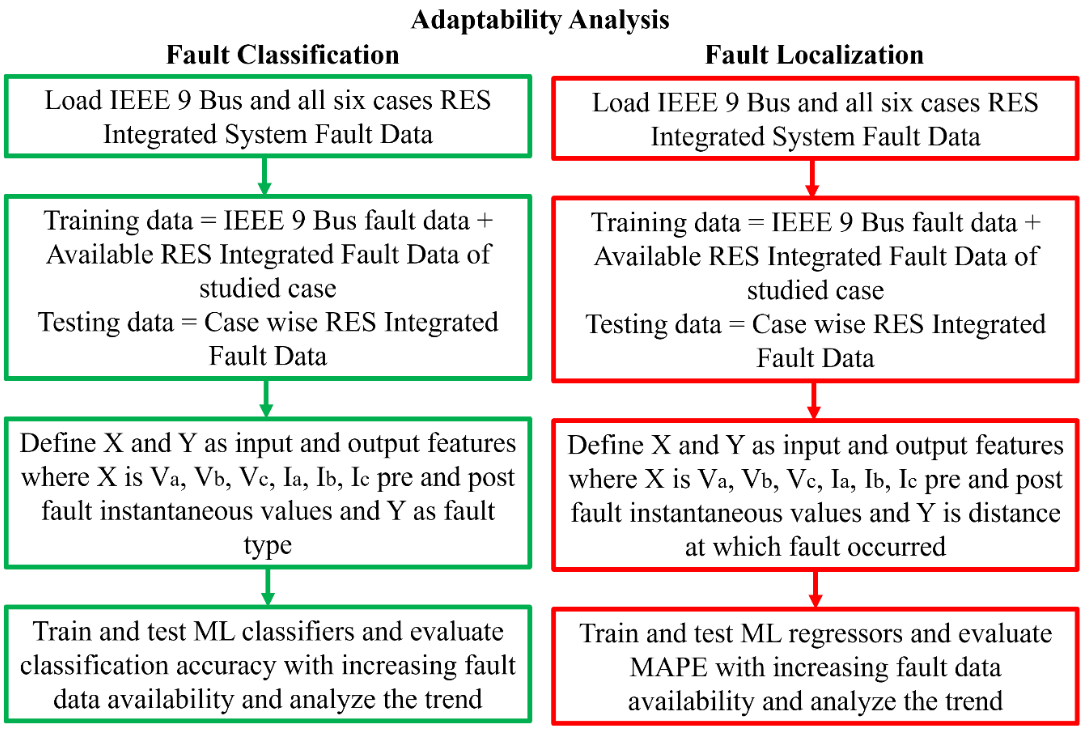

Adaptability Analysis
Fault Classification
Fault Localization
Load IEEE 9 Bus and all six cases RES Integrated System Fault Data
Training data = IEEE 9 Bus fault data + Available RES Integrated Fault Data of studied case
Testing data = Case wise RES Integrated Fault Data
Define X and Y as input and output features where X is Va, Vb, Vc, Ia, Ib, Ic pre and post fault instantaneous values and Y as fault type
Train and test ML classifiers and evaluate classification accuracy with increasing fault data availability and analyze the trend
Load IEEE 9 Bus and all six cases RES Integrated System Fault Data
Training data = IEEE 9 Bus fault data + Available RES Integrated Fault Data of studied case
Testing data = Case wise RES Integrated Fault Data
Define X and Y as input and output features where X is Va, Vb, Vc, Ia, Ib, Ic pre and post fault instantaneous values and Y is distance at which fault occurred
Train and test ML regressors and evaluate MAPE with increasing fault data availability and analyze the trend

Figure 5.6 Adaptability analysis procedural workflow for fault classification and localization.

While the training set remained constant for all test cases in impact analysis, in adaptability analysis, each training set contained available fault data samples from RES integrated systems of the same size being tested. Consequently, ML models had limited samples to learn the fault patterns of the altered network topology and transfer their learning to updated system fault diagnosis. Therefore, the adaptability analysis analyzes the adaptability of models, aiding in the selection of appropriate models based on data availability and network topology. Figure 5.6 presents the workflow for performing adaptability analysis of ML models for fault classification and localization post RES integration.

5.6.1 Fault Classification

Testing the models to transfer their learning from conventional power system fault diagnosis to RES integrated system fault diagnosis, using only 5% of RES integrated system fault data for training, revealed that many models exhibited performance matching to that they exhibited for the conventional power system. ET, Bagging, RF, and XGBoost demonstrated fault classification accuracies comparable to those observed in the previous chapter, as listed in Table 5.4 and depicted in Figure 5.7. This underscores their strong learning ability, as they can adapt to network changes with minimal training data. In contrast, although KNN's performance improved compared to impact analysis, it failed to reach the baseline performance obtained for conventional power system due to inadequate training data, highlighting the limitations in KNN's learning capacity despite its effectiveness as a classifier with generalization abilities for changed scenarios. MLP, AdaBoost, and Gaussian NB performed poorly, like conventional power system performance, whereas LR, Ridge, and DT performed optimally. As all cases included fault samples from the RES integration case in their training, there was a notable enhancement in the classification accuracy of all well performing classifiers compared to impact analysis performance. The F1-score of the studied classifiers, when trained with 5% RES integrated fault data inclusions, closely resembled the performance obtained for conventional power system performance; hence, it is not included in Table 5.4.

Table 5.4 Testing accuracy (%) of ML classifiers for adaptability analysis with 5% RES integrated fault data.

ML Classifier	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
SVM	81.92	81.53	81.47	83.61	81.83	80.47
DT	98.93	99.46	99.11	99.46	99.29	100
ET	100	100	100	100	100	100
RF	100	100	100	100	100	100
AdaBoost	64.23	50.06	41.27	50.2	40.62	29.97
LR	92.45	91.18	94.29	91.62	90.83	90.26
Ridge	95.86	97.96	96.69	93.93	94.2	92.89
GNB	49.7	46.77	41.89	56.27	59.58	42.75
KNN	98.43	97.04	97.84	97.99	96.27	96.51
Bagging	99.64	100	100	99.64	100	100
XGBoost	100	100	100	100	100	99.64
MLP	28.93	79.52	83.22	38.64	80.11	77.3

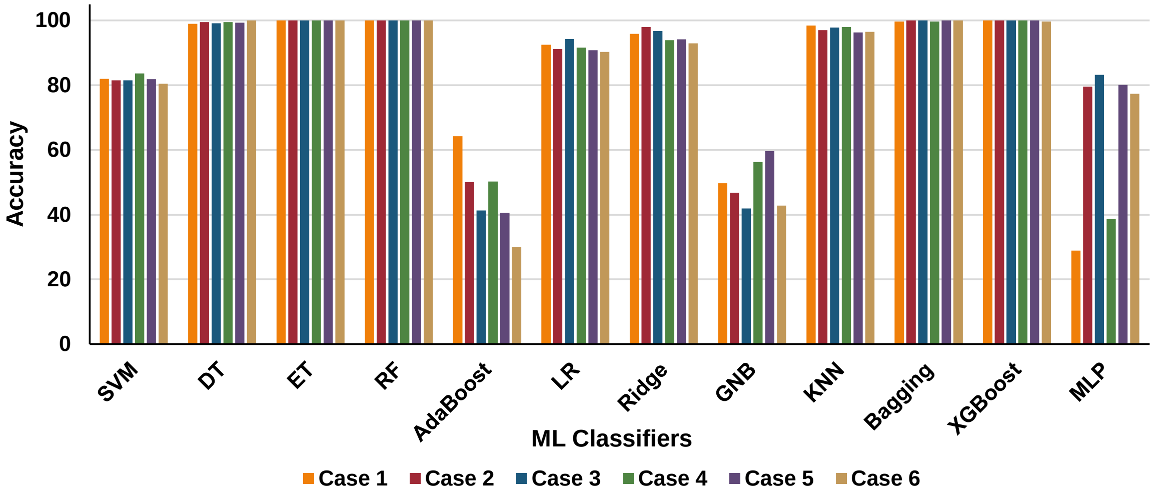


Figure 5.7 Improvement in classification accuracy of ML Models with 5% data in training.

Table 5.5 Performance of ML classifiers for adaptability analysis with 0.75% RES integrated fault data.

Models	RES Case	Testing Accuracy (%)	F1 - score				
			SLG	DLG	LL	LLLG	LLL
SVM	Case 1	75.84	0.75	0.90	0.45	0.99	0.64
	Case 6	74.37	0.75	0.89	0.40	0.97	0.61
DT	Case 1	86.82	0.89	0.83	0.81	1.00	0.94
	Case 6	83.19	0.81	0.89	0.74	0.98	0.88
RF	Case 1	97.19	0.97	0.98	0.95	1.00	1.00
	Case 6	96.63	1.00	0.94	0.94	1.00	1.00
AdaBoost	Case 1	49.82	0	0.60	0.55	0.89	0.38
	Case 6	38.40	0.44	0.53	0.32	0	0
ET	Case 1	99.70	0.99	1.00	0.99	1.00	1.00
	Case 6	95.88	0.99	0.93	0.93	1.00	1.00
LR	Case 1	87.65	0.79	0.92	0.84	1.00	0.97
	Case 6	63.51	0.42	0.80	0.55	0.98	0.64
Ridge	Case 1	77.88	0.72	0.86	0.74	0.86	0.75
	Case 6	62.54	0.53	0.63	0.69	0.75	0.56
Gaussian NB	Case 1	40.84	0	0.57	0.46	0	0
	Case 6	9	0	0	0	0	0.17
KNN	Case 1	82.05	0.91	0.88	0.67	0.94	0.54
	Case 6	81.14	0.90	0.84	0.70	0.89	0.66
Bagging	Case 1	92.74	0.94	0.91	0.90	1.00	0.94
	Case 6	89.28	0.91	0.89	0.84	0.96	0.93
XGBoost	Case 1	97.20	0.97	0.98	0.95	1.00	1.00
	Case 6	91.79	0.92	0.93	0.87	1.00	0.94
MLP	Case 1	62.25	0.49	0.72	0.61	0.87	0.32
	Case 6	36.28	0	0	0.46	0.98	0

Therefore, the F1-score for all fault types for 0.75% RES integrated fault data inclusions has been detailed in Table 5.5. The table presents the F1-score for RES integration cases 1 and 6, i.e., RES integration 10MW at bus 7 and 30MW at both bus 7 and 5. Analysis of the F1 score in the table concludes that SVM faces challenges in classifying non-ground faults, i.e., LL and LLL faults. However, SVM shows good performance for LLLG faults. Further, XGBoost, RF, and ET achieved testing accuracies surpassing 90% with a mere 0.75% inclusion of RES integrated fault data. Moreover, the performance of ET exceeded RF and XGBoost for the Case 1 RES integration.

The ML classifiers' learning ability to gradual data availability has been analyzed by varying the fault data percentage of the RES integrated system during training. The percentages of new fault data (FD) included at each step during training are 0.25%, 0.5%, 1%, 2%, 4%, and 8%. The improvement in classification testing accuracy with the increasing percentage of new FD in training has been depicted using radar plots given in Figure 5.8 for all six integration cases. A radar chart, also referred to as a spider plot or star plot, is a two-dimensional graphical method for representing multivariate data where each axis refers to variables having the same data length [236]. However, the arrangement and angles of the axes usually do not carry specific information. In Figure 5.8, the six axes represent the considered fault data inclusions (FDIs) of the RES integrated system used in training the ML models. Whereas the radial distance from the center represents the classification accuracy in percentage.

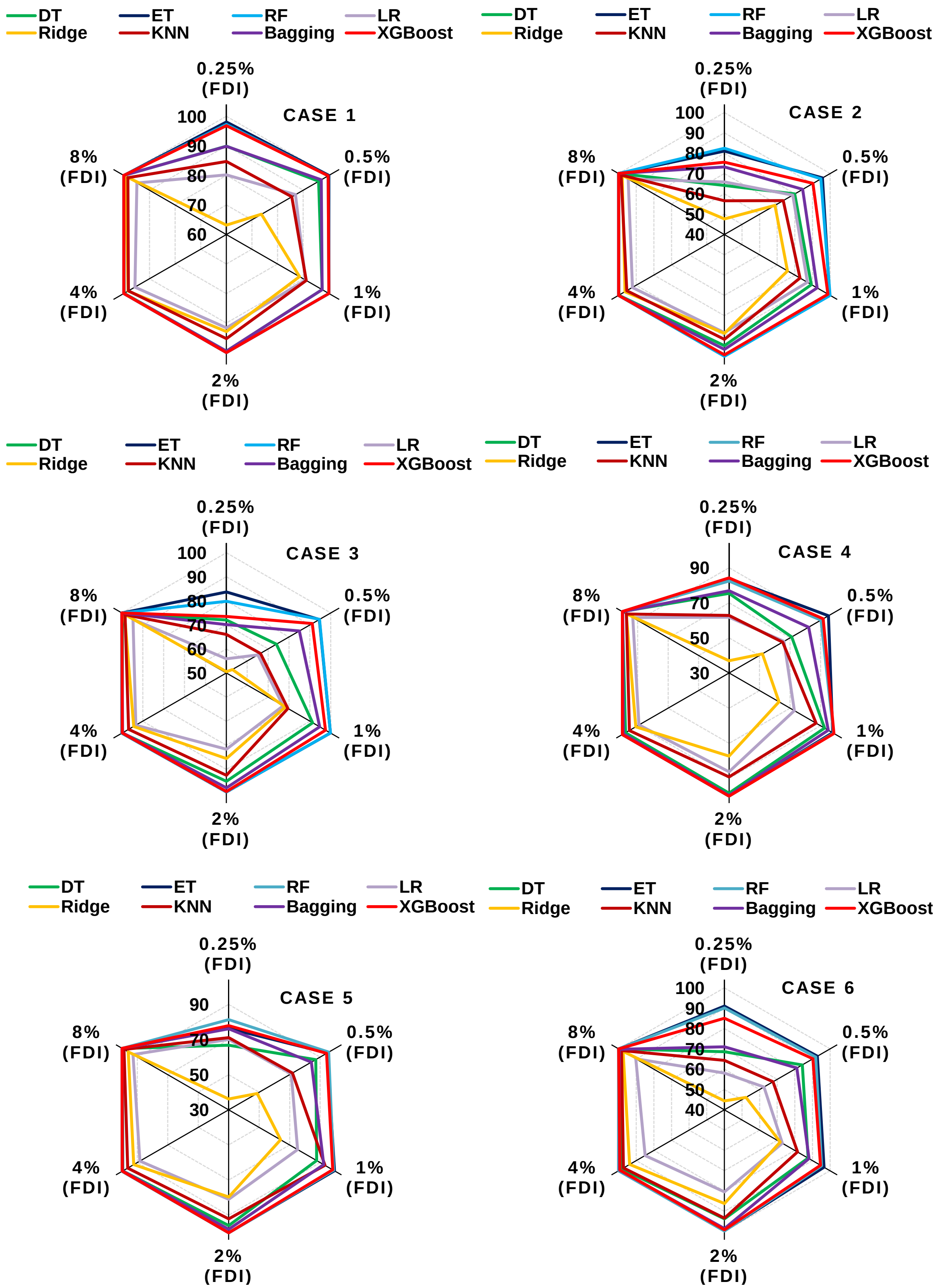


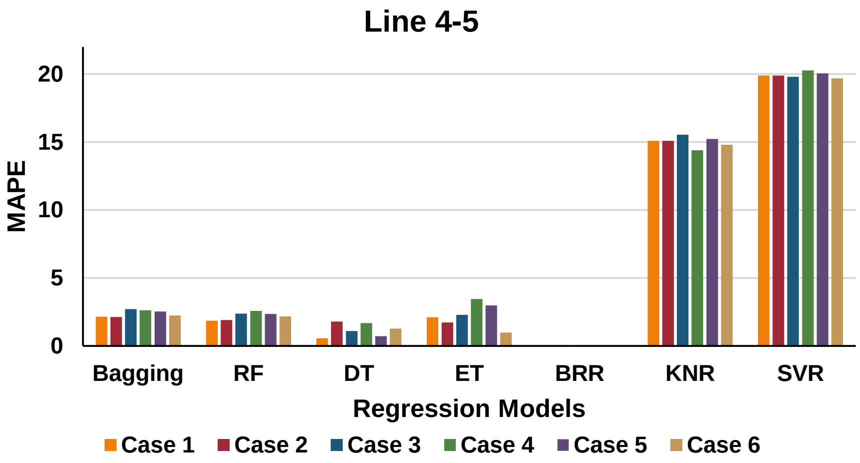
Figure 5.8 Radar charts for illustrating improving classification accuracy with % available fault data.

During the adaptability test for RES integration case 1, which entailed adding only a 10MW RES on Bus 7, RF, ET, and XGBoost achieved nearly 100% accuracy with just a 0.5% increase in new FD inclusion in their training, while DT and Bagging demonstrated similar classification performance that obtained for conventional power system with 1% increase in new FD inclusion in their training. From case 2 onwards, where RES is added at the two selected buses, it was observed that RF, ET, and XGBoost achieved nearly 100% accuracy with just a 1% inclusion of new FD in their training. Additionally, the plots indicate that the performance of other models also improves as the amount of RES integrated FD in their training data increases. However, it can be concluded that RF, ET, and XGBoost classifiers are particularly adept at quickly learning and adapting to new scenarios with minimal changed topology data.

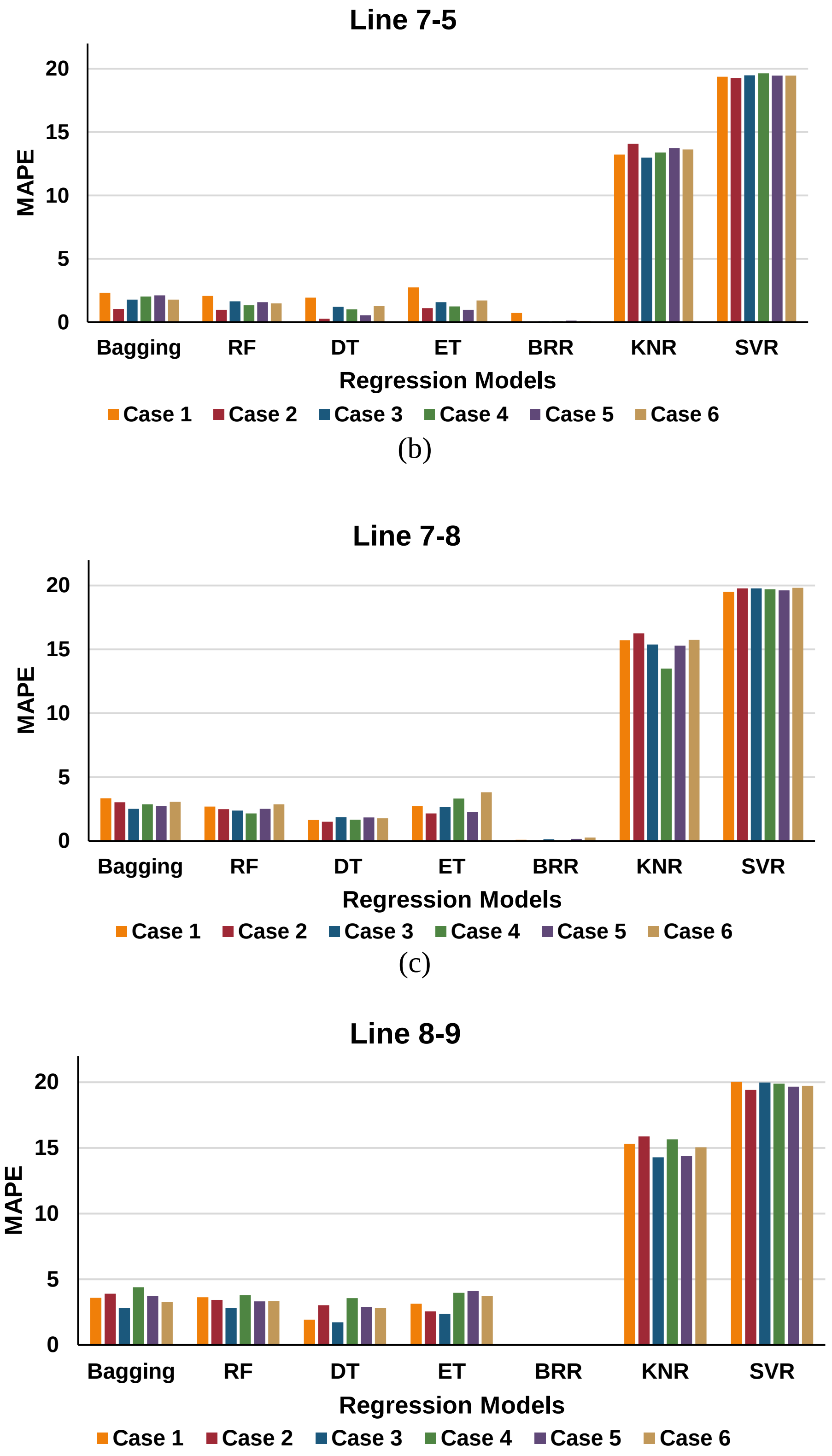
Bagging classification accuracy reached nearly 100% with a 2% new FD from case 2 onwards, while DT reached almost 100% accuracy with a 4% new FD in training. The LR and Ridge classifier performance also improved with the inclusion of increased new fault data in their training. However, the accuracy could only reach 94% for LR and 98% for Ridge, even after adding 12% new fault data to their training, which is not shown in the plots. Combining predictions, leveraging parallel computing capabilities, and robustness to overfitting are key factors contributing to the adaptability of RF and ET models. Additionally, the gradient boosting and regularization features of XGBoost further enhance its capacity to quickly adapt to new fault data scenarios for fault classification [237], [238]. The performance of SVM, AdaBoost, Gaussian NB, and MLP classifiers showed only little improvement; therefore, their results are not included in the plots.

5.6.2 Fault Localization

When analyzing all six RES integration cases for fault localization, the Mean Absolute Percentage Error (MAPE) obtained for ML models is listed in Table 5.6 for 8% RES integrated FD in their training. The column plots of the obtained results shown in Figure 5.9 indicate that the BRR model performs the best among the other models, consistent with its performance in conventional power system. Bagging, RF, RT, and ET performed satisfactorily. However, the performance of KNR and SVR remains poor. Thus, it can be concluded that the BRR model learns most effectively with minimal fault data under such structural changes in the power networks. However, in the absence of RES integrated fault data it cannot be used.



(a)



(b)

(c)

(d)

Figure 5.9 Fault location estimation MAPE on adaptability analysis post RES integration with 5% RES fault data in training for Lines: (a) Line 4-5, (b) Line 7-5, (c) Line 7-8 and (d) Line 8-9

Table 5.6 Location estimation MAPE of regression models for adaptability analysis with 8% RES integrated fault data in training.

Line	Models	Bagging	RF	DT	ET	BRR	KNR	SVR
Line 4-5	**Case 1**	2.16	1.88	0.566	2.12	0.007	15.10	19.89
	Case 2	2.15	1.92	1.805	1.73	0.018	15.10	19.89
	Case 3	2.73	2.38	1.108	2.28	0.046	15.54	19.80
	Case 4	2.63	2.59	1.696	3.46	0.042	14.41	20.27
	Case 5	2.533	2.36	0.732	3.00	0.027	15.22	20.06
	Case 6	2.260	2.18	1.288	1.00	0.020	14.80	19.68
Line 7-5	**Case 1**	2.321	2.07	1.934	2.75	0.730	13.23	19.38
	Case 2	1.036	0.97	0.285	1.11	0.053	14.08	19.27
	Case 3	1.784	1.64	1.212	1.57	0.086	12.97	19.48
	Case 4	2.025	1.32	1.019	1.23	0.073	13.38	19.64
	Case 5	2.112	1.58	0.555	0.97	0.124	13.73	19.46
	Case 6	1.781	1.50	1.285	1.71	0.106	13.63	19.47
Line 7-8	**Case 1**	3.342	2.69	1.641	2.71	0.095	15.71	19.49
	Case 2	3.015	2.48	1.505	2.15	0.023	16.24	19.77
	Case 3	2.509	2.38	1.866	2.65	0.123	15.38	19.76
	Case 4	2.860	2.14	1.658	3.31	0.031	13.50	19.71
	Case 5	2.736	2.52	1.844	2.27	0.163	15.29	19.62
	Case 6	3.063	2.86	1.766	3.81	0.278	15.73	19.82
Line 8-9	**Case 1**	2.811	3.63	1.930	3.14	0.015	15.30	20.01
	Case 2	3.275	3.43	3.034	2.55	0.036	15.86	19.42
	Case 3	3.590	2.80	1.720	2.36	0.052	14.27	19.96
	Case 4	3.895	3.31	3.565	3.97	0.035	15.64	19.87
	Case 5	3.740	3.34	2.894	4.09	0.022	14.36	19.66
	Case 6	4.384	3.37	2.813	3.72	0.027	15.03	19.72

The adaptability analysis of ML models in locating transmission line faults was conducted by gradually increasing the percentage of new fault data samples mixed with conventional power system fault data for training the ML models. Thus, all the regression models were trained for all RES integration cases by incorporating 2%, 4%, 8%, and 12% RES integrated fault data samples. This was done to analyze which regression model learns quickly after RES integration and to understand the effect of integrated RES size (MW) on the adaptability performance of studied regression models. The test results for Bayesian ridge regression, Bagging, RF, ET, and RT are presented in Figures 5.10 to 5.13 for all six cases. The axes represent the percentage fault data inclusions (FDIs) of the RES integrated system used in the training of ML models, while the distance from the center represents the MAPE of fault localization. The improvement in localization performance of the ML regression model with the increasing percentage of new fault data in training can be visualize from radar charts of Figures 5.10 to 5.13 as with increasing FDI percentage the cut on axes reduces for each model. The small red portion in the center of the plot depicts that BRR exhibited the least MAPE for all FDI conditions considered while training.

The MAPE for BRR reduced considerably with only 2% inclusion of new fault data in its training set, which is much lower than that of other models. Therefore, a separate plot for the adaptability trend of the BRR model has been presented in Figure 5.14. BRR reached performance saturation with only 2% of data; in contrast, Bagging, RF, and ET reached performance saturation at 8% inclusion of new fault data in their training sets. KNR showed only minor improvement even after including 12% new fault data in its training and SVR showed negligible improvement hence they have not been included in the plots. Thus, the BRR demonstrates the efficient transfer of learning with minimal inclusion of new data.

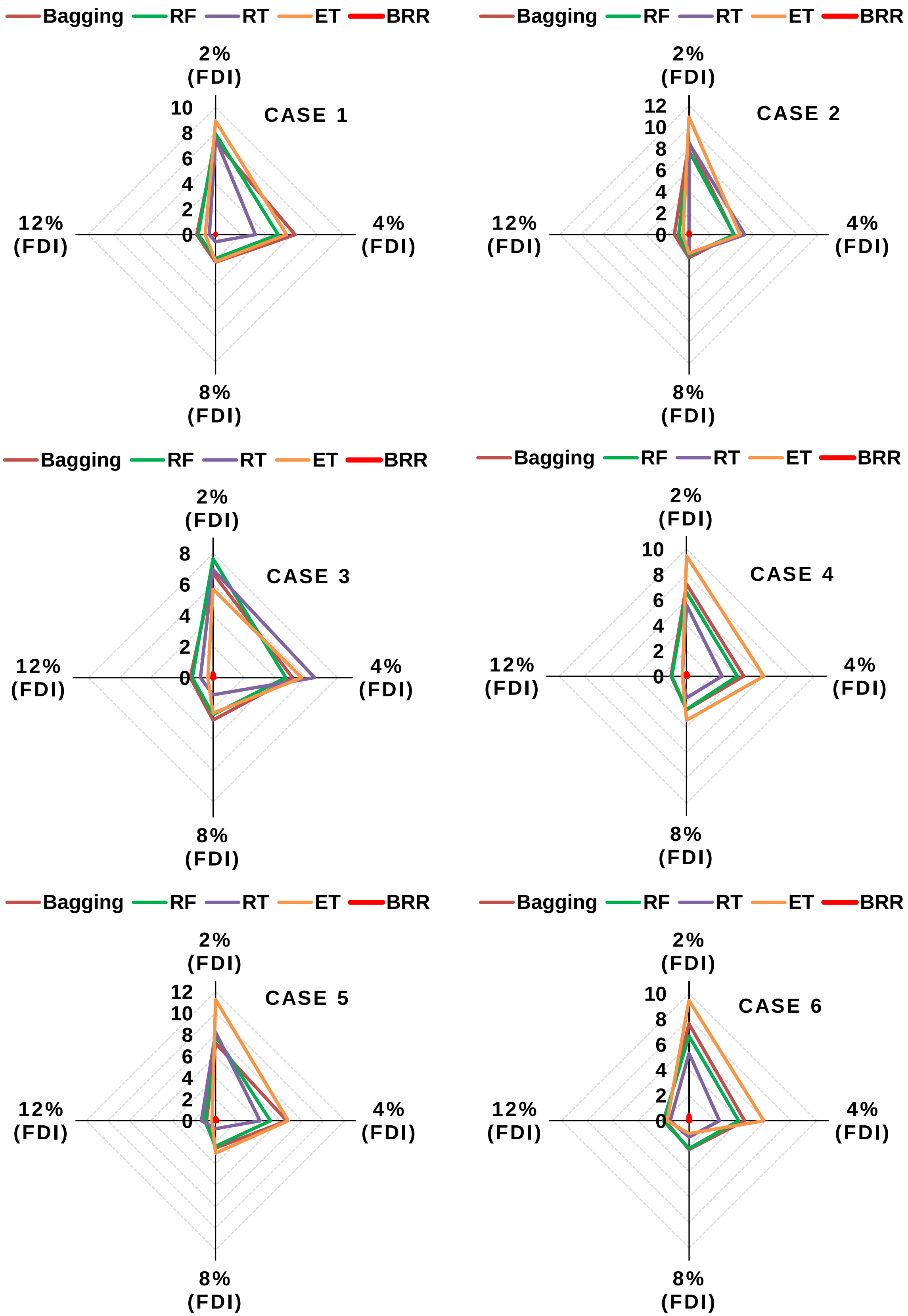


Figure 5.10 Case wise models localization performance for Line 4-5 using radar chart: MAPE Vs. Percentage fault data inclusion.

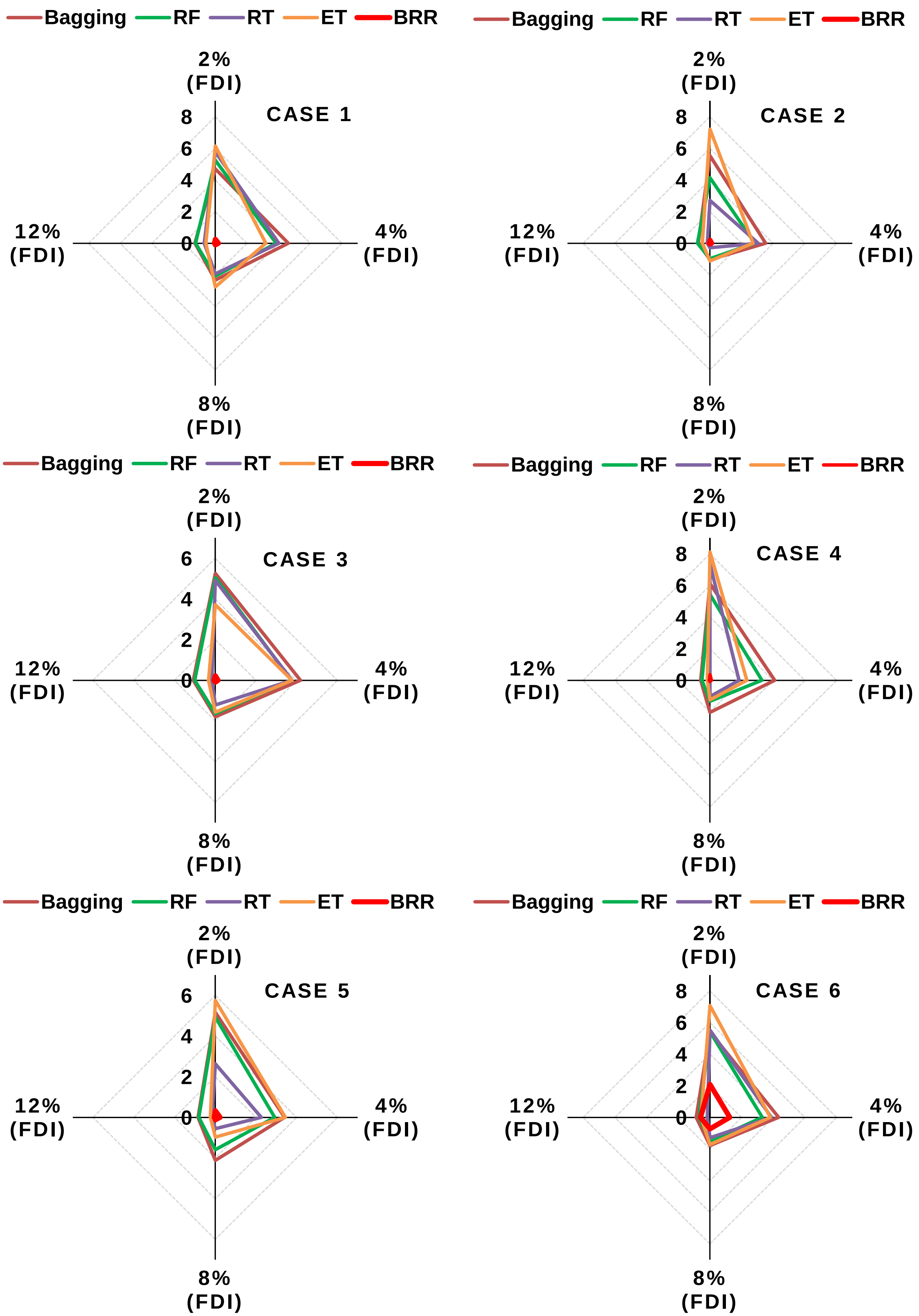


Figure 5.11 Case wise models localization performance for Line 7-5 using radar chart: MAPE Vs. Percentage fault data inclusion.

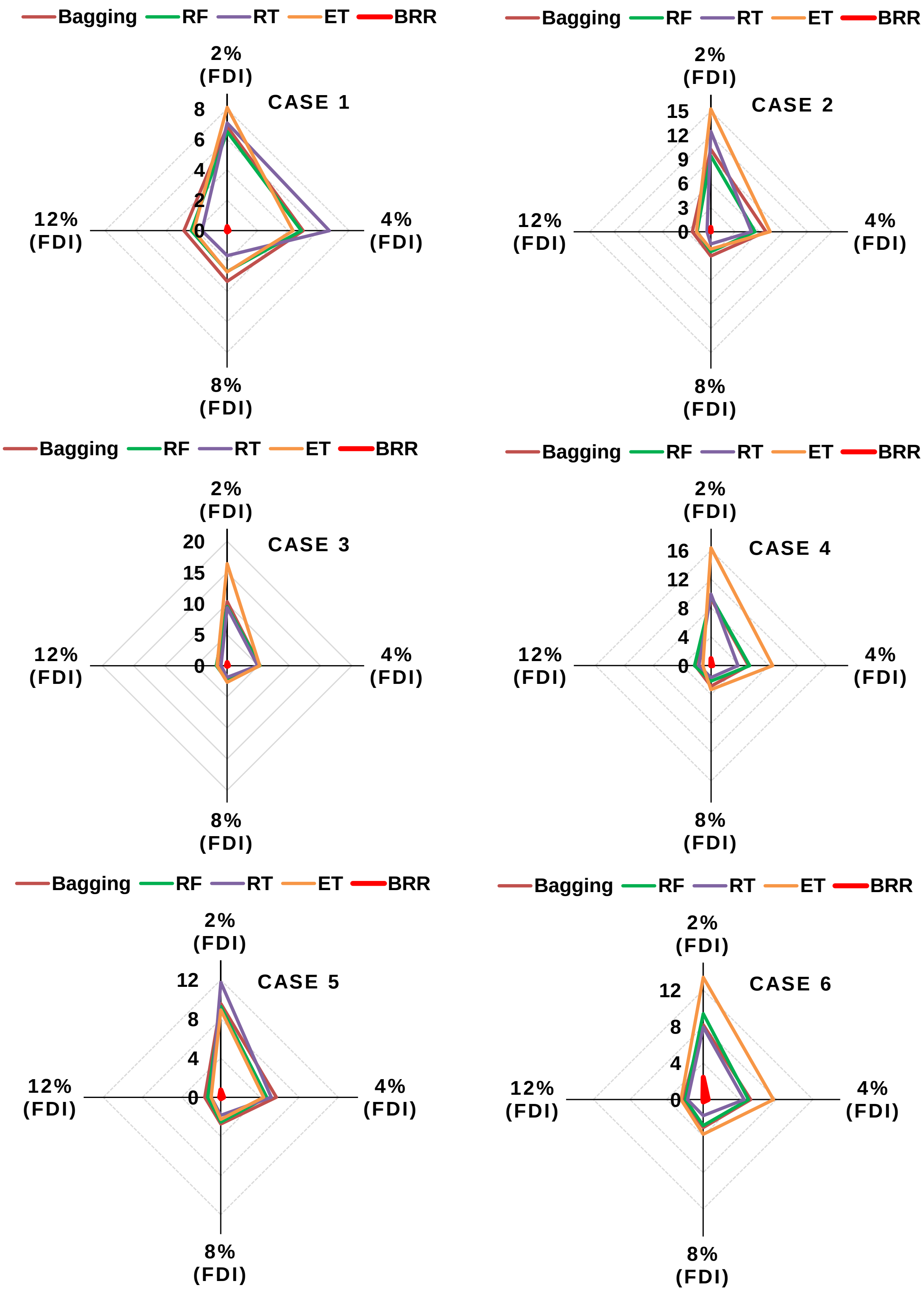


Figure 5.12 Case wise models localization performance for Line 7-8 using radar chart: MAPE Vs. Percentage fault data inclusion.

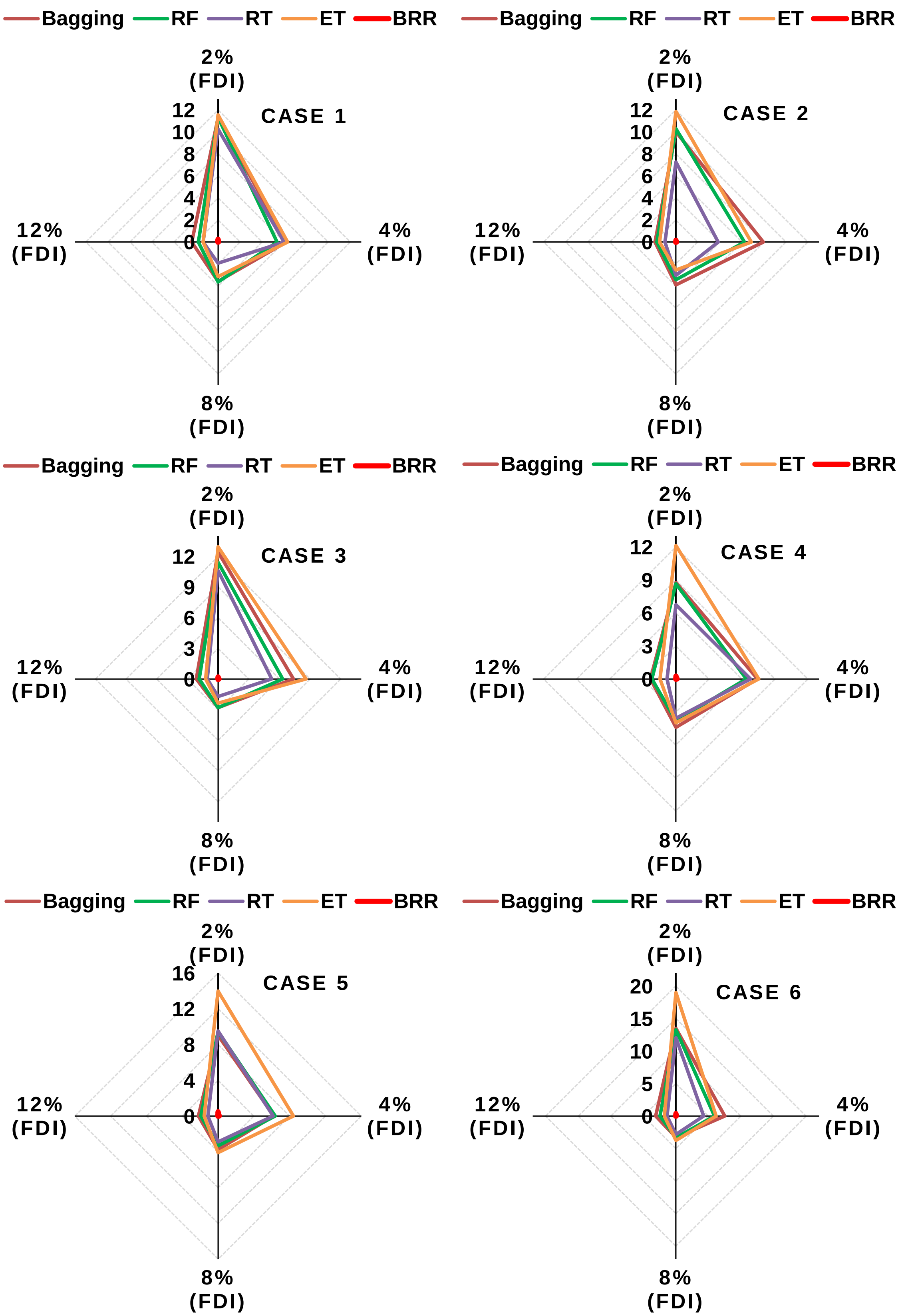
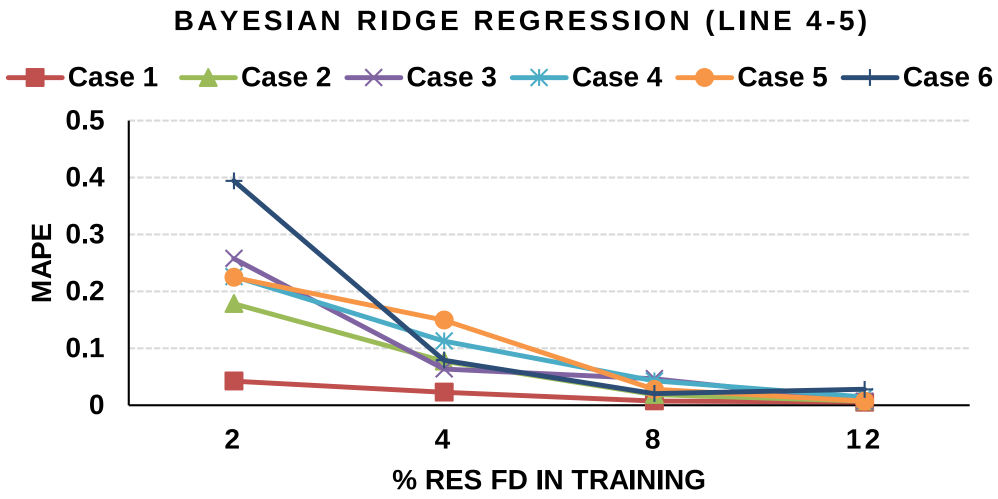
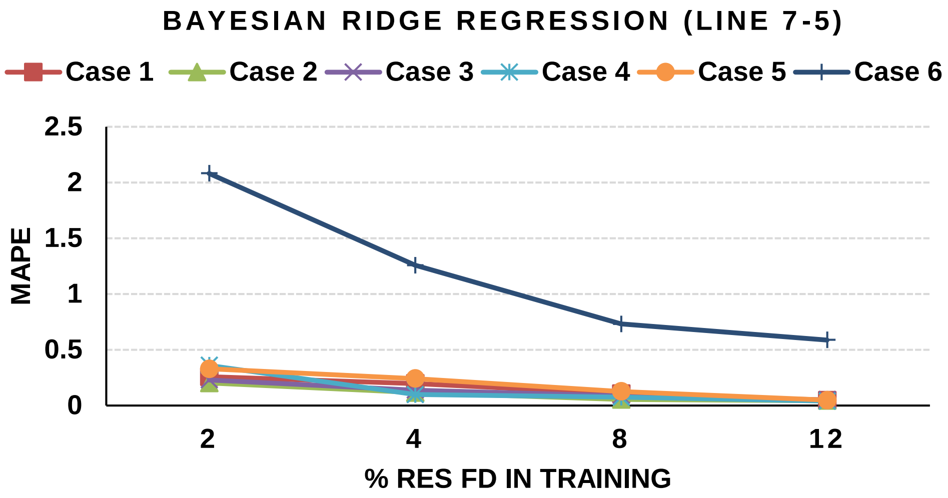


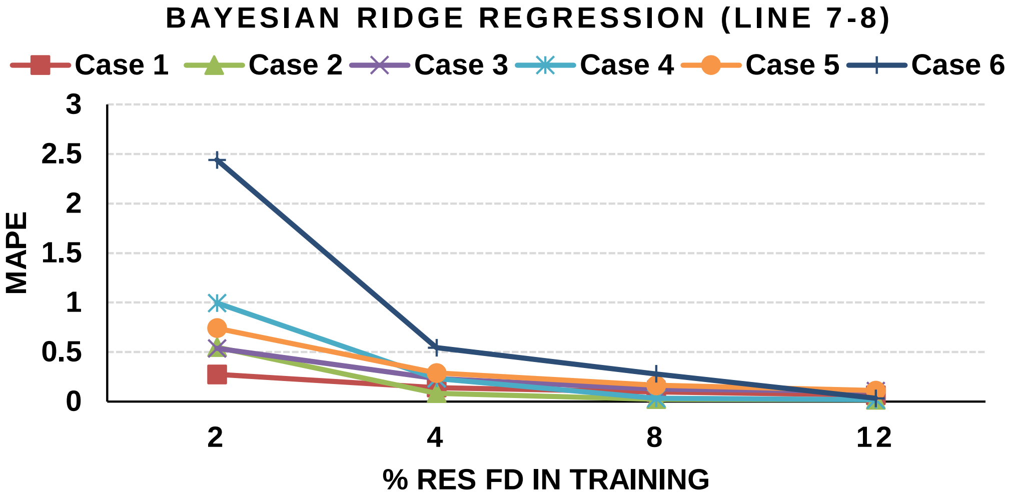
Figure 5.13 Case wise models localization performance for Line 8-9 using radar chart: MAPE Vs. Percentage fault data inclusion.


BAYESIAN RIDGE REGRESSION (LINE 4-5)
Case 1
Case 2
Case 3
Case 4
Case 5
Case 6
MAPE
0.5
0.4
0.3
0.2
0.1
0
2
4
8
12
% RES FD IN TRAINING

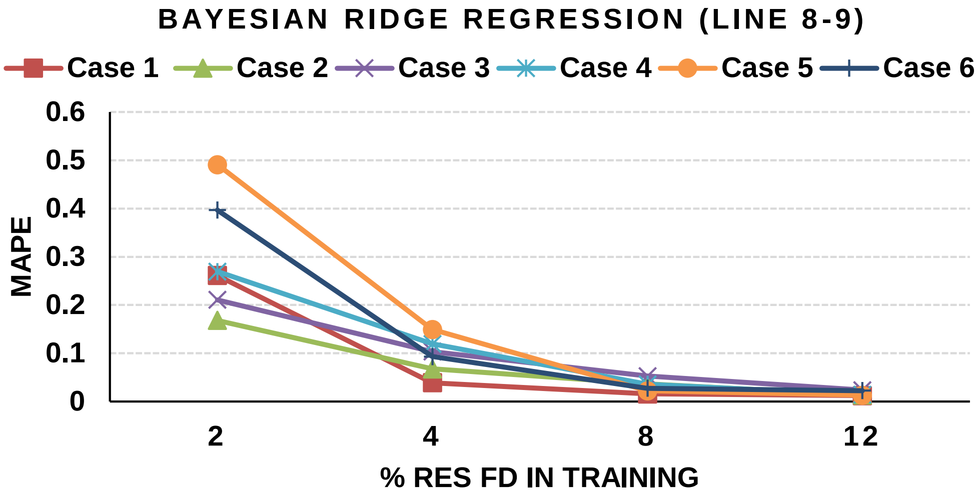
(a)


BAYESIAN RIDGE REGRESSION (LINE 7-5)
Case 1
Case 2
Case 3
Case 4
Case 5
Case 6
MAPE
2.5
2
1.5
1
0.5
0
2
4
8
12
% RES FD IN TRAINING

(b)


BAYESIAN RIDGE REGRESSION (LINE 7-8)
Case 1
Case 2
Case 3
Case 4
Case 5
Case 6
MAPE
3
2.5
2
1.5
1
0.5
0
2
4
8
12
% RES FD IN TRAINING

(c)



(d)

Figure 5.14 Learning trends of Bayesian Ridge Regression model for different sizes RES: MAPE Vs available fault data: (a) Line 4-5, (b) Line 7-5, (c) Line7-8 and (d) Line 8-9.

5.7 Result Analysis

The impact analysis outcomes revealed that on new RES integration, when ML models trained on conventional power system fault data were tested for RES integrated system faults, significant degradation in fault classification and localization performance was observed across all ML models. This study confirms that even integrating a 10MW RES plant into an IEEE 9 bus system, with a total load capacity of 315MW and 115MVAR, brings about significant changes in the fault signature. Additionally, as the size of RES integration increases, the classification accuracy of all ML models exhibits a diminishing trend. Furthermore, a marked decrease in accuracy is observed with an increase in the number of lines directly connected to the BUS integrating RES, i.e., if the incoming RES is added at a new location (BUS). The KNN classification model demonstrated relatively better results for fault classification, making it a reliable average classifier for unseen event scenarios. The accuracy of the KNN classifier remains relatively stable due to its reliance on the similarity principle. Its non-parametric approach requiring no training phase makes it suitable for changed systems [42]. Hence, without proper study, ML models cannot be entirely relied upon for power system fault diagnosis under such structural changes.

Finally, the adaptability analysis outcomes revealed that when the availability of fault data is increased gradually, ML models regain their initial performance, as demonstrated for conventional power system. The analysis identified XGBoost, RF, and ET as the fastest learning classification models, as shown by the trend plots of testing accuracy versus data availability. Although KNN's performance improved compared to impact analysis, it failed to reach the baseline performance of conventional power system due to inadequate training data, highlighting the limitations in KNN's learning capacity despite its effectiveness as a classifier with generalization abilities for changed scenarios. The intrinsic capability for incremental learning by the XGBoost algorithm supports XGBoost results outcomes. In incremental learning, new models are trained on new data while retaining the knowledge learned from previous data [239]. XGBoost supports incremental learning primarily due to its boosting algorithm and the nature of its implementation. Since XGBoost builds trees sequentially, it's conducive to updating the model with new data. XGBoost is highly optimized for speed and efficiency [212]. The design and implementation of XGBoost make it well suited for incremental learning tasks, allowing it to efficiently adapt to new data while leveraging the knowledge gained from previous iterations [239]. Therefore, it is suitable for classifying transmission line faults post RES integrations with gradual fault data availability. Similarly, Bayesian ridge regression emerged as the fastest learning model for fault localization, as indicated by the trend plots of MAPE versus data availability. BRR is based on a Bayesian approach to linear regression. In Bayesian inference, the prior beliefs about the parameters of the model get updated as new data becomes available. This incremental updating of beliefs aligns well with the concept of incremental learning required post RES integrations. BRR naturally incorporates regularization to prevent overfitting. By updating the posterior distribution of the parameters with new data, the regularization term can adapt to the changing characteristics of the dataset, ensuring that the model remains regularized even during incremental learning. Overall, the Bayesian framework, efficient computation, regularization, and memory efficiency make BRR well suited for incremental learning tasks [240]. Thus, the proposed BRR model can be utilized for automatic fault localization of transmission network faults using voltage and current measurements at transmission line buses post RES integrations with gradual fault data availability.

5.8 Chapter Summary

This chapter presented a two-facet performance analysis of potential ML models for classification and location estimation of transmission network faults. Firstly, the impact analysis of different size RES integrations on the performance of these models showed that the classification accuracy of all tested classifiers except KNN falls into the 40-50% range, making them unsuitable for such scenarios. Fault localization results of all regressors are also significantly impacted by new RES integration, including KNR, and degradation increases with an increase in the size of integrated RES. KNN’s performance falls from the previously obtained accuracy of 98.86% in conventional power system to 90.67% on 10MW RES integration at bus 7 and 80.39% on 30MW RES integrations at both the optimal integration buses. Thus, the KNN model can be relied upon for scenarios where system topology has changed, and no-fault data is available for the changed transmission network.

Lastly, the adaptability testing of these models was performed considering the practical case of gradual fault data availability over time with different sizes of RES integration. XGBoost, ET, and RF were found to be the best performers, giving 100% classification accuracy with just 0.5% new scenario's fault samples in their training on 10MW RES integration at bus 7 and 2% new fault samples of the maximum allowed 30MW RES placed at bus 7 and 5. Similarly, Bayesian ridge regression outperformed other compared regression models in the adaptability testing requiring only 2% changed network fault samples to give 1% MAPE. Although when no fault data is available for changed topology, it could not fit the model, however, as soon as it gets a few data of changed topology, it can very quickly learn the changed fault patterns and, hence, can be used for practical power system fault localization.

Chapter 6 PERFORMANCE ASSESSMENT OF MACHINE LEARNING MODELS FOR FAULT CLASSIFICATION AND LOCALIZATION UNDER INCREASING RES PENETRATIONS

6.1 Introduction

Machine learning (ML) models have been extensively researched and are deemed suitable for automated power system fault diagnosis, enabling rapid restoration of power networks. However, the increasing penetration of renewable energy sources (RES) into power systems poses challenges to ML-based fault diagnosis. The intermittent nature of RES, such as solar and wind power, causes significant fluctuations in the power fed into the grid, impacting the short circuit levels of transmission lines. This integration alters network topology and fault signatures, raising concerns about the continued accuracy of ML models trained on fault data from lower RES penetration levels, especially as fault data for new penetration levels is not immediately available following new RES integrations. Additionally, as new fault data at increased RES penetration levels become available over time, there is a need to analyze the incremental learning performance of ML models. Incremental learning allows continual adaptation to new fault data. This study investigates the adaptability and learning ability of various ML models using an incremental learning approach, focusing on real power system scenarios where fault data is initially unavailable or gradually becomes available with increasing RES penetration. Fault conditions are simulated under varying temperatures and irradiance levels specific to solar PV integrations to generate comprehensive fault datasets. Our findings demonstrate that the proposed XGBoost, among the tested ML classifiers and regression models, exhibits superior performance in accurately classifying and locating transmission line faults amidst ongoing RES integrations.

Thus, in this chapter, incremental learning is applied to various ML models to test their fault classification and localization capabilities. Initially, the models are trained using only the fault data from lower penetration levels to assess their adaptability to increasing RES penetration levels. Subsequently, fault data from current penetration levels are incrementally included in the training to examine the models' learning performance. Initial investigations revealed that ensemble methods performed better than other ML models. Consequently, models such as SVM, RT, Bagging, RF, and XGBoost are tested for fault classification and localization to analyze their adaptability and learning competence. The results indicate that ensemble methods demonstrate superior adaptability and learning ability due to their advantages, including reduced variance and bias, minimized overfitting, the capacity to handle com-plex data with good interpretability, and computational efficiency [241]. The main objectives of the proposed study are as follows:

1) To investigate the adaptability of studying ML classifiers for transmission line fault classification with increasing penetration of RES when fault data for the current penetration is unavailable.
2) To evaluate the learning capability of ML classifiers using an incremental learning approach, focusing on fault classification amidst increasing RES penetration where fault data for the current penetration gradually becomes available over time.
3) To assess the adaptability of considered ML regressors for precise localization of transmission line faults under increasing RES penetration conditions under the unavailability of the current penetration fault data.

4) To analyze the learning potential of ML regressors through an incremental learning approach, aiming to enhance fault localization accuracy as fault data for increasing RES penetration levels becomes progressively accessible.

6.2 Background

The advent of renewable energy sources (RESs) has caused a paradigm shift in the global power system. Traditional rotating synchronous generators are being replaced by inverter-based renewable energy systems, and this trend is projected to continue in the future [242]. Penetration of RES into the electric grid has been scaling drastically over the past decade [243]. The increasing penetration levels of RES-based power generation from variable and less predictable sources, such as wind and solar energy, necessitate the re-exploration of several aspects of power systems, such as grid frequency stability [242], flexibility [244], voltage stability [245], steady state and transient analysis [246], and protection [12]. High PV penetration affects the operation of protective devices as they change the direction of power flow. Furthermore, with RES penetration levels reaching up to 40%, the short circuit current can increase to as much as seven times the normal level [12]. The increase in short circuit current level with RES integration and increase in their penetration level has been illustrated using Figure 1.4 of chapter 1. This significant rise can disrupt the normal operation of relays that were designed for normal short circuit currents, necessitating the upgradation of protection devices [12], [13]. Nevertheless, the inclusion of large-scale solar plants in the power system reduces the power generation from thermal plants in the area, thus reducing the dependency on conventional coal and gas-based resources for power generation. This shift ultimately helps reduce the emission of greenhouse gases.

"Furthermore, due to their advantages, such as significant reduction in transmission losses, improvement in voltage profiles, and alleviation of the burden on natural resources, their installation is rapidly increasing [234].

The inclusion of RES also produces undesirable effects on power systems, primarily in the installation area. The grid's inertia is gradually reducing as synchronous generators are being replaced by low-inertia RES power plants, which might lead to considerable issues with grid frequency stability [242]. Eftekharnejad et al. [246] conducted an analysis of the impact of increased penetration of photovoltaic generation on power systems and found that the voltage magnitude is the most impacted power system parameter. They observed overvoltage occurrences at transmission line buses for penetration levels of 20% and above. During transient events in systems with high PV penetration, a significant decrease in voltage was noted post-fault. Furthermore, their study analyzed the impact of integrating photovoltaic (PV) systems at penetration levels of up to 50% by reducing the proportion of conventional power generation. The analysis revealed that high levels of PV penetration have both beneficial and detrimental impacts on the transient stability of the system. The degree of impact depends on the level of PV penetration, system topology, fault type, and fault location [246].

Increasing the penetration level of RES into power systems also brings many challenges to the existing protection and maintenance schemes [12], [13]. The dependence of RES on weather conditions causes the power fed to the grid to keep changing, resulting in fluctuations in the short circuit level of the transmission lines. RES introduces bidirectional power flow in the lines, increases the short circuit current level, and necessitates the upgradation of protection devices [13]. However, limited works have been reported in the literature for ML-based power system fault diagnosis with RES integration. An IEEE 13 bus test system, incorporating two distributed generators (DGs) of 1.8 MVA and 2.6 MVA, was used to study single-line-to-ground (SLG) faults, faulty phase, and faulty segment identification under varying DG penetration levels of 20%, 30%, and 50%. This study employed NN, SVM, bagging, and AdaBoost. The results indicate that as the penetration of DG increases, the accuracy of faulty phase and faulty segment identification decreases [67]. The findings highlight that increasing DG penetration presents challenges in maintaining accuracy in faulty phase and segment identification. Another study presented the SLG fault detection under varying distributed energy resource (DER) penetration levels of 30%, 40%, 50%, and 60% into the distribution network using support vector data description (SVDD) with increasing data volumes [16]. Furthering this research, a study on the IEEE 123 test feeder with 50% DER penetration into the distribution network examined the SLG fault phase and faulty segment identification, providing essential data and methodologies for tackling real-world challenges [65]. Additionally, a study on a 5-bus test system of an 11kV medium voltage distribution line with two DGs employed SVM, KNN, and bagging for fault classification, and GPR, regression trees (RT), support vector regression (SVR), and linear regression for fault localization [66]. Although focused on a specific fault impedance, this research lays the groundwork for expanding datasets to reflect a wider range of fault attributes. Moreover, fault classification work on a 20kV distribution network with two DGs of 9 MW utilized SVM, DT, KNN, and convolutional neural networks (CNN), demonstrating the effectiveness of various machine learning approaches in complex power systems [17]. Another noteworthy study on an IEEE 9 bus system, incorporating a 120 MW wind energy source, employed ANN and SVM for system fault detection and classification, underscoring the potential of these models in renewable-integrated systems [64]. Additionally, research on the IEEE 39 bus system with a 1.5 kVA wind generator at Bus 39 focused on faulty line identification using deep learning techniques, contributing to the growing body of knowledge on advanced fault diagnosis [63]. These studies collectively advance the understanding of fault detection, classification, and localization in power systems with renewable energy sources. While they have made substantial progress, there remains an opportunity to address practical challenges, such as the immediate availability of fault data following increased RES penetration, to further enhance the robustness and reliability of power systems.

Based on the literature review, it is evident that ML fault classifiers perform effectively in power systems. However, their performance in systems integrated with RES has not been thoroughly tested. The impact of increasing RES penetration on ML classifier performance in power networks is still largely unknown. With the increasing penetration of RES into transmission networks, it is essential to assess ML models' performance under increasing RES penetrations before deploying them in real-world systems. Most fault localization techniques focus on identifying faulty lines or sections within transmission and distribution networks using classification methods. Accurately pinpointing the exact fault location on a line with ML regressors could significantly improve maintenance efficiency. Currently, no research addresses the precise fault location on transmission lines in RES-integrated systems with increasing RES penetration using ML-based classification or regression techniques. This suggests the necessity for research studies on ML regression-based fault location estimation in transmission lines under increasing RES penetration conditions.

Further, the variability in power generation from RES, influenced by weather conditions like temperature and irradiance fluctuations, impacts the power output from RES plants. These fluctuations lead to variations in fault current levels, which should be considered when analyzing RES-integrated transmission networks. Existing ML-based power system fault diagnosis literature often overlooks these crucial aspects. An increase in the RES penetration level of an existing transmission network significantly alters its topology and fault characteristics, depending on the level of increase. Large fluctuations in RES power generation can cause substantial deviations in fault currents, potentially increasing ML model misclassification rates and fault location errors. To date, no study has examined ML models' performance for transmission line fault classification and localization after increasing RES penetration of varying levels. Research is needed to understand how ML models for fault diagnosis in existing power systems are affected by increasing RES penetration. Additionally, there is a shortage of studies on ML-based fault diagnosis in RES-integrated power systems. The studies available generally assume long-standing RES integration and sufficient fault data representing diverse variations for ML model training.

When the existing level of RES penetration increases, fault data availability can become limited, and accumulating diverse data on the increased penetration level, including various fault locations and types, may take years. Therefore, analyzing ML models' performance considering these practical issues is crucial to ensure uninterrupted power transfer and meet the evolving needs of power system protection and maintenance. Furthermore, as RES integration progresses, it is observed that SLG faults are more prevalent compared to triple line faults (LLLG and LLL), yet diverse fault data encompassing various locations, types, fault inception angles, and resistances remains scarce for training ML models at each penetration level of RES. This gap underscores the necessity for comprehensive analysis in power system fault classification and localization with increasing RES penetration, acknowledging the challenges of limited data availability.

Power flow in transmission lines is generally unidirectional; however, the integration of RES at various points creates tapping points and facilitates bi-directional power flow. Moreover, RES can supply fault currents during faults, leading to increased fault currents in the lines. With a 40% increase in RES penetration, the short circuit current may surge up to seven times its normal level [12]. As a result, the fault signature at a specific location will change with the addition of a new RES unit, even if the fault condition and location remain unchanged. This raises concerns about ML models' sustained accuracy for diagnosing faults in real-world power systems amidst ongoing RES integrations. The performance of ML models becomes questionable as they are trained using fault data from lower RES penetration levels, mainly since fault data for higher penetration levels isn't immediately accessible following new RES integrations. New fault data at increased RES penetration levels become available only over time. Despite this, there is a scarcity of literature on ML-based power system fault diagnosis with RES integrations considering fault data unavailability.

Therefore, in this chapter, ML models' adaptability and learning competence have been analyzed for fault classification and localization when the penetration level of existing RES is increased at the point of connection in the transmission network for varying weather conditions. This has been achieved by utilizing an incremental learning approach. The simple ML approach utilizes a complete dataset for training. However, in the context of incremental or continual learning, the data is received in a sequential manner or in several steps, and the underlying distribution of data evolves over time [247]. Incremental learning allows machine learning models to learn from large-scale dynamic stream data and build up a knowledge base over time to improve future learning and decision-making processes [248]. An incremental learning approach has been adopted for power system fault detection in [157] and [16] utilizing shape preserving scheme and incremental SVDD respectively.

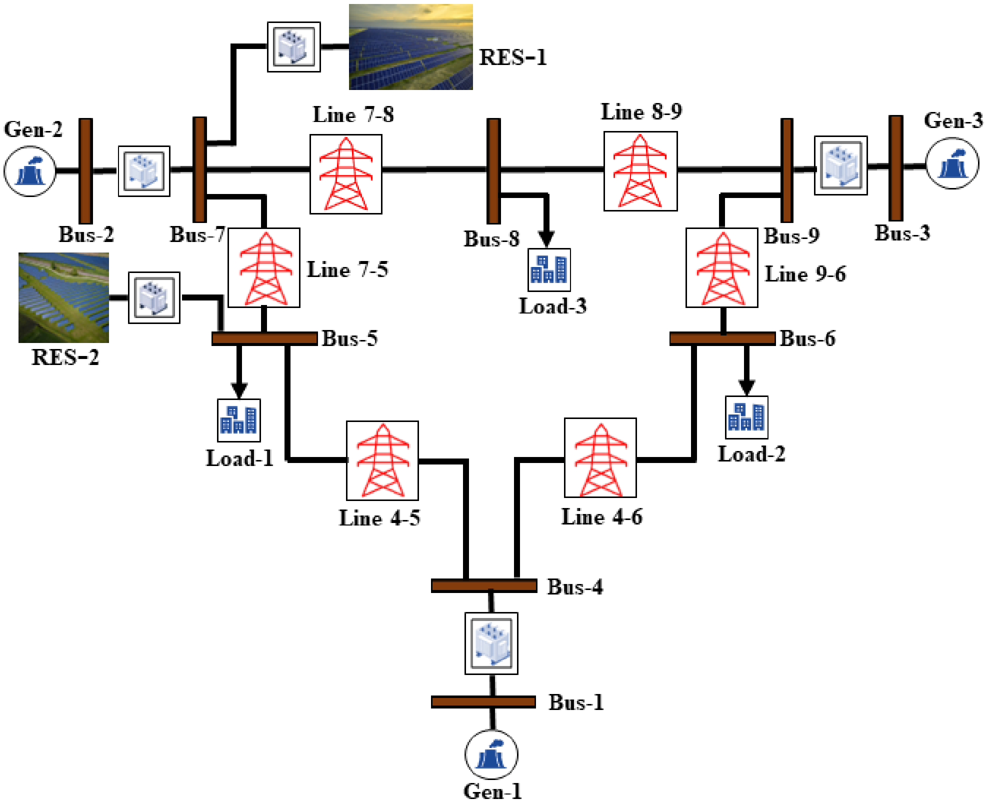


Figure 6.1 Standard IEEE 9 Bus System with RES (Solar PV Plants) at Bus-7 and Bus-5.

6.3 Proposed Methodology

The standard IEEE 9 Bus system with RES integration fault data has been taken under study to conduct the proposed study. The schematic diagram of the system under study has been presented in Figure 6.1. This chapter investigates two practical possibilities for analyzing the performance of ML models for fault diagnosis with increasing RES penetration.

1) Adaptability Investigation: When the fault data of the increased penetration level is unavailable, and ML models can only be trained using old penetration level fault data.

2) Learning competence investigation: After the increase in penetration level of RES, fault data gets collected gradually over time. Thus, ML models have also been analyzed for their learning competence using old and available current penetration level fault data. The learning competence investigation reveals how fast a model learns with the least available increased penetration level fault data.

The schematic diagram for the proposed methodology has been depicted in Figure 6.2 which illustrates RES integrated fault dataset and stages involved in the proposed adaptability and learning competence investigation. The figure also depicts the utilization of RES integrated fault data in relevant manner to conduct the proposed investigations. The proposed study schematic diagram starts with illustrating six RES integration penetration levels (PL) referred as PL-1, PL-2, PL-3, PL-4, PL-5 and PL-6. The RES integration has been done considering optimal size and placement study available in literature [10]. Following on the training and testing of ML models for both adaptability and learning competence investigations has also been illustrated.

It can be seen in the schematic diagram that adaptability investigation uses old penetration level's fault data while learning competence investigation uses both old penetration and available current penetration fault data which gradually increases in stages. Therefore, to perform the proposed investigations of ML models, the procedure adopted is as follows:

1. Selection of a suitable test system for the study.
2. Identification of the optimal location of RES placement and maximum allowable penetration level.

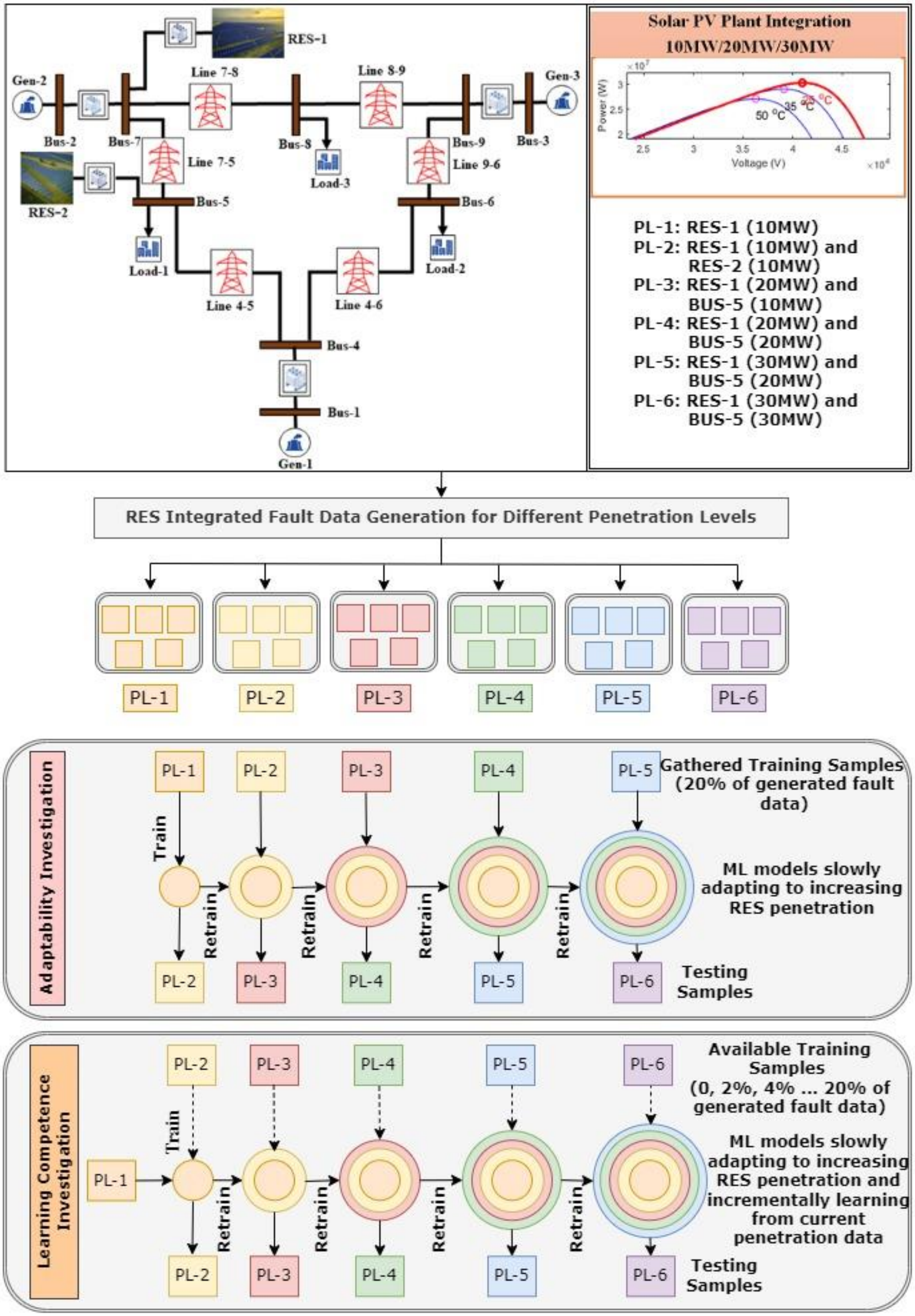

RES-1
Gen-2
Line 7-8
Line 8-9
Gen-3
Bus-2
Bus-7
Bus-8
Bus-9
Bus-3
Line 7-5
Load-3
Line 9-6
RES-2
Bus-5
Bus-6
Load-1
Load-2
Line 4-5
Line 4-6
Bus-4
Bus-1
Gen-1
Solar PV Plant Integration
10MW/20MW/30MW
Power (W)
Voltage (V)
50 °C
35 °C
25 °C
PL-1: RES-1 (10MW)
PL-2: RES-1 (10MW) and RES-2 (10MW)
PL-3: RES-1 (20MW) and BUS-5 (10MW)
PL-4: RES-1 (20MW) and BUS-5 (20MW)
PL-5: RES-1 (30MW) and BUS-5 (20MW)
PL-6: RES-1 (30MW) and BUS-5 (30MW)
RES Integrated Fault Data Generation for Different Penetration Levels
PL-1
PL-2
PL-3
PL-4
PL-5
PL-6
Adaptability Investigation
Train
Retrain
Gathered Training Samples (20% of generated fault data)
ML models slowly adapting to increasing RES penetration
Testing Samples
Learning Competence Investigation
Available Training Samples (0, 2%, 4% ... 20% of generated fault data)
ML models slowly adapting to increasing RES penetration and incrementally learning from current penetration data

Figure 6.2 Schematic diagram of the proposed study.

3. Generation of extensive fault data at different penetration levels incorporating fault attributes and varying weather conditions.
4. Investigating the adaptability of selected ML models to old penetration fault data for fault classification and localization in the absence current penetration level fault data for increasing RES penetration level.
5. Investigating the learning competence of selected ML models with gradual availability of current penetration fault data over time for fault classification and localization.

6.4 Dataset Description

In the proposed study of this chapter fault data for only RES integrated system has been used. The study aims to analyze the ML models' adaptability and learning capability with increase in RES penetration of existing RES. The study assumes that prior to the increase in RES penetration few fault data samples have been gathered over time which is referred to as old penetration fault data. The dataset encompasses various penetration levels, starting from 10 MW at Bus-7 (PL-1) and extending to 30 MW at both Bus-7 and Bus-5. Fault data is generated by varying fault attributes and adjusting power generation from solar plants across different temperatures and irradiance values. The six different combinations of penetration levels (PL-1 to PL-6), are described below:

PL – 1: Bus-7 (10MW) RES
PL – 2: Bus-7 (10MW) and Bus-5 (10MW) RES
PL – 3: Bus-7 (20MW) and Bus-5 (10MW) RES
PL – 4: Bus-7 (20MW) and Bus-5 (20MW) RES
PL – 5: Bus-7 (30MW) and Bus-5 (20MW) RES
PL – 6: Bus-7 (30MW) and Bus-5 (30MW) RES

Table 6.1 Data Description for Proposed Performance Analysis.

Study	Penetration Level Testing	Training – Testing Dataset Description
Adaptability Investigation	PL-2	Train – With PL-1 fault data Test – With Bus-7 (10MW) and Bus-5 (10MW) RES fault data
	PL-3	Train – With PL-1 + PL-2 fault data Test – With Bus-7 (20MW) and Bus-5 (10MW) RES fault data
	PL-4	Train – With PL-1 + PL-2 + PL-3 fault data Test – With Bus-7 (20MW) and Bus-5 (20MW) RES fault data
	PL-5	Train – With PL-1 + PL-2 + PL-3 + PL-4 fault data Test – With Bus-7 (30MW) and Bus-5 (20MW) RES fault data
	PL-6	Train – With PL-1 + PL-2 + PL-3 + PL-4 + PL-5 fault data Test – With Bus-7 (30MW) and Bus-5 (30MW) RES fault data
Learning Competence Investigation	PL-2	Train – With PL-1 + Available Bus-7 (10MW) and Bus-5 (10MW) RES fault data Test – With Bus-7 (10MW) and Bus-5 (10MW) RES fault data
	PL-3	Train – With PL-1 + PL-2 + Available Bus-7 (20MW) and Bus-5 (10MW) RES fault data Test – With Bus-7 (20MW) and Bus-5 (10MW) RES fault data
	PL-4	Train – With PL-1 + PL-2 + PL-3 + Available Bus-7 (20MW) and Bus-5 (20MW) RES fault data Test – With Bus-7 (20MW) and Bus-5 (20MW) RES fault data
	PL-5	Train – With PL-1 + PL-2 + PL-3 + PL-4 + Available Bus-7 (30MW) and Bus-5 (20MW) RES fault data Test – With Bus-7 (30MW) and Bus-5 (20MW) RES fault data
	PL-6	Train – With PL-1 + PL-2 + PL-3 + PL-4 + PL-5 + Available Bus-7 (30MW) and Bus-5 (30MW) RES fault data Test – With Bus-7 (30MW) and Bus-5 (30MW) RES fault data

A comprehensive dataset comprising 25,344 fault data samples has been generated for RES integrated IEEE 9 Bus system, incorporating diverse combinations of fault attributes and RES penetrations. Specifically, the dataset includes 4,224 fault data samples for each penetration level, dedicated to fault classification, along with 1,056 fault data samples for each transmission line for fault localization purposes as discussed in chapter 3. To incorporate the limited fault data collection aspect for any RES penetration level prior to increase in penetration level the study utilizes only 20% of fault data for training in the proposed studies. Thus, 20% of 4224 samples have been taken from each penetration level for training in both investigations. Further, Table 6.1 enlists the train test data split considered while training models for investigating 1) adaptability for unavailability of current penetration fault data and 2) learning competence with gradual fault data availability. The adaptability investigation refers to exploring the capability of ML models to classify and locate faults of higher RES penetrated power systems from models trained on lower penetration level RES fault data. Further, the learning competence investigation refers to exploring the fast and accurate learning ML model when fault data for the current penetration level is gradually gathered over time.

6.5 Adaptability Investigation of ML models to Old Penetration Fault Data

In the proposed adaptability investigation, ML models are trained for fault classification and localization using only old fault data from previous penetration levels to assess their performance on increased penetration level fault data for fault classification and localization. For instance, to classify and locate faults at penetration level PL-4, the models are trained using 20% of the total fault data generated from penetration levels PL-1, PL-2, and PL-3. The utilization of fault data has been illustrated using Figure 6.3 adaptability investigation block where N represents the current RES penetration level fault data and N-1 represents previous penetration level fault data. This approach ensures that all available data from previous penetration levels are utilized during training when testing for fault classification and localization at any given penetration level.

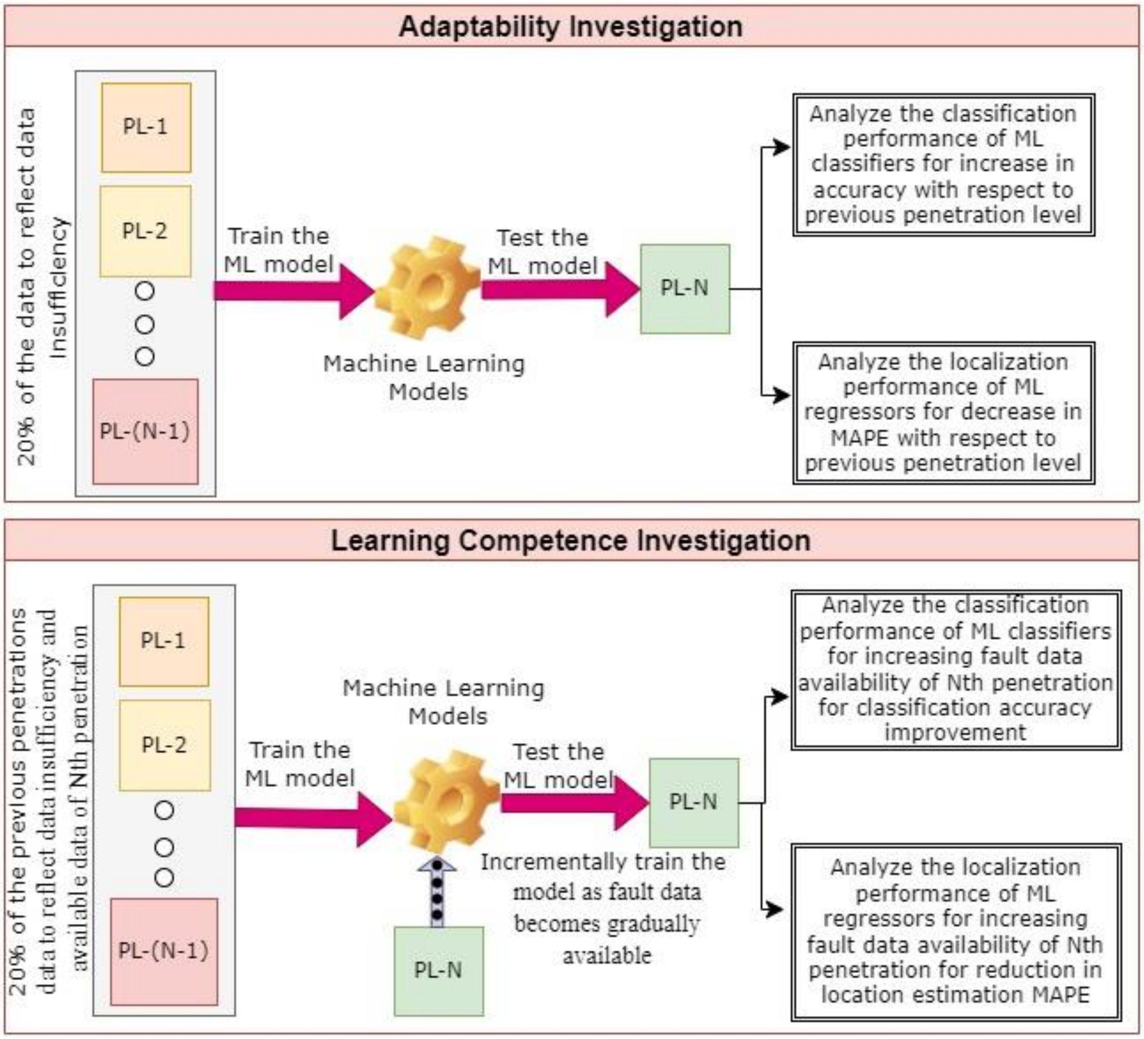

Adaptability Investigation
20% of the data to reflect data insufficiency
PL-1
PL-2
PL-(N-1)
Train the ML model
Machine Learning Models
Test the ML model
PL-N
Analyze the classification performance of ML classifiers for increase in accuracy with respect to previous penetration level
Analyze the localization performance of ML regressors for decrease in MAPE with respect to previous penetration level
Learning Competence Investigation
20% of the previous penetrations data to reflect data insufficiency and available data of Nth penetration
PL-1
PL-2
PL-(N-1)
Train the ML model
Machine Learning Models
Test the ML model
PL-N
PL-N
Incrementally train the model as fault data becomes gradually available
Analyze the classification performance of ML classifiers for increasing fault data availability of Nth penetration for classification accuracy improvement
Analyze the localization performance of ML regressors for increasing fault data availability of Nth penetration for reduction in location estimation MAPE

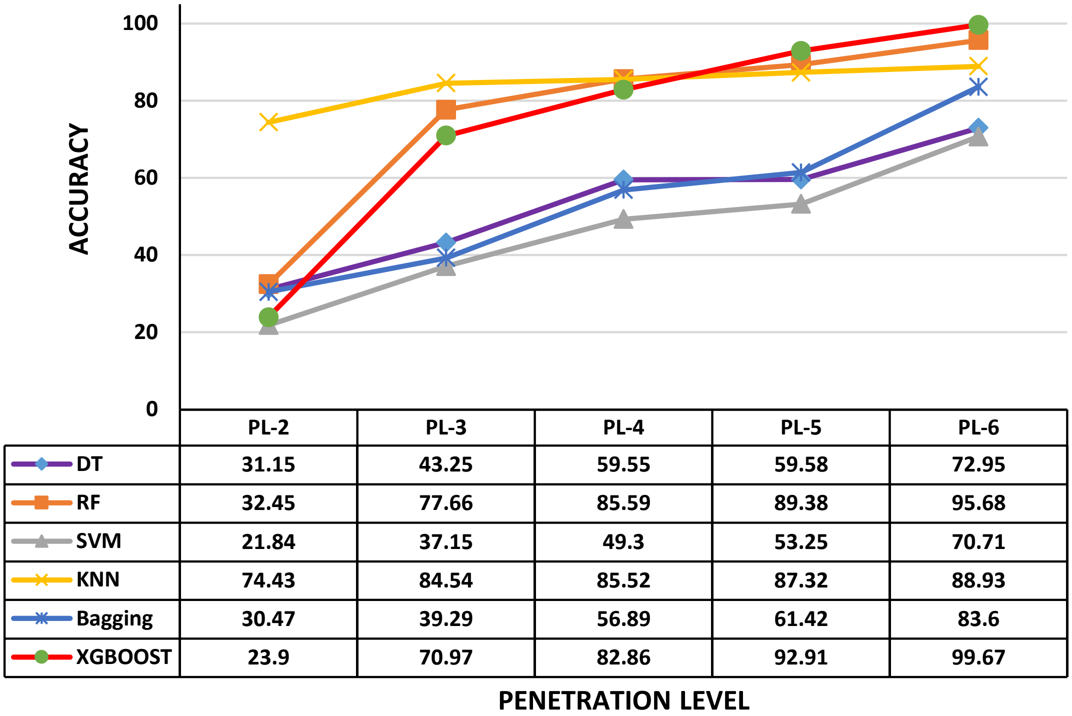
Figure 6.3 Illustration of adaptability and learning competence investigation.

A total of 4224 fault data samples have been generated for each fault penetration level, encompassing fault data from four transmission lines. However, due to constraints in data availability, particularly concerning the gradual development of large solar parks and the limited time between their operational stages, only 20% of the complete data is utilized for the adaptability investigation. This consideration helps to address data scarcity issues after RES integration. Consequently, for fault classification, 20% of the fault data translates to 840 data samples, while for fault localization, 210 data samples for each line. As a result, relatively more fault data is available for fault classification compared to fault localization.

6.5.1 Fault Classification

When the ML model's performance is tested for fault classification for various penetration levels using fault data of lower penetration levels, the testing accuracy obtained is very low for the initial penetration level, as shown in Figure 6.4. The accuracy for PL-2 is minimum for all models as training is done from fault data of 10 MW RES included at bus 7 of the IEEE 9 bus system and tested for two RES included in the studied network at bus 7 and 5 of 10 MW each. Thus, accuracy is very low for this penetration level as training is done with one RES fault data while testing is done on two RES fault data. However, KNN performed as an average classifier for all penetration levels. Similarly, its accuracy is lowest for PL-2 and almost remains constant for higher penetration levels.

As the penetration level increases, the training data has fault data for all the lower penetration level fault data; hence, when no data for the current penetration level is available for training, accuracy still improves with increasing penetration level. As ML models slowly capture the changing fault signature with increasing penetration levels. Thus, it is wise to use all the previous penetration level's fault data for training while testing for any level of RES penetration because the model slowly adapts to the increasing penetration level of RES. From the plot, it can be deduced that although RF classification accuracy is higher than XGBoost and Bagging for lower penetration levels, however, from PL-5, XGBoost accuracy surpasses the RF accuracy. Hence, when available training data is much less, KNN can be used as a classifier that generalizes the classification accuracy irrespective of considered RES penetration level. When sufficient data is available for training, RF and XGBoost can be used for fault classification even when tested for increased penetration levels. SVM and DT performed worse than ensemble methods, showing the efficacy of ensemble methods.



	PL-2	PL-3	PL-4	PL-5	PL-6
DT	31.15	43.25	59.55	59.58	72.95
RF	32.45	77.66	85.59	89.38	95.68
SVM	21.84	37.15	49.3	53.25	70.71
KNN	74.43	84.54	85.52	87.32	88.93
Bagging	30.47	39.29	56.89	61.42	83.6
XGBOOST	23.9	70.97	82.86	92.91	99.67

Figure 6.4 Testing accuracy improvement on adaptability investigation in the absence of current penetration level data with increasing penetration level.

6.5.2 Fault Localization

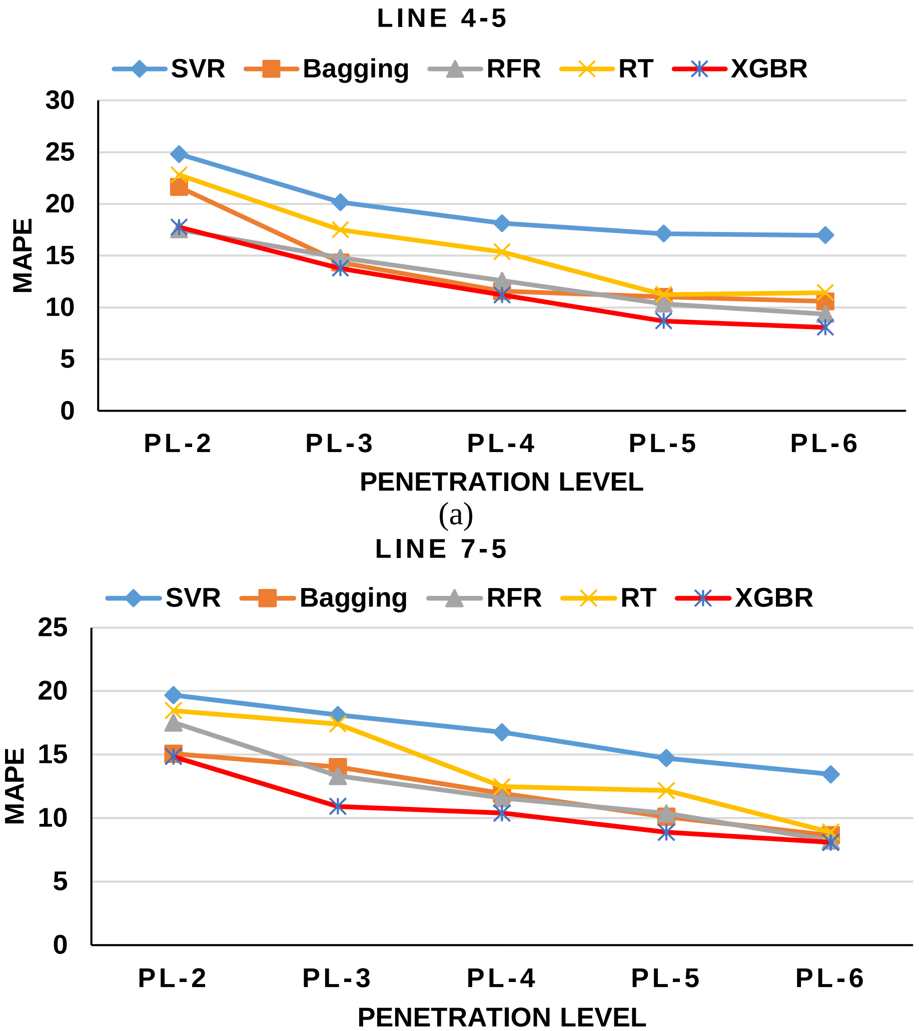
When the ML model's performance is tested for fault localization at various penetration levels using fault data of lower penetration levels, the mean absolute percentage error (MAPE) for fault location estimation on four lines of the studied network obtained is very high for initial penetration levels. However, the MAPE decreases with an increase in penetration level, as shown in Figure 6.5 and tabulated in Table 6.2. This indicates that models slowly adapt to the change in penetration level from old penetration level fault data.

Table 6.2 MAPE reduction in location estimation while adaptability investigation with increasing penetration level of RES.

	Model	PL-2	PL-3	PL-4	PL-5	PL-6
Line 4-5	**SVR**	24.817	20.158	18.127	17.124	16.968
	Bagging	21.645	14.346	11.577	11.023	10.586
	RFR	17.536	14.806	12.587	10.36	9.363
	RT	22.814	17.506	15.368	11.24	11.435
	XGBR	17.763	13.77	11.202	8.695	8.068
Line 7-5	**SVR**	19.687	18.113	16.758	14.712	13.452
	Bagging	15.065	14.015	11.927	10.092	8.635
	RFR	17.526	13.314	11.578	10.364	8.22
	RT	18.466	17.415	12.466	12.161	8.886
	XGBR	14.847	10.924	10.394	8.879	8.086
Line 7-8	**SVR**	26.785	24.57	22.785	20.847	19.417
	Bagging	24.252	23.94	20.068	13.329	13.008
	RFR	25.398	24.061	19.439	13.959	10.807
	RT	28.125	23.962	21.143	20.239	19.185
	XGBR	19.812	19.136	17.067	13.265	9.101
Line 8-9	**SVR**	21.425	20.741	19.715	18.512	16.789
	Bagging	19.7	19.124	18.115	16.203	15.821
	RFR	20.096	19.013	18.295	17.535	16.331
	RT	19.547	18.85	17.915	17.297	16.738
	XGBR	19.111	17.264	13.963	12.794	11.621

Moreover, XGBR and RFR displayed better adaptability to fault localization for the increase in the penetration level of RES. The MAPE from PL-2 to PL-6 for XGBR ranged from 17%

to 8%, for RT 22% to 11%, for Bagging 21% to 10%, for RFR 17% to 9% and for SVR 24% to 16% for line 4-5. Thus, XGBoost gave better results than RFR even without current penetration fault data in training. Similar superior performance for XGBoost has been obtained for the other three lines as shown in Figure 6.5.

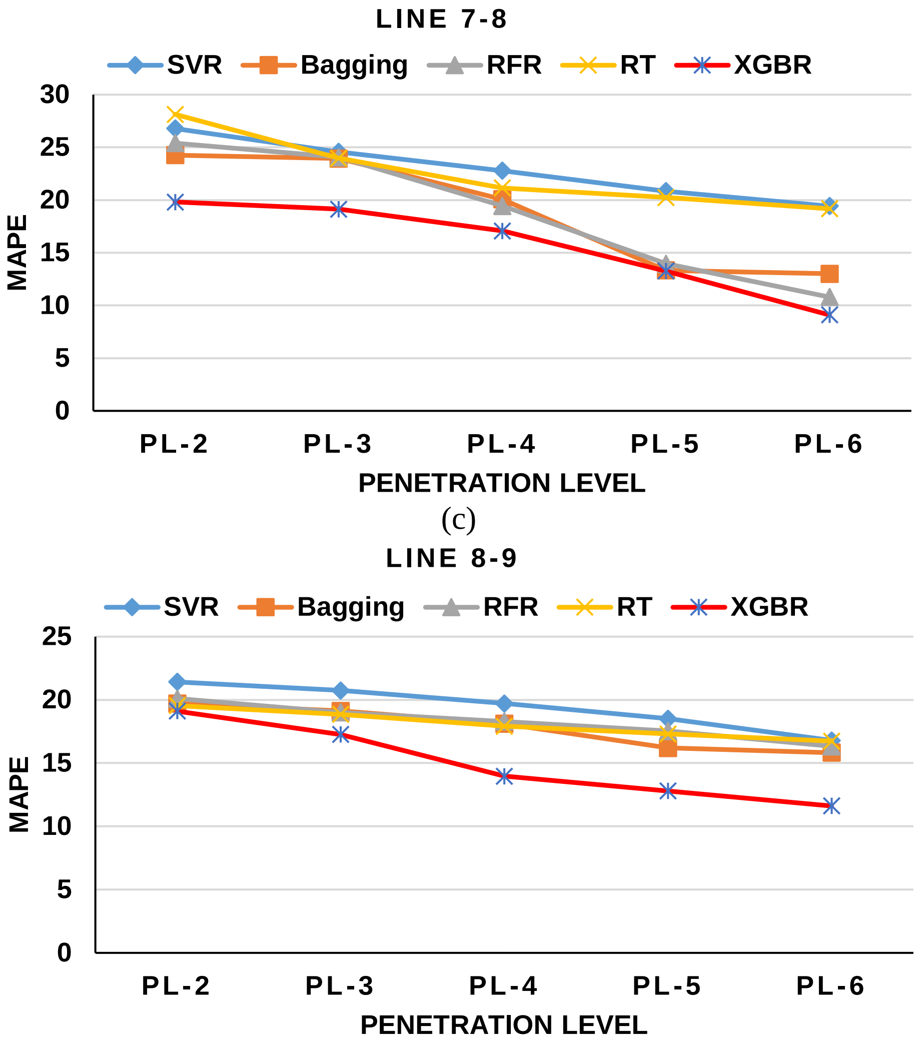


LINE 4-5

(a)

LINE 7-5

(b)



(c)

(d)

Figure 6.5 Location estimation error reduction on adaptability investigation in the absence of current penetration level data with increasing penetration level.

6.6 Learning Competence Investigation of ML models with Current Penetration Fault Data Availability

The investigation into learning competence assesses the performance of ML models in real-world scenarios characterized by the gradual accumulation of fault data over time. Consequently, ML models are trained using both historical fault data from previous penetration levels and the available fault data from the current penetration level. As the fault data collected by data recorders/PMUs increases gradually following the installation of new

RES units, the study adjusts the consideration of fault data availability for the current penetration level in stepwise increments. The performance testing of ML models begins with a minimum of 0.5% data availability for the current penetration level and increments by 0.5% of the complete data in subsequent steps. Thus, the study examines ML models performance for fault classification with data availability ranging from 0.5% to 5% of the current penetration level. Similarly, for fault localization, the learning competence of regression models is evaluated using data availability ranging from 2% to 20% of the current penetration level. For instance, to classify and locate faults at penetration level PL-4, the models are trained using 20% of the total fault data generated at penetration levels PL-1, PL-2, and PL-3, along with the available fault data from the current penetration level i.e., PL-4, which ranges from 0.5% to 5% for classification and 2% to 20% for localization.

6.6.1 Fault Classification

When ML classification models are tested for learning competence with gradual availability of fault data over time for increased penetration level, XGBoost emerged as the fastest learning classifier, as depicted in Figure 6.6 using radar plot. Figure 6.6 displays the eleven axes that reflect the fault data inclusions (FDIs) where fault data is increased incrementally while training the ML models. The radial distance from the center corresponds to the categorization accuracy, expressed as a percentage. The classification accuracies obtained have also been tabulated in Tables 6.3, 6.4, 6.5, 6.6 and 6.7 for all five penetration levels under the test. As seen from plots, XGBoost performance surpassed RF for higher penetration level fault classification.

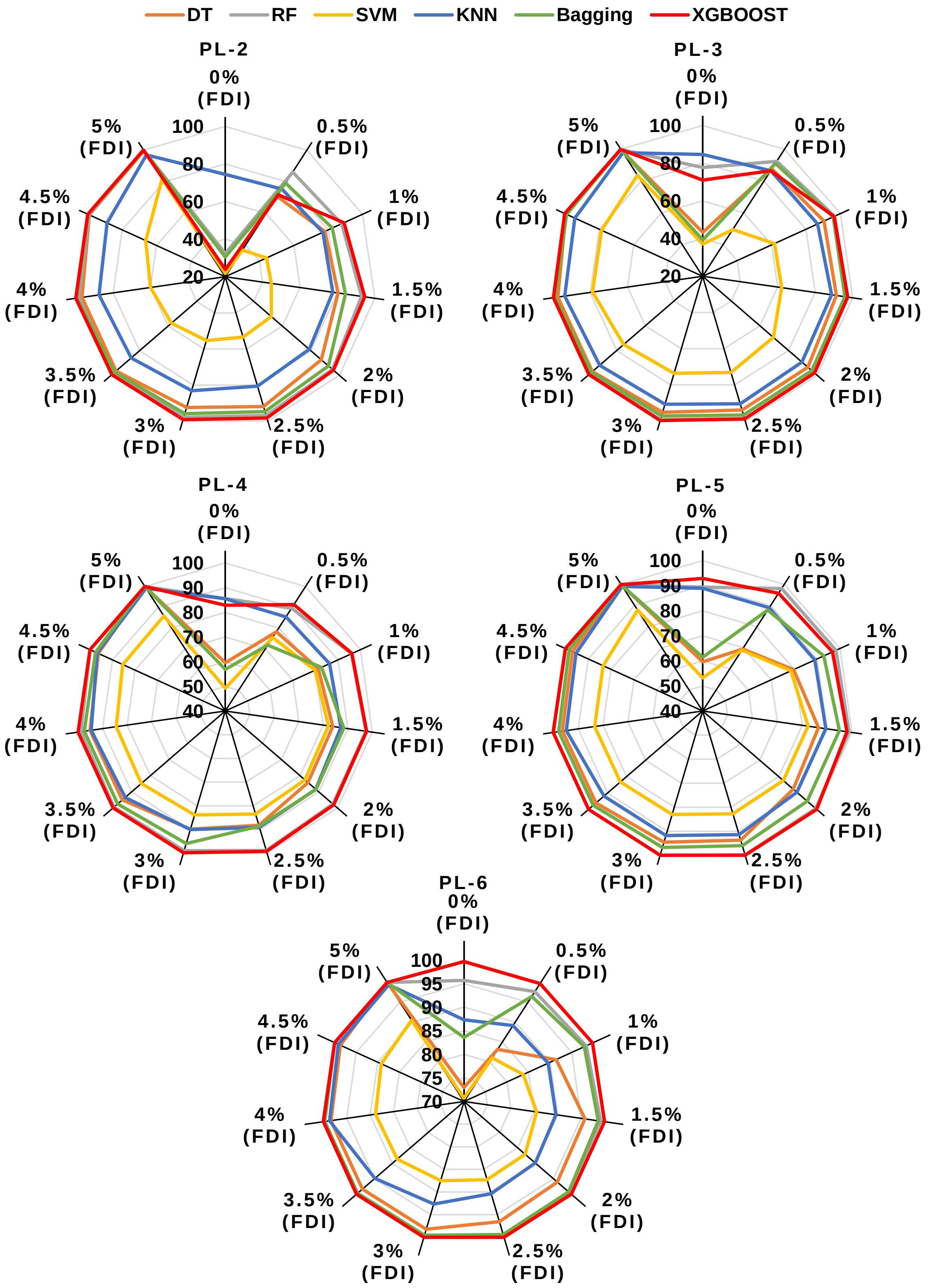


Figure 6.6 Classification accuracy Vs current penetration data availability for various penetration levels.

Moreover, from the detailed analyses of tables and plots in Figure 6.6, it can be deduced that ML models are adapted to fault patterns with the increase in penetration level of RES with very few fault data of current penetration. As per Table 6.3 for test case 10 MW at both the buses (PL-2), the accuracy ranges from 24% to 99% for XGBoost, 32% to 97% for RF, 30% to 96% for Bagging, 31% to 92% for DT, and 21% to 55% for SVM classifier from no-fault data to 3% fault data of total generated fault data of current penetration level in training.

Table 6.3 Penetration level 2: Classification accuracy improvement with increasing fault data availability.

Model	**0**	**0.5 %**	**1.0 %**	**1.5 %**	**2.0 %**	**2.5 %**	**3.0 %**	**3.5 %**	**4.0 %**	**4.5 %**	**5.0 %**
DT	31.1	70.7	78.1	80.4	87.3	91.9	92.4	96.4	96.6	99.1	99.6
RF	32.4	86.2	87.8	92.9	95.1	96.8	97.8	98.1	98.4	99.2	100
SVM	21.8	36.8	44.0	44.6	52.2	53.4	55.3	57.6	59.9	66.2	81.8
KNN	74.4	75.5	76.9	77.4	78.8	80.5	83.0	85.8	87.5	88.8	97
Bagging	30.4	79.2	82.3	84.3	92.3	94.6	95.8	97.4	99.4	99.8	99.8
XGBoost	23.9	71.8	89.2	94.6	95.7	98.2	99.1	99.4	99.9	100	100

Similarly, according to Table 6.4 for test case 20 MW at bus 7 and 10 MW at bus 5 (PL-3), the accuracy ranges from 70% to 99% for XGBoost, 77% to 98% for RF, 39% to 96% for Bagging, 43% to 93% for DT, and 37% to 69% for SVM classifier from no fault data to only 2% fault data in the training of current penetration level. Thus, XGBoost performance is superior to other models. Further, the superior performance exhibited by XGBoost classifier for increasing penetration levels PL-4, PL-5 and PL-6 can be seen from Table 6.5, 6.6 and 6.7 and plots of Figure 6.6. The SVM model performed worse than other models as its performance is highly dependent on feature extraction techniques and in the presented study no feature extraction technique has been utilized.

Table 6.4 Penetration level 3: Classification accuracy improvement with increasing fault data availability.

Model	0	0.5 %	1.0 %	1.5 %	2.0 %	2.5 %	3.0 %	3.5 %	4.0 %	4.5 %	5.0 %
DT	43.2	87.7	90.4	91.4	93.4	93.7	95.1	97	97.2	99.4	99.8
RF	77.6	92.4	96.5	97.1	98.1	98.7	99.1	99.1	99.3	100	100
SVM	37.1	49.4	61.7	62.1	69.4	73	73.6	75.1	78.6	78.7	83.6
KNN	84.5	86.5	86.7	88.8	89.4	90.4	90.6	91.9	93.6	94.2	97.9
Bagging	39.3	91.1	95.8	95.8	96.2	96.6	97.1	97.5	98.6	98.9	100
XGBoost	70.9	86.5	96.2	97.3	98.1	98.8	99.7	99.4	99.5	100	100

Table 6.5 Penetration level 4: Classification accuracy improvement with increasing fault data availability.

Model	0	0.5 %	1.0 %	1.5 %	2.0 %	2.5 %	3.0 %	3.5 %	4.0 %	4.5 %	5.0 %
DT	59.5	78.2	81.3	83.8	84.2	88.2	89.9	94.9	95.1	96.3	100
RF	85.6	89.5	95.9	97.4	97.5	98.8	98.9	99.4	98.8	100	100
SVM	49.3	75.4	79.8	82.1	82.6	83.4	83.8	84.8	84.4	85.4	85.8
KNN	85.5	85.3	86.3	87.2	88.4	89	90	93.4	94.7	96.8	99.1
Bagging	56.9	71.7	82.7	88.4	88.5	88.6	96	97.4	97.5	97.5	100
XGBoost	82.8	91.3	96.2	97.6	97.9	99.2	99.8	99.9	100	100	100

Table 6.6 Penetration level 5: Classification accuracy improvement with increasing fault data availability.

Model	**0**	**0.5 %**	**1.0 %**	**1.5 %**	**2.0 %**	**2.5 %**	**3.0 %**	**3.5 %**	**4.0 %**	**4.5 %**	**5.0 %**
DT	59.5	69.4	79.8	86.5	87.6	93.6	94.6	96.0	96.2	96.8	100
RF	89.3	98.1	98.2	98.8	99.4	99.8	100	100	100	100	100
SVM	53.2	68.8	78.5	82.3	82.3	82.7	83.0	83.3	83.3	83.5	87.8
KNN	87.3	88.9	89	89.4	89.5	91.4	91.7	91.8	94.7	95.3	99.1
Bagging	61.4	87.8	92.9	94.9	94.9	96	96.7	97.4	97.8	98.3	100
XGBoost	92.9	95.8	96.7	97.9	100	100	100	100	100	100	100

Table 6.7 Penetration level 6: Classification accuracy improvement with increasing fault data availability.

Model	**0**	**0.5 %**	**1.0 %**	**1.5 %**	**2.0 %**	**2.5 %**	**3.0 %**	**3.5 %**	**4.0 %**	**4.5 %**	**5.0 %**
DT	72.9	83.1	91.4	95.8	96.1	96.5	98.2	98.3	98.3	98.6	100
RF	95.6	97.7	98.4	99.1	99.4	99.8	100	100	100	100	100
SVM	70.7	81.0	83.8	85.5	87	87.3	87.4	88.6	88.8	89.2	90.4
KNN	88.9	89.1	89.5	89.6	89.9	90.3	92.6	94.8	98.6	99.1	99.4
Bagging	83.6	96.5	97.9	98.6	99.2	99.4	99.5	99.7	99.7	100	100
XGBoost	99.6	99.7	99.8	100	100	100	100	100	100	100	100

6.6.2 Fault Localization

The fault data for the increased RES penetration is obtained gradually over time. Thus, it is mandatory to investigate the model that can quickly perform best with the least data available. Hence, investigation for regression models for fault localization is done for their learning capability testing. When the ML model's performance is tested for fault localization at various penetration levels using fault data of lower penetration levels and available fault data samples of current penetration level, the mean absolute percentage error (MAPE) for fault location estimation on four lines of the studied network obtained is relatively high for initial penetration levels than higher penetration levels. However, MAPE decreases drastically for ensemble methods with fault data inclusions of current penetration levels. This indicates that ensemble methods have quick tendency to learn.

When the ML regression models are tested for fast learning ability on including some fault data of current penetration level with available old penetration level fault data, the MAPE of XGBoost reduces faster than other compared models. The assumed available fault data for old penetration for the proposed study has been taken as 20% of the total fault data generated. From the plots shown in Figures 6.7, 6.8, 6.9, 6.10 and 6.11 displaying MAPE of PL-2, PL-3, PL-4, PL-5 and PL-6 for four lines, XGBR, RFR, RT, and Bagging line plots seem to overlap; however, the XGBR line plot is visibly lower than other compared model line plots. The 0% data represents the no fault data inclusion of current penetration level which is same as adaptability investigation case. Thus, MAPE for 0% data inclusion is highest for all plots.

Moreover, location estimation MAPE for all penetration levels have also been tabulated in Tables 6.8, 6.9, 6.10, 6.11 and 6.12 for PL-2, PL-3, PL-4, PL-5 and PL-6 respectively. XGBoost performed better than other compared models as portrayed from all penetration levels tables. Although Bagging, RFR, and RT MAPE are in the same range as XGBoost, it still has the lowest MAPE compared to other models. The investigation found that XGBoost is the first model to reach the lowest MAPE for most cases, and on reaching 20% of current penetration fault data, it exhibited very low MAPE. Thus, XGBoost exhibited superior performance in adapting and learning new scenario fault data. Hence, XGBoost classifier and regressor can be used for power system fault classification and localization.

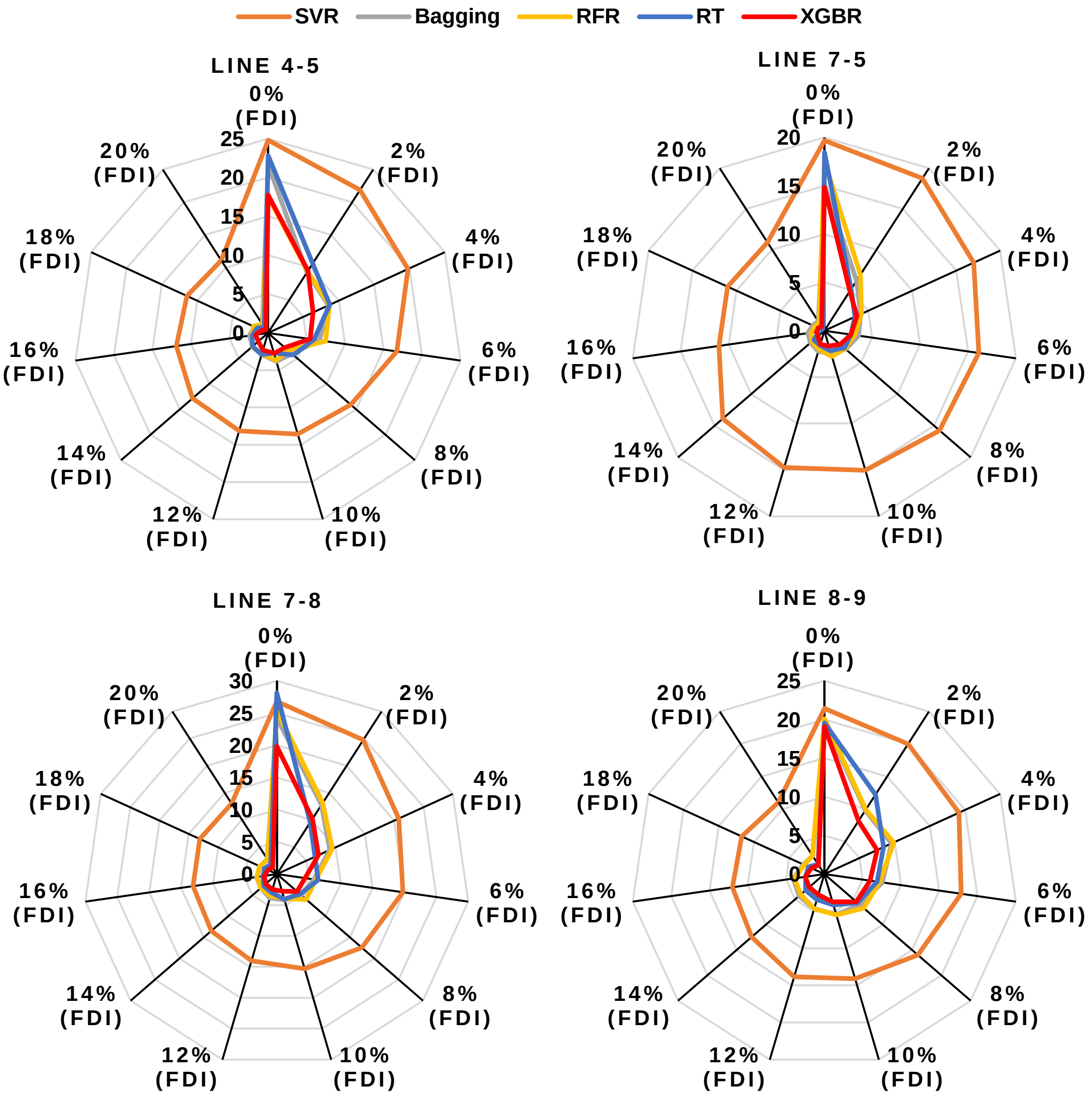


Figure 6.7 PL-2: Radar chart demonstrating MAPE reduction with increasing current penetration fault data inclusion in training.

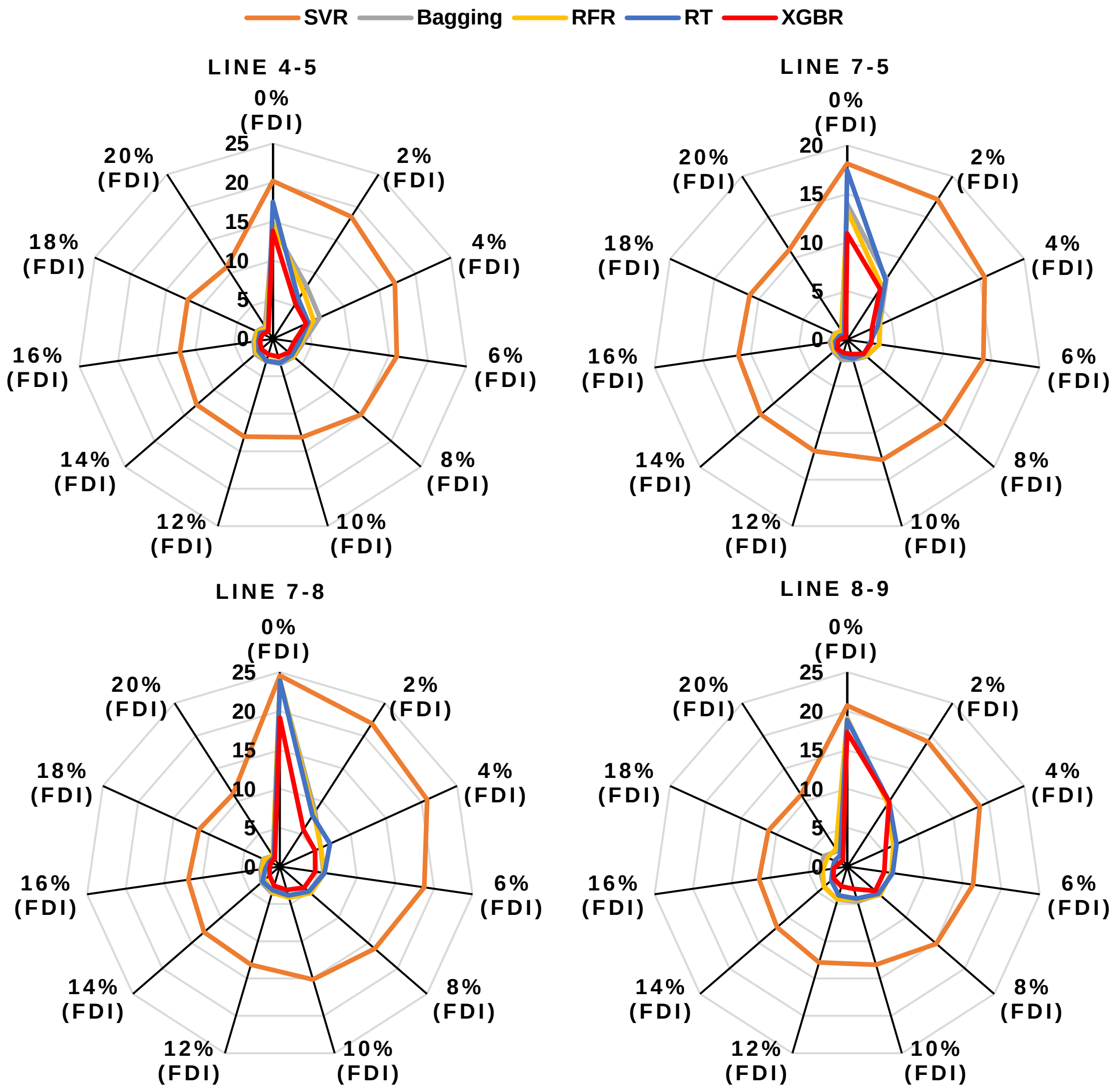


Figure 6.8 PL-3: Radar chart demonstrating MAPE reduction with increasing current penetration fault data inclusion in training.

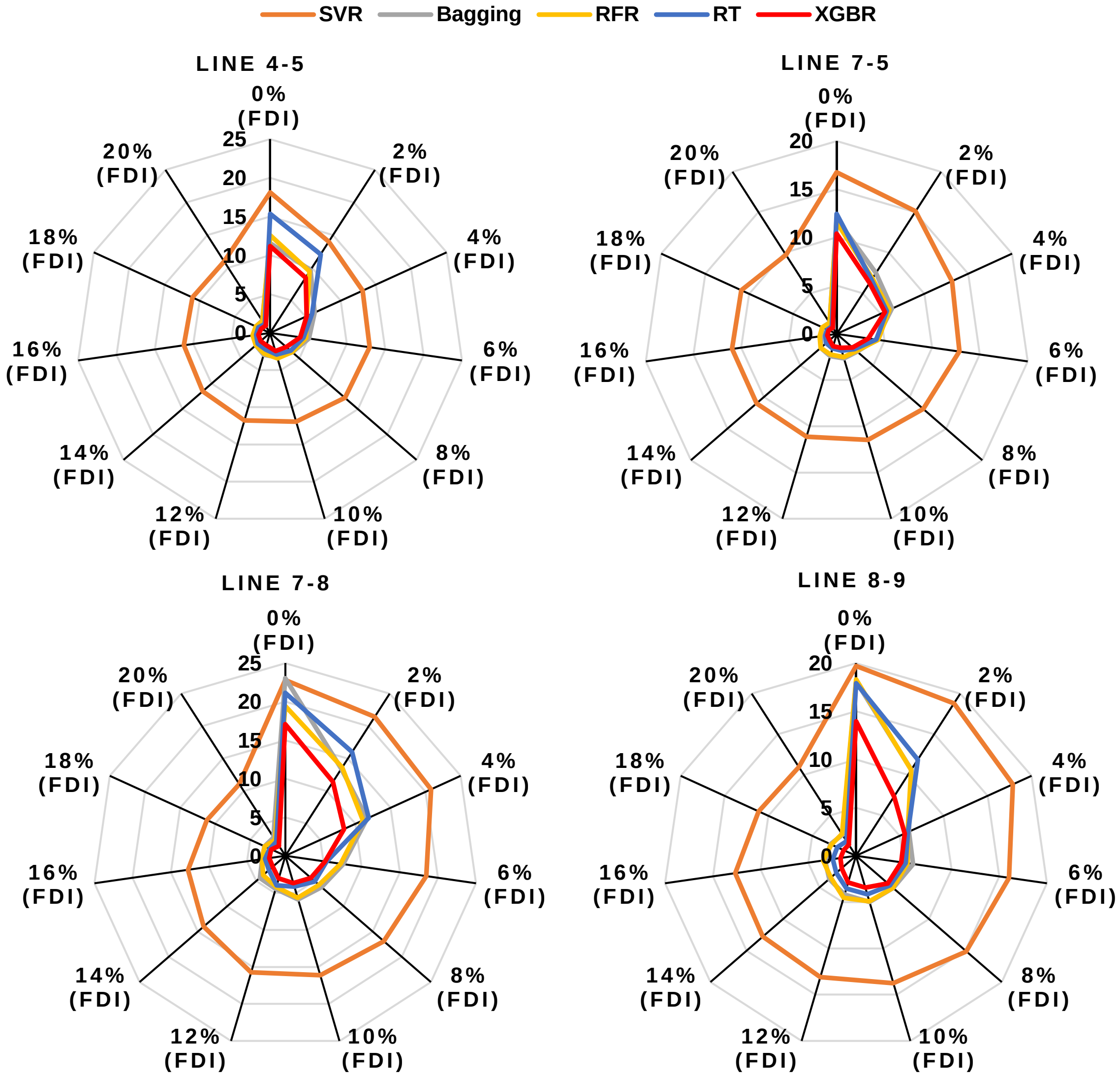


Figure 6.9 PL-4: Radar chart demonstrating MAPE reduction with increasing current penetration fault data inclusion in training.

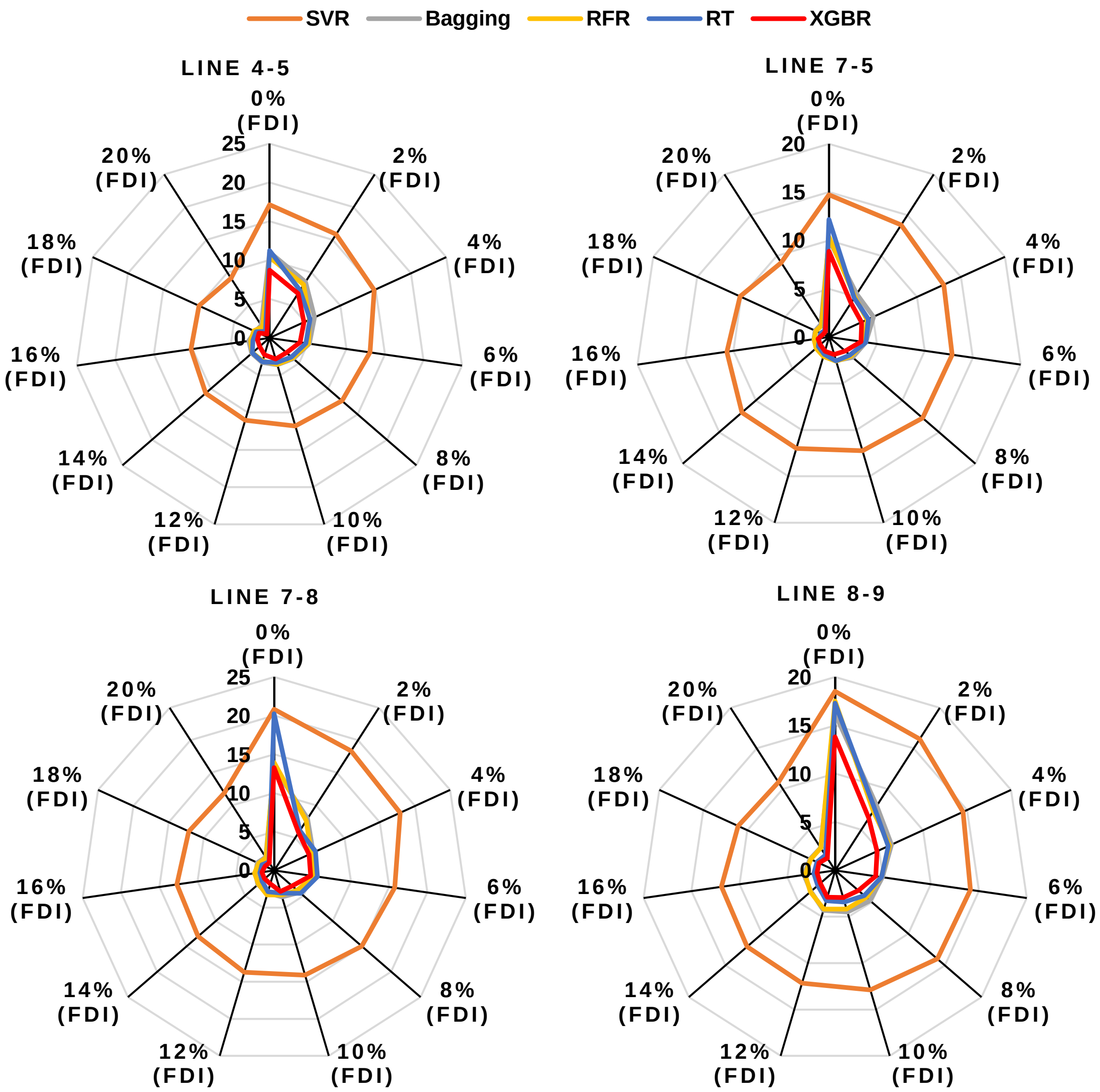


Figure 6.10 PL-5: Radar chart demonstrating MAPE reduction with increasing current penetration fault data inclusion in training.

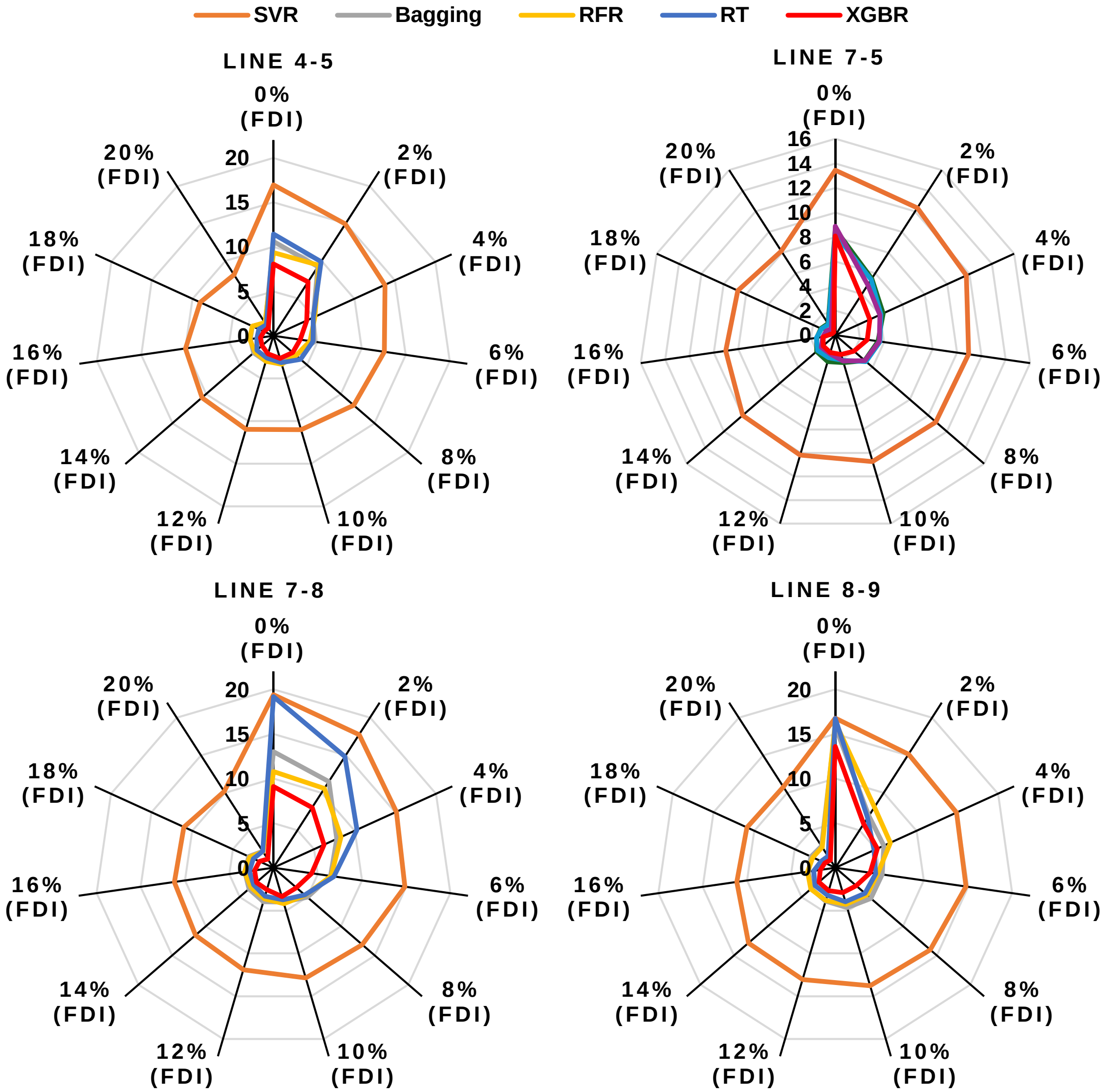


Figure 6.11 PL-6: Radar chart demonstrating MAPE reduction with increasing current penetration fault data inclusion in training.

Table 6.8 PL 2: Reduction in localization MAPE with increasing data availability.

	Model	0	2%	4%	6%	8%	10%	12%	14%	16%	18%	20%
Line 4-5	SVR	24.8	21.8	19.7	16.7	14.1	13.5	13.1	12.8	11.9	11.4	11.1
	Bagging	21.6	9.27	8.52	6.69	4.10	3.82	2.89	2.66	2.41	1.91	1.34
	RFR	17.5	9.42	8.68	7.46	3.48	3.67	2.84	2.73	2.08	2.06	1.28
	RT	22.8	10.6	8.70	5.99	4.33	2.69	2.93	2.64	2.03	1.52	0.87
	XGBR	17.7	9.50	6.38	5.49	2.88	2.75	2.31	1.65	1.59	0.83	0.48
Line 7-5	SVR	19.6	18.7	16.9	16.1	15.7	15.0	14.7	13.8	11.0	10.9	10.8
	Bagging	15.1	6.13	4.14	3.55	2.92	2.45	2.21	1.89	1.71	1.32	1.15
	RFR	17.5	6.88	4.20	3.15	2.82	2.76	2.00	1.70	1.39	1.08	1.01
	RT	18.4	5.11	3.47	2.77	2.72	2.14	1.56	1.42	0.53	0.28	0.27
	XGBR	14.8	4.88	3.70	2.77	2.16	1.64	1.46	0.87	0.80	0.69	0.53
Line 7-8	SVR	26.7	24.7	20.7	19.7	17.4	15.3	14.0	13.4	13.1	13.1	12.9
	Bagging	24.2	12.8	9.04	5.82	5.06	4.16	3.83	3.33	3.16	2.86	2.78
	RFR	25.4	13.1	9.48	6.14	6.01	3.97	3.56	3.31	3.12	2.97	2.66
	RT	28.1	9.55	6.52	6.48	4.74	4.08	2.99	2.73	2.14	2.07	1.79
	XGBR	19.8	10.1	7.09	4.40	4.11	2.76	2.47	2.28	1.97	1.51	1.19
Line 8-9	SVR	21.4	20.0	19.1	17.8	15.9	14.1	13.8	12.4	11.9	11.7	11
	Bagging	19.7	9.81	9.57	7.57	6.04	5.46	4.64	4.32	4.06	3.00	2.82
	RFR	20.1	9.89	9.87	7.23	6.71	5.49	4.66	4.00	3.92	2.87	2.86
	RT	19.5	12.2	8.41	6.90	5.69	4.18	3.42	3.20	2.45	2.27	1.52
	XGBR	19.1	8.15	7.48	5.96	5.44	3.72	2.79	2.58	2.50	1.85	1.38

Table 6.9 PL 3: Reduction in localization MAPE with increasing data availability.

	Model	0	2%	4%	6%	8%	10%	12%	14%	16%	18%	20%
Line 4-5	SVR	20.1	18.5	17.1	15.9	14.8	13.1	13.0	12.9	12.0	11.9	10.8
	Bagging	14.3	7.97	6.55	3.88	3.56	3.36	3.02	3.01	2.50	2.27	1.82
	RFR	14.8	7.24	5.72	3.63	3.51	3.06	2.91	2.78	2.28	2.18	1.54
	RT	17.5	6.02	4.96	3.40	3.22	3.22	2.97	2.43	1.87	1.81	1.16
	XGBR	13.7	5.33	4.65	2.74	2.68	2.40	2.10	1.99	1.67	1.46	1.11
Line 7-5	SVR	18.1	17.2	15.5	14.1	12.9	12.8	11.9	11.7	11.3	11.1	11.0
	Bagging	14.0	7.48	3.66	3.31	2.61	2.16	2.15	1.86	1.81	1.45	1.07
	RFR	13.3	6.56	3.71	3.33	2.53	1.91	1.78	1.72	1.50	1.35	1.09
	RT	17.4	7.30	3.48	2.36	2.28	2.05	1.76	1.44	1.24	0.89	0.59
	XGBR	10.9	6.22	2.82	2.50	2.24	1.56	1.42	1.39	0.95	0.54	0.28
Line 7-8	SVR	24.5	21.8	20.7	18.7	16.1	15.1	13.1	12.8	11.8	11.4	11.1
	Bagging	23.9	8.41	5.77	5.69	4.92	3.95	3.57	3.11	2.49	2.38	1.66
	RFR	24.0	8.10	5.81	5.79	5.22	4.14	3.47	2.58	2.34	2.19	1.66
	RT	23.9	7.78	7.08	5.80	4.90	3.86	3.15	2.95	1.87	1.53	1.44
	XGBR	19.1	5.63	4.96	4.59	4.15	3.13	2.56	1.76	1.34	1.24	1.23
Line 8-9	SVR	20.7	19.1	18.7	16.3	15.1	13.1	12.8	11.9	11.4	11.1	10.9
	Bagging	19.1	9.72	6.71	5.76	5.47	4.67	4.40	3.91	3.54	3.30	2.36
	RFR	19.0	9.62	6.51	5.72	5.60	4.44	4.41	3.96	3.17	2.78	2.70
	RT	18.8	9.97	6.99	5.88	5.33	4.29	3.86	2.72	1.88	1.79	1.72
	XGBR	17.2	9.97	5.42	4.79	4.74	3.00	2.64	2.30	1.75	1.10	1.01

Table 6.10 PL 4: Reduction in localization MAPE with increasing data availability.

	Model	0	2%	4%	6%	8%	10%	12%	14%	16%	18%	20%
Line 4-5	**SVR**	18.1	13.9	13.1	12.9	12.7	11.9	11.7	11.5	11.2	11.0	10.9
	Bagging	11.5	9.63	6.30	5.06	3.80	3.24	2.97	2.46	2.16	2.00	1.88
	RFR	12.5	9.41	5.19	4.66	3.75	3.44	2.80	2.42	2.28	2.03	1.82
	RT	15.3	12.04	5.96	4.36	3.52	2.93	2.21	2.12	1.81	1.66	1.31
	XGBR	11.2	8.46	5.19	3.94	2.77	2.46	1.69	1.64	1.55	1.27	1.09
Line 7-5	**SVR**	16.7	15.1	13.1	12.8	11.8	11.4	11.1	11.0	10.9	10.8	9.74
	Bagging	11.9	7.51	6.31	4.15	2.81	2.64	2.39	2.20	1.63	1.61	1.40
	RFR	11.5	6.82	6.01	4.33	2.73	2.52	2.24	2.22	1.77	1.67	1.37
	RT	12.4	6.68	5.72	4.15	2.42	1.81	1.58	1.47	1.20	1.06	0.94
	XGBR	10.3	6.28	5.48	3.30	2.18	1.51	1.33	1.00	0.97	0.94	0.78
Line 7-8	**SVR**	22.7	21.4	20.7	18.4	16.8	16.1	15.7	13.9	12.7	11.1	11.0
	Bagging	23.0	13.4	11.6	7.60	6.22	5.89	4.45	4.35	3.01	2.98	2.83
	RFR	19.4	13.6	11.0	7.25	5.77	5.66	4.20	3.81	3.05	2.94	2.48
	RT	21.1	16.0	11.8	5.39	5.07	4.16	3.99	2.79	2.63	2.23	2.15
	XGBR	17.0	11.4	8.40	5.17	4.38	3.68	3.05	2.13	2.10	2.01	1.51
Line 8-9	**SVR**	19.7	18.7	17.8	16.0	15.1	13.7	13.1	12.7	12.6	11.1	10.9
	Bagging	18.1	10.6	6.04	5.98	5.09	4.99	4.19	3.66	3.24	2.92	2.60
	RFR	18.3	10.5	5.8	5.40	5.07	4.90	4.52	3.47	3.43	2.82	2.70
	RT	17.9	11.8	5.90	5.22	4.67	4.14	3.53	2.69	2.45	2.13	1.82
	XGBR	13.9	7.29	5.56	4.74	4.35	3.41	2.86	1.96	1.64	1.40	1.39

Table 6.11 PL 5: Reduction in localization MAPE with increasing data availability.

	Model	0	2%	4%	6%	8%	10%	12%	14%	16%	18%	20%
Line 4-5	**SVR**	17.1	15.8	14.7	13.0	12.3	11.8	11.0	10.8	10.1	9.95	9.15
	Bagging	11.0	8.54	6.39	5.18	4.06	3.63	3.32	3.03	2.70	1.97	1.93
	RFR	10.3	8.15	5.64	5.06	3.94	3.56	3.09	2.90	2.44	2.21	1.66
	RT	11.2	7.27	5.74	4.74	3.75	3.31	3.22	2.95	2.14	1.89	0.99
	XGBR	8.70	6.79	4.88	3.99	2.91	2.82	2.29	1.66	1.55	1.50	0.58
Line 7-5	**SVR**	14.7	13.7	13.0	12.8	12.7	12.2	12.0	11.9	10.6	10.1	9.15
	Bagging	10.0	5.30	5.05	3.82	3.21	2.54	2.08	1.72	1.58	1.57	1.56
	RFR	10.3	4.92	4.32	3.72	2.99	2.58	2.09	1.78	1.57	1.52	1.42
	RT	12.1	4.79	4.47	3.83	2.82	2.56	1.85	1.45	1.03	0.95	0.83
	XGBR	8.88	4.17	3.70	3.34	2.20	1.88	1.53	1.19	1.15	0.69	0.69
Line 7-8	**SVR**	20.8	18.3	17.8	15.7	14.9	14.1	13.7	13.0	12.7	12.1	11.9
	Bagging	13.3	7.99	5.53	5.18	4.37	3.58	3.05	2.79	2.44	2.39	2.05
	RFR	13.9	7.67	5.29	5.24	3.91	3.31	3.29	2.69	2.55	2.22	1.95
	RT	20.2	6.14	5.88	5.65	4.52	3.03	2.89	2.08	1.86	1.73	1.29
	XGBR	13.2	5.82	4.93	4.78	3.06	2.82	1.82	1.65	1.62	1.31	1.15
Line 8-9	**SVR**	18.5	16.1	14.5	14.1	13.9	12.8	12.1	12.0	11.9	11.0	10.8
	Bagging	16.2	7.94	6.39	4.93	4.77	4.50	4.32	3.32	3.26	2.85	2.77
	RFR	17.5	7.30	6.18	4.81	4.26	4.18	4.17	3.40	3.12	2.77	2.76
	RT	17.3	7.55	6.05	4.83	3.96	3.41	3.27	2.38	2.24	2.07	1.85
	XGBR	13.7	6.43	4.76	4.24	3.16	2.95	2.88	2.02	1.90	1.83	1.51

Table 6.12 PL 6: Reduction in localization MAPE with increasing data availability.

	Model	0	2%	4%	6%	8%	10%	12%	14%	16%	18%	20%
Line 4-5	**SVR**	16.9	14.9	13.7	12.6	11.9	11.0	10.9	10.6	9.99	9.03	8.15
	Bagging	10.5	9.26	4.91	4.64	3.36	3.06	2.99	2.95	2.63	2.61	1.68
	RFR	9.36	9.44	5.13	4.14	3.36	3.31	3.00	2.77	2.70	2.61	1.75
	RT	11.4	9.90	4.90	4.48	4.02	3.16	2.59	2.45	1.86	1.81	1.48
	XGBR	8.07	7.20	4.12	3.04	2.86	2.64	2.12	1.69	1.48	1.17	1.04
Line 7-5	**SVR**	13.4	12.3	11.7	10.9	10.8	10.7	10.1	9.99	9.03	8.75	8.15
	Bagging	8.64	5.48	4.29	3.49	3.16	2.32	2.25	2.02	1.53	1.30	1.12
	RFR	8.22	5.40	4.08	3.67	3.28	2.13	1.91	1.91	1.57	1.24	1.03
	RT	8.89	4.87	3.98	3.62	3.18	2.19	1.59	1.52	1.06	0.83	0.63
	XGBR	8.09	3.81	3.07	2.61	1.98	1.62	1.47	1.33	1.01	0.29	0.28
Line 7-8	**SVR**	19.4	17.7	15.1	14.8	13.1	12.8	11.9	11.5	11.2	11.0	10.2
	Bagging	13.0	11.4	7.80	6.46	5.01	4.01	3.97	3.63	3.22	3.03	2.20
	RFR	10.8	10.5	8.32	6.53	4.70	4.17	3.61	3.33	3.26	2.93	2.15
	RT	19.1	14.8	10.3	6.85	4.63	3.73	3.30	2.92	2.71	2.47	2.21
	XGBR	9.10	8.02	6.22	4.34	3.40	3.37	2.54	2.52	2.12	1.71	1.18
Line 8-9	**SVR**	16.7	15.1	14.9	14.7	14.0	13.7	13.0	12.8	11.1	10.8	10.7
	Bagging	15.8	6.93	5.95	5.28	5.23	4.66	3.89	3.57	3.05	3.02	2.88
	RFR	16.3	8.09	6.74	4.76	4.64	4.26	3.80	3.65	3.13	2.77	2.75
	RT	16.7	6.67	4.75	4.55	4.35	3.97	3.18	3.00	2.45	1.78	1.55
	XGBR	13.6	5.84	5.15	3.92	3.08	2.91	2.67	2.45	1.71	1.29	1.05

6.7 Result Analysis

XGBoost has emerged as the fastest learning and most adaptable classifier and regressor for power system fault classification and localization for increasing penetration of existing RES. Moreover, its training and testing time is also very low, showing its fast-computing capability and supremacy over other ensemble techniques. The gradient boosting and regularization features of XGBoost enhance its capacity to quickly adapt to new fault data scenarios for fault classification and regression. The intrinsic capability for incremental learning by the XGBoost algorithm supports XGBoost results outcomes. In incremental learning, new models are trained on new data while retaining the knowledge learned from previous data [239]. XGBoost supports incremental learning primarily due to its boosting algorithm and the nature of its implementation. Since XGBoost builds trees sequentially, it's conducive to updating the model with new data. XGBoost is highly optimized for speed and efficiency [212]. The design and implementation of XGBoost make it well suited for incremental learning tasks, allowing it to efficiently adapt to new data while leveraging the knowledge gained from previous iterations [239]. Therefore, it is suitable for classifying and locating transmission line faults for increasing RES penetration levels with gradual fault data availability.

6.8 Chapter Summary

The study presented in this chapter delved into the adaptability and learning competence of tree-based models for fault classification and localization in power systems experiencing increasing penetration of RES. Primarily, the models' adaptability is examined for fault classification with increasing penetration levels, considering that no fault data is available at the increased penetrated level. Models are trained using 20% of the total generated fault data of previous penetration levels to incorporate real power system scenarios of limited data availability. These classifiers are also investigated for learning competence, considering the increasing availability of current penetration level fault data for their training. From the analysis of the results obtained for adaptability testing at different penetration levels, it is found that KNN performed as an average classifier somewhat irrespective of the level of RES penetration. However, all other classifiers performed unreliably at lower penetration levels. In general, the performance of all classifiers showed improvement with an increase in the level of RES penetration. Performances of XGBoost and RF improved significantly with an increase in penetration level. XGBoost exhibited maximum improvement in classification accuracy up to 99.6%. In comparison, RF accuracy reached 95.6% at the maximum considered level of RES penetration without including the current penetration level fault data in their training. The learning competence testing found that the classification performance of considered models showed significant improvement with incremental inclusion of current penetration fault data as per availability gradually over time. Further, from the analysis of learning trends of tested classifiers, XGBoost and RF are found to be the quickest learners giving around 99% classification accuracy with minimum availability of fault data for the current penetration levels. However, at higher penetration levels, XGBoost outperforms other classifiers in the learning competence testing by giving 100% classification accuracy even at the least availability of fault data.

Further, for fault localization, in the adaptability testing of ML models, initially, higher MAPE is observed for all the models. However, MAPE decreases significantly with the increase in RES penetration, especially for XGBoost. Further, when these models are investigated for learning competence, XGBoost performance is found to be superior to other compared models at all penetration levels. Thus, among all tested models, the XGBoost regressor is found to be the best ML model for the localization of power system faults under increasing RES penetration levels. It quickly learns changing fault patterns introduced due to varying power generation from RES, as observed with the least available fault data at different penetration levels. Although XGBoost and RF outperformed other tested models with the gradual availability of fault data, this study suggests that KNN results should be included in the decision-making of the fault type at lower penetration levels for no data availability of current penetration levels.

Chapter 7 CONCLUSIONS AND FUTURE SCOPE

7.1 Introduction

The dissertation presents a comprehensive analysis of machine learning (ML) models for power system fault diagnosis in the context of ongoing integrations of renewable energy sources (RES). Through the exploration of various fault diagnosis techniques, from conventional methodologies to intelligent fault diagnosis paradigms, the thesis has provided valuable insights into the evolution of fault diagnosis in power systems.

By examining existing ML-based approaches for power system fault diagnosis and conducting a detailed analysis of their characteristics and research trends, this work contributes to the understanding of the status of research in the field. Furthermore, the utilization of the IEEE 9 Bus system fault database allows for a rigorous assessment of the suitability of ML models for fault classification and localization.

A key highlight of the thesis is the investigation into the impact of RES integration on ML-based fault diagnosis. By examining the adaptability of various ML models post RES integration and assessing their performance under increasing RES penetration, this study offers valuable insights into the challenges and opportunities associated with the evolving power system landscape.

In conclusion, the findings presented in this thesis contribute to the advancement of ML-based fault diagnosis methodologies in power systems, particularly in the context of RES integration. The identification of future research scopes opens avenues for further exploration and refinement of ML-based approaches to address the evolving challenges in power system fault diagnosis.

7.2 Machine Learning Based Power System Fault Diagnosis Research Advancements and Perspectives

A comprehensive review of ML based power system fault detection, classification, and localization has been presented in this chapter. A brief discussion on conventional and modern power systems, power system monitoring, power system faults, and various fault diagnostic techniques has been covered in the beginning of the thesis. Further, the evolution of fault localization from conventional impedance-based and traveling wave-based methods to ML-based approaches has been outlined. Additionally, a structured framework for addressing problems using ML paradigms, along with various performance metrics and dimensionality reduction techniques, is provided.

The chapter delves into the extensive literature survey covering unsupervised and supervised learning techniques for fault diagnosis. Further, the taxonomical tabulations of research literature facilitate easy reference retrieval and provide insight into current research status and trends. Moreover, the advantages and disadvantages of fault diagnosis techniques have also been discussed. Finally, the chapter identifies gaps within the literature, and suggests several unexplored models and areas that can be utilized for power system fault diagnosis.

7.3 Machine Learning Models Assessment for Conventional Power System Fault Classification and Localization

This chapter extensively tested various baseline and potential models for classifying and localizing conventional transmission network faults. The findings suggest that ML models excel in power system fault classification and localization, given sufficient data availability.

Consequently, this study affirms the feasibility of ML-based automatic fault diagnosis for transmission networks using voltage and current measurements at transmission line buses.

The tests demonstrated that DT, RF, ET, bagging and XGBoost perform satisfactorily for conventional power system fault classification using instantaneous voltage and current measurements from buses. Notably, XGBoost showcased superior performance attributed to its faster execution due to its multithreading parallel computing capabilities. Moreover, BRR emerged as the superior model for fault localization compared to bagging, RF, RT, ET, and SVR models. These results underscore the potential of ML techniques for enhancing fault diagnosis efficiency in power systems, paving the way for more reliable and automated fault classification and localization methodologies. Further, the models have also been explored for fault classification and localization performance, with dimensionality reduction techniques. The investigation revealed that SVM, MLP, and KNN classification performance improved with dimensionality reduction. However, ensemble methods are capable of dealing with high dimensional data effectively without requiring dimensionality reduction. Moreover, fault localization performance degraded significantly with dimensionality reduction techniques.

7.4 Performance and Adaptability Analysis of Machine Learning Models for Transmission Network Fault Classification and Localization with RES Integrations

This chapter presented a two-facet performance analysis of potential ML models for classification and location estimation of transmission network faults with RES integrations. Firstly, the impact analysis of different size RES integrations on the performance of ML models showed that the classification accuracy of all tested classifiers except KNN degrades severely, making them unsuitable for a changed network topology. The fault localization results of all regressors are also significantly impacted by the new RES integration, including KNR. Moreover, the larger the size of the integrated RES, the greater the degradation. KNN's performance alone was within a considerable range for fault classification. Thus, the KNN model can be relied upon for scenarios where system topology has changed, and no-fault data is available for the changed transmission network.

Further, the adaptability testing of these models post RES integration with fault data availability over time revealed that XGBoost, ET, and RF are quick classification learners. Similarly, Bayesian ridge regression outperformed other compared regression models in the adaptability testing, however, it failed to fit the model during the impact analysis. Thus, BRR can learn very quickly the changed fault patterns with data availability.

7.5 Performance Assessment of Machine Learning Models for Fault Classification and Localization Under Increasing RES Penetrations

The study presented in this chapter delved into the adaptability and learning competence investigation of ML models for fault classification and localization in power systems experiencing increasing penetration of existing RES. Initially, ML models' adaptability for fault classification across various RES penetration levels was investigated, considering the absence of fault data from the increased penetration level. Models were trained using fault data from previous penetration levels. This analysis revealed that KNN demonstrated consistent performance as an average classifier, regardless of the RES penetration level. However, other classifiers exhibited unreliable performance at lower penetration levels, with overall improvements observed as RES penetration increased. Notably, XGBoost and RF demonstrated significant enhancements in classification accuracy as RES penetration levels rose, with XGBoost achieving maximum accuracy. The adaptability testing of regressor models for fault localization demonstrated poor performance, giving high MAPE for all models, which moderately decreased with increasing RES penetration, particularly for the XGBoost regressor.

Subsequently, we investigated the learning competence of tested classifiers as fault data availability from the current penetration level increases utilizing incremental learning approach. XGBoost and RF emerged as the quickest learners. XGBoost notably outperformed other classifiers at higher penetration levels, achieving 100% classification accuracy even with limited fault data availability. Furthermore, XGBoost consistently outperformed other tested models across all penetration levels in the learning competence investigation. Thus, among the tested models, XGBoost emerged as the most effective ML model for fault classification and localization in power systems experiencing increasing RES penetration levels. Its ability to quickly adapt to changing fault patterns introduced by varying RES power generation, even with minimal fault data availability, highlights its superiority. While XGBoost and RF showcased superior performance with gradual fault data availability, we suggest considering KNN results in fault type decision-making at lower penetration levels when data from the current penetration level is unavailable.

7.6 Future Scope

While a variety of unsupervised and supervised ML techniques have been applied to fault diagnosis, heterogeneous ensemble techniques remain underexplored. Additionally, more advanced learning approaches, such as transfer learning, commonly used in other engineering applications, warrant deeper exploration in power system fault diagnosis applications. Advanced neural network-based transfer learning is the focus of the proposed study's future work. Further, the investigation proposed can also be investigated for other engineering systems, e.g., electrical motor fault diagnosis undergoing structural changes due to integrating new components or changing motor ratings.

Despite the remarkable achievements of ML techniques in power system fault diagnosis, several key issues remain unresolved and can be examined in future research work. The effectiveness of research work heavily relies on the quality and quantity of available datasets and the characteristics of the test system. Researchers often resort to simulated data due to the lack of realistic power system data. While synthetic data generated from simulations may yield promising results, it fails to fully capture the dynamic behavior of real-world power systems, especially with the integration of renewable energy sources. The prevalent use of synthetic data in research, with the assumption of achieving similar performance in real-world scenarios, overlooks challenges such as imbalanced data. Imbalanced datasets, common in real-time scenarios, are not adequately addressed by researchers, although techniques like undersampling and oversampling could potentially mitigate this issue.

REFERENCES

[1] G. Papaefthymiou, M. Houwing, M. P. C. Weijnen, and L. van der Sluis, “Distributed generation vs bulk power transmission,” in *2008 First International Conference on Infrastructure Systems and Services: Building Networks for a Brighter Future (INFRA)*, 2008, pp. 1–6. doi: 10.1109/INFRA.2008.5439691.

[2] S. Junlakarn and M. Ilic, “Distribution system reliability options and utility liability,” *IEEE Trans. Smart Grid*, vol. 5, no. 5, pp. 2227–2234, 2014, doi: 10.1109/TSG.2014.2316021.

[3] H. A. Tokel, R. Al Halaseh, G. Alirezaei, and R. Mathar, “A new approach for machine learning-based fault detection and classification in power systems,” in *2018 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, Feb. 2018, pp. 1–5. doi: 10.1109/ISGT.2018.8403343.

[4] Y. Wang, C. Chen, J. Wang, and R. Baldick, “Research on Resilience of Power Systems Under Natural Disasters—A Review,” *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1604–1613, Mar. 2016, doi: 10.1109/TPWRS.2015.2429656.

[5] H. Panahi, M. Sanaye-Pasand, and M. Davarpanah, “Three-Terminal Lines Fault Location Using Two Main Terminals Data in the Presence of Renewable Energy Sources,” *IEEE Trans. Smart Grid*, vol. 14, no. 3, pp. 2085–2095, May 2023, doi: 10.1109/TSG.2022.3216908.

[6] R. Vaish, U. D. Dwivedi, S. Tewari, and S. M. Tripathi, “Machine learning applications in power system fault diagnosis: Research advancements and perspectives,” *Eng. Appl. Artif. Intell.*, vol. 106, no. October, p. 104504, Nov. 2021, doi: 10.1016/j.engappai.2021.104504.

[7] N. Peng, L. Zhou, R. Liang, X. Xue, G. Piliposyan, and Y. Hu, “Fault location on double–circuit transmission lines by phase correction of fault recorder signals without accurate time synchronization,” *Electr. Power Syst. Res.*, vol. 181, no. July 2019, p. 106198, 2020, doi: 10.1016/j.epsr.2020.106198.

[8] F. Aminifar, S. Teimourzadeh, A. Shahsavari, M. Savaghebi, and M. S. Golsorkhi, “Machine learning for protection of distribution networks and power electronics-interfaced systems,” *Electr. J.*, vol. 34, no. 1, p. 106886, 2021, doi: 10.1016/j.tej.2020.106886.

[9] H. Zhan *et al.*, “Relay Protection Coordination Integrated Optimal Placement and Sizing of Distributed Generation Sources in Distribution Networks,” *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 55–65, Jan. 2016, doi: 10.1109/TSG.2015.2420667.

[10] K. Yoon, D. Choi, S. H. Lee, and J.-W. Park, “Optimal Placement Algorithm of Multiple DGs Based on Model-Free Lyapunov Exponent Estimation,” *IEEE Access*, vol. 8, pp. 135416–135425, 2020, doi: 10.1109/ACCESS.2020.3011162.

[11] C. Galvez and A. Abur, “Fault Location in Power Networks Using a Sparse Set of Digital Fault Recorders,” *IEEE Trans. Smart Grid*, vol. 13, no. 5, pp. 3468–3480,

Sep. 2022, doi: 10.1109/TSG.2022.3168904.

[12] D. S. Nair and R. T, “Investigation on Impact of Solar PV penetration on the Operation of Protective Relays in a Distribution System using Python,” in *2021 IEEE Conference on Technologies for Sustainability (SusTech)*, Apr. 2021, pp. 1–8. doi: 10.1109/SusTech51236.2021.9467474.

[13] M. A. Aftab, S. M. S. Hussain, I. Ali, and T. S. Ustun, “Dynamic protection of power systems with high penetration of renewables: A review of the traveling wave based fault location techniques,” *Int. J. Electr. Power Energy Syst.*, vol. 114, p. 105410, Jan. 2020, doi: 10.1016/j.ijepes.2019.105410.

[14] T. Bi, B. Yang, K. Jia, L. Zheng, Q. Liu, and Q. Yang, “Review on Renewable Energy Source Fault Characteristics Analysis,” *CSEE J. Power Energy Syst.*, vol. 8, no. 4, pp. 963–972, 2022, doi: 10.17775/CSEEJPES.2021.06890.

[15] KhareSaxena A, Saxena S, and Sudhakar K, “Solar energy policy of India: An overview,” *CSEE J. Power Energy Syst.*, 2020, doi: 10.17775/CSEEJPES.2020.03080.

[16] Z. Lin *et al.*, “One-Class Classifier Based Fault Detection in Distribution Systems with Varying Penetration Levels of Distributed Energy Resources,” *IEEE Access*, vol. 8, pp. 130023–130035, 2020, doi: 10.1109/ACCESS.2020.3009385.

[17] P. Rai, N. D. Londhe, and R. Raj, “Fault classification in power system distribution network integrated with distributed generators using CNN,” *Electr. Power Syst. Res.*, vol. 192, no. September, p. 106914, Mar. 2021, doi: 10.1016/j.epsr.2020.106914.

[18] V. A. Papaspiliotopoulos, G. N. Korres, V. A. Kleftakis, and N. D. Hatziargyriou, “Hardware-In-the-Loop Design and Optimal Setting of Adaptive Protection Schemes for Distribution Systems With Distributed Generation,” *IEEE Trans. Power Deliv.*, vol. 32, no. 1, pp. 393–400, Feb. 2017, doi: 10.1109/TPWRD.2015.2509784.

[19] R. Bansal, *Power System Protection in Smart Grid Environment*. Boca Raton : Taylor & Francis, a CRC title, part of the Taylor & Francis imprint, a member of the Taylor & Francis Group, the academic division of T&F Informa, plc, 2019.: CRC Press, 2019. doi: 10.1201/9780429401756.

[20] A. Silos-Sanchez, R. Villafafila-Robles, and P. Lloret-Gallego, “Novel fault location algorithm for meshed distribution networks with DERs,” *Electr. Power Syst. Res.*, vol. 181, no. July 2019, p. 106182, Apr. 2020, doi: 10.1016/j.epsr.2019.106182.

[21] V. Telukunta, J. Pradhan, A. Agrawal, M. Singh, and S. G. Srivani, “Protection challenges under bulk penetration of renewable energy resources in power systems: A review,” *CSEE J. Power Energy Syst.*, vol. 3, no. 4, pp. 365–379, Dec. 2017, doi: 10.17775/CSEEJPES.2017.00030.

[22] D. Tzelepis *et al.*, “Voltage and Current Measuring Technologies for High Voltage Direct Current Supergrids: A Technology Review Identifying the Options for

Protection, Fault Location and Automation Applications,” *IEEE Access*, vol. 8, pp. 203398–203428, 2020, doi: 10.1109/ACCESS.2020.3035905.

[23] 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.*, vol. 32, no. 3, pp. 1450–1459, Jun. 2017, doi: 10.1109/TPWRD.2016.2615886.

[24] S. Kulkarni, S. Member, S. Santoso, S. Member, T. A. Short, and S. Member, “Incipient Fault Location Algorithm for Underground Cables,” *IEEE Trans. Smart Grid*, vol. 5, no. 3, pp. 1165–1174, 2014.

[25] Z. M. Radojević, C. H. Kim, M. Popov, G. Preston, and V. Terzija, “New Approach for Fault Location on Transmission Lines Not Requiring Line Parameters,” *Int. Conf. Power Syst. Transients*, 2009, [Online]. Available: http://www.ipst.org/TechPapers/2009/IPST09Papers.htm

[26] J. Izykowski, “Power System Faults,” *Renew. Energy Syst.*, pp. 239–325, 2014, [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124104235000080

[27] E. De Santis, A. Rizzi, and A. Sadeghian, “A learning intelligent System for classification and characterization of localized faults in Smart Grids,” in *2017 IEEE Congress on Evolutionary Computation (CEC)*, Jun. 2017, pp. 2669–2676. doi: 10.1109/CEC.2017.7969631.

[28] S.-J. Huang and H.-H. Wan, “A Method to Enhance Ground-Fault Computation,” *IEEE Trans. Power Syst.*, vol. 25, no. 2, pp. 1190–1191, May 2010, doi: 10.1109/TPWRS.2009.2036331.

[29] Y. Q. Chen, O. Fink, and G. Sansavini, “Combined Fault Location and Classification for Power Transmission Lines Fault Diagnosis With Integrated Feature Extraction,” *IEEE Trans. Ind. Electron.*, vol. 65, no. 1, pp. 561–569, Jan. 2018, doi: 10.1109/TIE.2017.2721922.

[30] M.-F. Guo, X.-D. Zeng, D.-Y. Chen, and N.-C. Yang, “Deep-Learning-Based Earth Fault Detection Using Continuous Wavelet Transform and Convolutional Neural Network in Resonant Grounding Distribution Systems,” *IEEE Sens. J.*, vol. 18, no. 3, pp. 1291–1300, Feb. 2018, doi: 10.1109/JSEN.2017.2776238.

[31] R.-A. Tirnovan and M. Cristea, “Advanced techniques for fault detection and classification in electrical power transmission systems: An overview,” in *2019 8th International Conference on Modern Power Systems (MPS)*, May 2019, pp. 1–10. doi: 10.1109/MPS.2019.8759695.

[32] I. S. Baxevanos and D. P. Labridis, “Software Agents Situated in Primary Distribution Networks: A Cooperative System for Fault and Power Restoration Management,” *IEEE Trans. Power Deliv.*, vol. 22, no. 4, pp. 2378–2385, Oct. 2007, doi: 10.1109/TPWRD.2007.905463.

[33] F. Shen, Q. Wu, and Y. Xue, “Review of Service Restoration for Distribution

Networks,” *J. Mod. Power Syst. Clean Energy*, vol. 8, no. 1, pp. 1–14, 2020, doi: 10.35833/MPCE.2018.000782.

[34] K. Chen, J. Hu, Y. Zhang, Z. Yu, and J. He, “Fault Location in Power Distribution Systems via Deep Graph Convolutional Networks,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 1, pp. 119–131, Jan. 2020, doi: 10.1109/JSAC.2019.2951964.

[35] F. Zhang, Q. Liu, Y. Liu, N. Tong, S. Chen, and C. Zhang, “Novel Fault Location Method for Power Systems Based on Attention Mechanism and Double Structure GRU Neural Network,” *IEEE Access*, vol. 8, pp. 75237–75248, 2020, doi: 10.1109/ACCESS.2020.2988909.

[36] J. Liang, T. Jing, H. Niu, and J. Wang, “Two-Terminal Fault Location Method of Distribution Network Based on Adaptive Convolution Neural Network,” *IEEE Access*, vol. 8, pp. 54035–54043, 2020, doi: 10.1109/ACCESS.2020.2980573.

[37] O. A. Gashteroodkhani, M. Majidi, M. Etezadi-Amoli, A. F. Nematollahi, and B. Vahidi, “A hybrid SVM-TT transform-based method for fault location in hybrid transmission lines with underground cables,” *Electr. Power Syst. Res.*, vol. 170, no. January, pp. 205–214, 2019, doi: 10.1016/j.epsr.2019.01.023.

[38] M. Sahani and P. K. Dash, “Fault location estimation for series-compensated double-circuit transmission line using EWT and weighted RVFLN,” *Eng. Appl. Artif. Intell.*, vol. 88, no. December 2018, p. 103336, Feb. 2020, doi: 10.1016/j.engappai.2019.103336.

[39] M. Shafiullah, M. A. Abido, and Z. Al-Hamouz, “Wavelet-based extreme learning machine for distribution grid fault location,” *IET Gener. Transm. Distrib.*, vol. 11, no. 17, pp. 4256–4263, Nov. 2017, doi: 10.1049/iet-gtd.2017.0656.

[40] Yixin Cai, Mo-Yuen Chow, Wenbin Lu, and Lexin Li, “Evaluation of distribution fault diagnosis algorithms using ROC curves,” in *IEEE PES General Meeting*, Jul. 2010, pp. 1–6. doi: 10.1109/PES.2010.5588154.

[41] P. P. K. Chan, J. Zhu, Z.-W. Qiu, W. W. Y. Ng, and D. S. Yeung, “Comparision of different classifiers in fault detection in microgrid,” in *2011 International Conference on Machine Learning and Cybernetics*, Jul. 2011, no. July, pp. 1210–1213. doi: 10.1109/ICMLC.2011.6016932.

[42] A. Dasgupta, S. Debnath, and A. Das, “Transmission line fault detection and classification using cross-correlation and k-nearest neighbor,” *Int. J. Knowledge-based Intell. Eng. Syst.*, vol. 19, no. 3, pp. 183–189, Oct. 2015, doi: 10.3233/KES-150320.

[43] P. K. Mishra, A. Yadav, and M. Pazoki, “A Novel Fault Classification Scheme for Series Capacitor Compensated Transmission Line Based on Bagged Tree Ensemble Classifier,” *IEEE Access*, vol. 6, pp. 27373–27382, May 2018, doi: 10.1109/ACCESS.2018.2836401.

[44] Z. El Mrabet, N. Sugunaraj, P. Ranganathan, and S. Abhyankar, “Random Forest Regressor-Based Approach for Detecting Fault Location and Duration in Power

Systems," *Sensors*, vol. 22, no. 2, p. 458, Jan. 2022, doi: 10.3390/s22020458.

[45] A. de Souza Gomes, M. A. Costa, T. G. A. de Faria, and W. M. Caminhas, "Detection and Classification of Faults in Power Transmission Lines Using Functional Analysis and Computational Intelligence," *IEEE Trans. Power Deliv.*, vol. 28, no. 3, pp. 1402–1413, Jul. 2013, doi: 10.1109/TPWRD.2013.2251752.

[46] M. Mishra and P. K. Rout, "Detection and classification of micro-grid faults based on HHT and machine learning techniques," *IET Gener. Transm. Distrib.*, vol. 12, no. 2, pp. 388–397, 2018, doi: 10.1049/iet-gtd.2017.0502.

[47] T. S. Abdelgayed, W. G. Morsi, and T. S. Sidhu, "Fault Detection and Classification Based on Co-training of Semisupervised Machine Learning," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1595–1605, Feb. 2018, doi: 10.1109/TIE.2017.2726961.

[48] R. Godse and S. Bhat, "Mathematical Morphology-Based Feature-Extraction Technique for Detection and Classification of Faults on Power Transmission Line," *IEEE Access*, vol. 8, pp. 38459–38471, 2020, doi: 10.1109/ACCESS.2020.2975431.

[49] S. R. Samantaray and P. K. Dash, "High impedance fault detection in distribution feeders using extended kalman filter and support vector machine," *Eur. Trans. Electr. Power*, vol. 20, no. 3, pp. 382–393, Apr. 2010, doi: 10.1002/etep.321.

[50] M. Sarwar, F. Mehmood, M. Abid, A. Q. Khan, S. T. Gul, and A. S. Khan, "High impedance fault detection and isolation in power distribution networks using support vector machines," *J. King Saud Univ. - Eng. Sci.*, vol. 32, no. 8, pp. 524–535, Dec. 2020, doi: 10.1016/j.jksues.2019.07.001.

[51] J. C. Arouche Freire, A. R. Garcez Castro, M. S. Homci, B. S. Meiguins, and J. M. De Morais, "Transmission Line Fault Classification Using Hidden Markov Models," *IEEE Access*, vol. 7, pp. 113499–113510, 2019, doi: 10.1109/ACCESS.2019.2934938.

[52] M.-F. Guo, N.-C. Yang, and W.-F. Chen, "Deep-Learning-Based Fault Classification Using Hilbert–Huang Transform and Convolutional Neural Network in Power Distribution Systems," *IEEE Sens. J.*, vol. 19, no. 16, pp. 6905–6913, Aug. 2019, doi: 10.1109/JSEN.2019.2913006.

[53] H. Lala and S. Karmakar, "Detection and Experimental Validation of High Impedance Arc Fault in Distribution System Using Empirical Mode Decomposition," *IEEE Syst. J.*, vol. 14, no. 3, pp. 3494–3505, Sep. 2020, doi: 10.1109/JSYST.2020.2969966.

[54] G. W. Chang, Y.-H. Hong, and G.-Y. Li, "A Hybrid Intelligent Approach for Classification of Incipient Faults in Transmission Network," *IEEE Trans. Power Deliv.*, vol. 34, no. 4, pp. 1785–1794, Aug. 2019, doi: 10.1109/TPWRD.2019.2924840.

[55] M. Sahani and P. K. Dash, "Fault location estimation for series-compensated double-circuit transmission line using EWT and weighted RVFLN," *Eng. Appl.*

Artif. Intell., vol. 88, no. July 2018, p. 103336, 2020, doi: 10.1016/j.engappai.2019.103336.

[56] B. Patel, “A new FDOST entropy based intelligent digital relaying for detection, classification and localization of faults on the hybrid transmission line,” *Electr. Power Syst. Res.*, vol. 157, no. April, pp. 39–47, Apr. 2018, doi: 10.1016/j.epsr.2017.12.002.

[57] X. G. Magagula, Y. Hamam, J. A. Jordaan, and A. A. Yusuff, “A fault classification and localization method in a power distribution network,” in *2017 IEEE AFRICON*, Sep. 2017, pp. 1337–1343. doi: 10.1109/AFRCON.2017.8095676.

[58] K. Moloi, J. A. Jordaan, and Y. Hamam, “High Impedance Fault Classification and Localization Method for Power Distribution Network,” in *2018 IEEE PES/IAS PowerAfrica*, Jun. 2018, pp. 84–89. doi: 10.1109/PowerAfrica.2018.8520972.

[59] K. Moloi and A. A. Yusuff, “A Support Vector Machine Based Fault Diagnostic Technique In Power Distribution Networks,” in *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, Jan. 2019, pp. 229–234. doi: 10.1109/RoboMech.2019.8704768.

[60] R. Vaish and U. D. Dwivedi, “Comparative Study of Machine Learning Models for Power System Fault Identification and Localization,” in *2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)*, Feb. 2022, pp. 110–115. doi: 10.1109/ICRTCST54752.2022.9781861.

[61] A. N. Hasan, P. S. P. Eboule, and B. Twala, “The use of machine learning techniques to classify power transmission line fault types and locations,” *Proc. - 2017 Int. Conf. Optim. Electr. Electron. Equipment, OPTIM 2017 2017 Intl Aegean Conf. Electr. Mach. Power Electron. ACEMP 2017*, pp. 221–226, 2017, doi: 10.1109/OPTIM.2017.7974974.

[62] N. Sapountzoglou, J. Lago, and B. Raison, “Fault diagnosis in low voltage smart distribution grids using gradient boosting trees,” *Electr. Power Syst. Res.*, vol. 182, no. February, p. 106254, May 2020, doi: 10.1016/j.epsr.2020.106254.

[63] Y. Wang, Q. Cui, Y. Weng, D. Li, and W. Li, “Learning picturized and time-series data for fault location with renewable energy sources,” *Int. J. Electr. Power Energy Syst.*, vol. 147, no. December 2022, p. 108853, May 2023, doi: 10.1016/j.ijepes.2022.108853.

[64] H. Shah, N. Chothani, and J. Chakravorty, “Fault Detection and Classification in Interconnected System with Wind Generation Using ANN and SVM,” *Power Eng. Electr. Eng.*, vol. 20, no. 3, pp. 225–239, 2022, doi: 10.15598/aeee.v20i3.4483.

[65] Z. Lin *et al.*, “Data-Driven Fault Localization in Distribution Systems with Distributed Energy Resources,” *Energies*, vol. 13, no. 1, p. 275, Jan. 2020, doi: 10.3390/en13010275.

[66] A. Srivastava and S. K. Parida, “Recognition of Fault Location and Type in a

Medium Voltage System with Distributed Generation using Machine Learning Approach," *2019 20th Int. Conf. Intell. Syst. Appl. to Power Syst. ISAP 2019*, 2019, doi: 10.1109/ISAP48318.2019.9065994.

[67] H. M. Mesbah Maruf, F. Muller, M. S. Hassan, and B. Chowdhury, "Locating Faults in Distribution Systems in the Presence of Distributed Generation using Machine Learning Techniques," in *2018 9th IEEE International Symposium on Power Electronics for Distributed Generation Systems (PEDG)*, Jun. 2018, pp. 1–6. doi: 10.1109/PEDG.2018.8447728.

[68] M. T. Haque and A. M. Kashtiban, "Application of neural networks in power systems; A review," *Proc. - Wec 05 Fourth World Enformatika Conf.*, vol. 6, no. 6, pp. 53–57, 2005, doi: https://doi.org/10.5281/zenodo.1081503.

[69] M. Prakash, S. Pradhan, and S. Roy, "Soft computing techniques for fault detection in power distribution systems: A review," *Proceeding IEEE Int. Conf. Green Comput. Commun. Electr. Eng. ICGCCEE 2014*, 2014, doi: 10.1109/ICGCCEE.2014.6922376.

[70] S. A. Aleem, N. Shahid, and I. H. Naqvi, "Methodologies in power systems fault detection and diagnosis," *Energy Syst.*, vol. 6, no. 1, pp. 85–108, Mar. 2015, doi: 10.1007/s12667-014-0129-1.

[71] V. H. Ferreira *et al.*, "A survey on intelligent system application to fault diagnosis in electric power system transmission lines," *Electr. Power Syst. Res.*, vol. 136, pp. 135–153, Jul. 2016, doi: 10.1016/j.epsr.2016.02.002.

[72] K. Chen, C. Huang, and J. He, "Fault detection, classification and location for transmission lines and distribution systems: A review on the methods," *High Volt.*, vol. 1, no. 1, pp. 25–33, 2016, doi: 10.1049/hve.2016.0005.

[73] S. S. Gururajapathy, H. Mokhlis, and H. A. Illias, "Fault location and detection techniques in power distribution systems with distributed generation: A review," *Renew. Sustain. Energy Rev.*, vol. 74, no. February 2016, pp. 949–958, Jul. 2017, doi: 10.1016/j.rser.2017.03.021.

[74] D. P. Mishra and P. Ray, "Fault detection, location and classification of a transmission line," *Neural Comput. Appl.*, vol. 30, no. 5, pp. 1377–1424, Sep. 2018, doi: 10.1007/s00521-017-3295-y.

[75] A. Prasad, J. Belwin Edward, and K. Ravi, "A review on fault classification methodologies in power transmission systems: Part-II," *J. Electr. Syst. Inf. Technol.*, vol. 5, no. 1, pp. 61–67, May 2018, doi: 10.1016/j.jesit.2016.10.003.

[76] A. Raza, A. Benrabah, T. Alquthami, and M. Akmal, "A review of fault diagnosing methods in power transmission systems," *Appl. Sci.*, vol. 10, no. 4, 2020, doi: 10.3390/app10041312.

[77] N. Shahid, S. A. Aleem, I. H. Naqvi, and N. Zaffar, "Support Vector Machine based fault detection & classification in smart grids," *2012 IEEE Globecom Work. GC Wkshps 2012*, no. Ll, pp. 1526–1531, 2012, doi:

10.1109/GLOCOMW.2012.6477812.

[78] J. Liu, W. Fang, X. Zhang, and C. Yang, “An Improved Photovoltaic Power Forecasting Model With the Assistance of Aerosol Index Data,” *IEEE Trans. Sustain. Energy*, vol. 6, no. 2, pp. 434–442, 2015, doi: 10.1109/TSTE.2014.2381224.

[79] W. Khan, S. Walker, and W. Zeiler, “Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach,” *Energy*, vol. 240, p. 122812, Feb. 2022, doi: 10.1016/j.energy.2021.122812.

[80] R. B. Otto, F. P. Silva, M. B. Do Carmo, A. B. Piardi, and R. A. Ramos, “Simulation technologies applicable to microgrids,” *Renew. Energy Power Qual. J.*, vol. 18, no. 18, pp. 584–589, 2020, doi: 10.24084/repqj18.439.

[81] P. Pan, R. K. Mandal, and M. M. Rahman Redoy Akanda, “Fault Classification with Convolutional Neural Networks for Microgrid Systems,” *Int. Trans. Electr. Energy Syst.*, vol. 2022, pp. 1–21, 2022, doi: 10.1155/2022/8431450.

[82] 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*, vol. 5, no. 6, pp. 2947–2956, Nov. 2014, doi: 10.1109/TSG.2014.2330624.

[83] A. Jain, T. C. Archana, and M. B. K. Sahoo, “A Methodology for Fault Detection and Classification Using PMU Measurements,” in *2018 20th National Power Systems Conference (NPSC)*, Dec. 2018, pp. 1–6. doi: 10.1109/NPSC.2018.8771757.

[84] J. J. Q. Yu, Y. Hou, A. Y. S. Lam, and V. O. K. Li, “Intelligent Fault Detection Scheme for Microgrids With Wavelet-Based Deep Neural Networks,” *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1694–1703, Mar. 2019, doi: 10.1109/TSG.2017.2776310.

[85] A. Rangel-Damian, E. Melgoza-Vazquez, and H. F. Ruiz-Paredes, “Application of fault location methods in distribution circuits with SCADA,” *2017 IEEE Int. Autumn Meet. Power, Electron. Comput. ROPEC 2017*, vol. 2018-Janua, no. Ropec, pp. 1–6, 2018, doi: 10.1109/ROPEC.2017.8261652.

[86] C. Wang, C. X. Dou, X. Bin Li, and Q. Q. Jia, “A WAMS/PMU-based fault location technique,” *Electr. Power Syst. Res.*, vol. 77, no. 8, pp. 936–945, 2007, doi: 10.1016/j.epsr.2006.08.007.

[87] Y. Zhang and Z. Wang, “MAP fault localization based on wide area synchronous phasor measurement information,” *Int. J. Emerg. Electr. Power Syst.*, vol. 16, no. 1, pp. 75–81, 2015, doi: 10.1515/ijeeps-2014-0107.

[88] W. Fan and Y. Liao, “Wide area measurements based fault detection and location method for transmission lines,” *Prot. Control Mod. Power Syst.*, vol. 4, no. 1, 2019, doi: 10.1186/s41601-019-0121-9.

[89] O. A. Alimi, K. Ouahada, and A. M. Abu-Mahfouz, “A Review of Machine

Learning Approaches to Power System Security and Stability," *IEEE Access*, vol. 8, pp. 113512–113531, 2020, doi: 10.1109/ACCESS.2020.3003568.

[90] S. H. Asman, N. F. Ab Aziz, M. Z. A. Abd Kadir, and U. A. U. Amirulddin, "Fault signature analysis based on digital fault recorder in malaysia overhead line system," *PECon 2020 - 2020 IEEE Int. Conf. Power Energy*, pp. 188–193, 2020, doi: 10.1109/PECon48942.2020.9314417.

[91] M. Chantier, P. Pogliano, A. Aldea, G. Tornielli, T. Wyatt, and A. Jolley, "The use of fault-recorder data for diagnosing timing and other related faults in electricity transmission networks," *IEEE Trans. Power Syst.*, vol. 15, no. 4, pp. 1388–1393, 2000, doi: 10.1109/59.898117.

[92] K. M. Silva, B. A. Souza, and N. S. D. Brito, "Fault Detection and Classification in Transmission Lines Based on Wavelet Transform and ANN," *IEEE Trans. Power Deliv.*, vol. 21, no. 4, pp. 2058–2063, Oct. 2006, doi: 10.1109/TPWRD.2006.876659.

[93] H. Mirshekali, R. Dashti, A. Keshavarz, A. J. Torabi, and H. R. Shaker, "A Novel Fault Location Methodology for Smart Distribution Networks," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1277–1288, Mar. 2021, doi: 10.1109/TSG.2020.3031400.

[94] S. Das, S. Santoso, A. Gaikwad, and M. Patel, "Impedance-based fault location in transmission networks: theory and application," *IEEE Access*, vol. 2, no. January, pp. 537–557, 2014, doi: 10.1109/ACCESS.2014.2323353.

[95] The Mathworks and MATLAB, "Introducing Machine Learning," *Perspect. Ontol. Learn.*, vol. 18, no. January 2014, pp. 353–378, 2016.

[96] S. Shalev-shwartz, C. Science, S. Ben-david, and C. Science, *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

[97] T. Goswami and U. B. Roy, "Predictive Model for Classification of Power System Faults using Machine Learning," in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, Oct. 2019, vol. 2019-Octob, pp. 1881–1885. doi: 10.1109/TENCON.2019.8929264.

[98] M. H. D. M. Ribeiro, R. G. da Silva, V. C. Mariani, and L. dos S. Coelho, "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil," *Chaos, Solitons and Fractals*, vol. 135, 2020, doi: 10.1016/j.chaos.2020.109853.

[99] S. R. Moreno, V. C. Mariani, and L. dos S. Coelho, "Hybrid multi-stage decomposition with parametric model applied to wind speed forecasting in Brazilian Northeast," *Renew. Energy*, vol. 164, pp. 1508–1526, 2021, doi: 10.1016/j.renene.2020.10.126.

[100] P. P. Shinde and S. Shah, "A Review of Machine Learning and Deep Learning Applications," *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–6, 2018, doi: 10.1109/ICCUBEA.2018.8697857.

[101] H. V. H. Ayala, D. Habineza, M. Rakotondrabe, and L. dos Santos Coelho, "Nonlinear black-box system identification through coevolutionary algorithms and

radial basis function artificial neural networks," *Appl. Soft Comput.*, vol. 87, p. 105990, Feb. 2020, doi: 10.1016/j.asoc.2019.105990.

[102] Y. Zhao, R. Ball, J. Mosesian, J.-F. de Palma, and B. Lehman, "Graph-Based Semi-supervised Learning for Fault Detection and Classification in Solar Photovoltaic Arrays," *IEEE Trans. Power Electron.*, vol. 30, no. 5, pp. 2848–2858, May 2015, doi: 10.1109/TPEL.2014.2364203.

[103] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, 2018, doi: 10.1155/2018/7068349.

[104] L. El Ghaoui, G. C. Li, V. A. Duong, V. Pham, A. Srivastava, and K. Bhaduri, "Sparse machine learning methods for understanding large text corpora," *Proc. 2011 Conf. Intell. Data Understanding, CIDU 2011*, pp. 159–173, 2011.

[105] W. Jia, R. M. Shukla, and S. Sengupta, "Anomaly Detection using Supervised Learning and Multiple Statistical Methods," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Dec. 2019, pp. 1291–1297. doi: 10.1109/ICMLA.2019.00211.

[106] J. Kober, "Robot Learning," in *Encyclopedia of Systems and Control*, vol. 28, no. 5, London: Springer London, 2020, pp. 1–9. doi: 10.1007/978-1-4471-5102-9_100027-1.

[107] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mech. Syst. Signal Process.*, vol. 138, p. 106587, Apr. 2020, doi: 10.1016/j.ymssp.2019.106587.

[108] Y. Roh, G. Heo, and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1328–1347, Apr. 2021, doi: 10.1109/TKDE.2019.2946162.

[109] Y. Guo, Z. Yang, S. Feng, and J. Hu, "Complex Power System Status Monitoring and Evaluation Using Big Data Platform and Machine Learning Algorithms: A Review and a Case Study," *Complexity*, vol. 2018, 2018, doi: 10.1155/2018/8496187.

[110] X. Xu and Z. Meng, "A hybrid transfer learning model for short-term electric load forecasting," *Electr. Eng.*, vol. 102, no. 3, pp. 1371–1381, 2020, doi: 10.1007/s00202-020-00930-x.

[111] A. Hooshmand and R. Sharma, "Energy predictive models with limited data using transfer learning," *e-Energy 2019 - Proc. 10th ACM Int. Conf. Futur. Energy Syst.*, pp. 12–16, 2019, doi: 10.1145/3307772.3328284.

[112] S.-M. Jung, S. Park, S.-W. Jung, and E. Hwang, "Monthly Electric Load Forecasting Using Transfer Learning for Smart Cities," *Sustainability*, vol. 12, no. 16, p. 6364, Aug. 2020, doi: 10.3390/su12166364.

[113] R. B. Diwate and A. Sahu, "Data Mining Techniques in Association Rule: A Review," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 227–229, 2014,

[Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.642.5098&rep=rep1&type=pdf

[114] N. G. Lo, J. M. Flaus, and O. Adrot, “Review of Machine Learning Approaches in Fault Diagnosis applied to IoT Systems,” *2019 Int. Conf. Control. Autom. Diagnosis, ICCAD 2019 - Proc.*, 2019, doi: 10.1109/ICCAD46983.2019.9037949.

[115] C. He, D. Ge, M. Yang, N. Yong, J. Wang, and J. Yu, “A data-driven adaptive fault diagnosis methodology for nuclear power systems based on NSGAII-CNN,” *Ann. Nucl. Energy*, vol. 159, p. 108326, Sep. 2021, doi: 10.1016/j.anucene.2021.108326.

[116] S. Khatir, D. Boutchicha, C. Le Thanh, H. Tran-Ngoc, T. N. Nguyen, and M. Abdel-Wahab, “Improved ANN technique combined with Jaya algorithm for crack identification in plates using XIGA and experimental analysis,” *Theor. Appl. Fract. Mech.*, vol. 107, no. February, p. 102554, Jun. 2020, doi: 10.1016/j.tafmec.2020.102554.

[117] D. H. Nguyen-Le, Q. B. Tao, V. H. Nguyen, M. Abdel-Wahab, and H. Nguyen-Xuan, “A data-driven approach based on long short-term memory and hidden Markov model for crack propagation prediction,” *Eng. Fract. Mech.*, vol. 235, no. May, p. 107085, 2020, doi: 10.1016/j.engfracmech.2020.107085.

[118] R. Zenzen, S. Khatir, I. Belaidi, C. Le Thanh, and M. Abdel Wahab, “A modified transmissibility indicator and Artificial Neural Network for damage identification and quantification in laminated composite structures,” *Compos. Struct.*, vol. 248, p. 112497, 2020, doi: 10.1016/j.compstruct.2020.112497.

[119] S. Wang *et al.*, “Automatic laser profile recognition and fast tracking for structured light measurement using deep learning and template matching,” *Measurement*, vol. 169, p. 108362, Feb. 2021, doi: 10.1016/j.measurement.2020.108362.

[120] A. A. Silva, A. M. Bazzi, and S. Gupta, “Fault diagnosis in electric drives using machine learning approaches,” in *2013 International Electric Machines & Drives Conference*, May 2013, pp. 722–726. doi: 10.1109/IEMDC.2013.6556173.

[121] S. Manikandan and K. Duraivelu, “Fault diagnosis of various rotating equipment using machine learning approaches – A review,” *Proc. Inst. Mech. Eng. Part E J. Process Mech. Eng.*, vol. 235, no. 2, pp. 629–642, Apr. 2021, doi: 10.1177/0954408920971976.

[122] S. Pang, X. Yang, X. Zhang, and X. Lin, “Fault diagnosis of rotating machinery with ensemble kernel extreme learning machine based on fused multi-domain features,” *ISA Trans.*, vol. 98, no. March, pp. 320–337, Mar. 2020, doi: 10.1016/j.isatra.2019.08.053.

[123] B. Li and Y.-P. Zhao, “Group reduced kernel extreme learning machine for fault diagnosis of aircraft engine,” *Eng. Appl. Artif. Intell.*, vol. 96, no. July, p. 103968, Nov. 2020, doi: 10.1016/j.engappai.2020.103968.

[124] S. F. Stefenon *et al.*, “Wavelet group method of data handling for fault prediction in

electrical power insulators," *Int. J. Electr. Power Energy Syst.*, vol. 123, no. March, p. 106269, Dec. 2020, doi: 10.1016/j.ijepes.2020.106269.

[125] S. F. Stefenon, R. B. Grebogi, R. Z. Freire, A. Nied, and L. H. Meyer, "Optimized Ensemble Extreme Learning Machine for Classification of Electrical Insulators Conditions," *IEEE Trans. Ind. Electron.*, vol. 67, no. 6, pp. 5170–5178, Jun. 2020, doi: 10.1109/TIE.2019.2926044.

[126] Saurabh Tewari and U.D. Dwivedi, "A Real-World Investigation of Twin SVM for the Classification of Petroleum Drilling Data," in *In Proceeding of 2019 IEEE Region 10 Symposium (TENSYMP)*, 2019, pp. 90–95.

[127] M. Rostami, K. Berahmand, E. Nasiri, and S. Forouzandeh, "Review of swarm intelligence-based feature selection methods," *Eng. Appl. Artif. Intell.*, vol. 100, no. April, p. 104210, Apr. 2021, doi: 10.1016/j.engappai.2021.104210.

[128] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, 2006, doi: 10.1007/s10462-007-9052-3.

[129] M. Jamil, S. K. Sharma, and R. Singh, "Fault detection and classification in electrical power transmission system using artificial neural network," *Springerplus*, vol. 4, no. 1, pp. 334–347, Dec. 2015, doi: 10.1186/s40064-015-1080-x.

[130] P. N. and P. B. M. Gowrishankar, "Transmission lines fault detection using discrete wavelet transform and artificial neural network," *Middle-East J. Sci. Res.*, vol. 24, no. 4, pp. 1112–1121, 2016, [Online]. Available: https://www.researchgate.net/publication/335619816_Transmission_Line_Fault_Detection_and_Classification_Using_Discrete_Wavelet_Transform_and_Artificial_Neural_Network

[131] S. K. Shukla, E. Koley, and S. Ghosh, "Protection Scheme for Shunt Faults in Six-Phase Transmission System Based on Wavelet Transform and Support Vector Machine," in *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Dec. 2017, pp. 1–5. doi: 10.1109/ICCIC.2017.8524340.

[132] Q. H. Alsafasfeh, I. Abdel-Qader, and A. M. Harb, "Fault Classification and Localization in Power Systems Using Fault Signatures and Principal Components Analysis," *Energy Power Eng.*, vol. 04, no. 06, pp. 506–522, 2012, doi: 10.4236/epe.2012.46064.

[133] M. Paluszek and S. Thomas, "MATLAB Machine Learning Toolboxes," in *Practical MATLAB Deep Learning*, Berkeley, CA: Apress, 2017. doi: 10.1007/978-1-4842-5124-9_2.

[134] M. He and J. Zhang, "A Dependency Graph Approach for Fault Detection and Localization Towards Secure Smart Grid," *IEEE Trans. Smart Grid*, vol. 2, no. 2, pp. 342–351, Jun. 2011, doi: 10.1109/TSG.2011.2129544.

[135] J. Doria-Garcia, C. Orozco-Henao, L. U. Iurinic, and J. D. Pulgarín-Rivera, "High

impedance fault location: Generalized extension for ground faults," *Int. J. Electr. Power Energy Syst.*, vol. 114, no. November 2018, p. 105387, 2020, doi: 10.1016/j.ijepes.2019.105387.

[136] Y. Xu and R. Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning," *J. Anal. Test.*, vol. 2, no. 3, pp. 249–262, 2018, doi: 10.1007/s41664-018-0068-2.

[137] A. Dasgupta, Y. V. Sun, I. R. König, J. E. Bailey-Wilson, and J. D. Malley, "Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience," *Genet. Epidemiol.*, vol. 35, no. S1, pp. S5–S11, 2011, doi: 10.1002/gepi.20642.

[138] A. Saxena *et al.*, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, no. July, pp. 664–681, 2017, doi: 10.1016/j.neucom.2017.06.053.

[139] A. Yadav and A. Swetapadma, "Fault analysis in three phase transmission lines using k-nearest neighbor algorithm," in *2014 International Conference on Advances in Electronics Computers and Communications*, Oct. 2014, pp. 1–5. doi: 10.1109/ICAECC.2014.7002474.

[140] L. Xu and M. Y. Chow, "Power distribution systems fault cause identification using logistic regression and artificial neural network," *Proc. 13th Int. Conf. Intell. Syst. Appl. to Power Syst. ISAP'05*, vol. 2005, pp. 163–168, 2005, doi: 10.1109/ISAP.2005.1599256.

[141] S. M. Srinivasan, T. Truong-Huu, and M. Gurusamy, "Machine Learning-Based Link Fault Identification and Localization in Complex Networks," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6556–6566, Aug. 2019, doi: 10.1109/JIOT.2019.2908019.

[142] Z. Elamrani Abou Elassad, H. Mousannif, H. Al Moatassime, and A. Karkouch, "The application of machine learning techniques for driving behavior analysis: A conceptual framework and a systematic literature review," *Eng. Appl. Artif. Intell.*, vol. 87, no. March 2019, p. 103312, 2020, doi: 10.1016/j.engappai.2019.103312.

[143] S. Tewari, U. D. Dwivedi, and S. Biswas, "A Novel Application of Ensemble Methods with Data Resampling Techniques for Drill Bit Selection in the Oil and Gas Industry," *Energies*, vol. 14, no. 2, p. 432, 2021, doi: 10.3390/en14020432.

[144] S. Tewari and U. D. Dwivedi, "A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies," *J. Pet. Explor. Prod. Technol.*, vol. 10, no. 5, pp. 1849–1868, 2020, doi: 10.1007/s13202-020-00839-y.

[145] S. Tewari and U. D. Dwivedi, "Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs," *Comput. Ind. Eng.*, vol. 128, pp. 937–947, Feb. 2019, doi: 10.1016/j.cie.2018.08.018.

[146] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean Absolute

Percentage Error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, Jun. 2016, doi: 10.1016/j.neucom.2015.12.114.

[147] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, Oct. 2006, doi: 10.1016/j.ijforecast.2006.03.001.

[148] R. Fezai *et al.*, "Effective Random Forest-Based Fault Detection and Diagnosis for Wind Energy Conversion Systems," *IEEE Sens. J.*, vol. 21, no. 5, pp. 6914–6921, 2021, doi: 10.1109/JSEN.2020.3037237.

[149] A. Mukherjee, P. K. Kundu, and A. Das, "Power system fault identification and localization using multiple linear regression of principal component distance indices," *Int. J. Appl. Power Eng.*, vol. 9, no. 2, p. 113, Aug. 2020, doi: 10.11591/ijape.v9.i2.pp113-126.

[150] H. Zhao, "Neural component analysis for fault detection," *Chemom. Intell. Lab. Syst.*, vol. 176, pp. 11–21, 2018, doi: 10.1016/j.chemolab.2018.02.001.

[151] M. Sarlak and S. M. Shahrtash, "High impedance fault detection in distribution networks using support vector machines based on wavelet transform," in *2008 IEEE Canada Electric Power Conference*, Oct. 2008, pp. 1–6. doi: 10.1109/EPC.2008.4763380.

[152] H. Gharavi and B. Hu, "Space-Time Approach for Disturbance Detection and Classification," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5132–5140, 2018, doi: 10.1109/TSG.2017.2680742.

[153] E. De Santis, A. Rizzi, and A. Sadeghian, "A cluster-based dissimilarity learning approach for localized fault classification in Smart Grids," *Swarm Evol. Comput.*, vol. 39, pp. 267–278, Apr. 2018, doi: 10.1016/j.swevo.2017.10.007.

[154] T. K. Barik and V. A. Centeno, "K-Medoids Clustering of Setting Groups in Directional Overcurrent Relays for Distribution System Protection," *2020 IEEE Kansas Power Energy Conf. KPEC 2020*, 2020, doi: 10.1109/KPEC47870.2020.9167531.

[155] Y. G. Zhang, Z. P. Wang, J. F. Zhang, and J. Ma, "Fault localization in electrical power systems: A pattern recognition approach," *Int. J. Electr. Power Energy Syst.*, vol. 33, no. 3, pp. 791–798, 2011, doi: 10.1016/j.ijepes.2011.01.018.

[156] H. Li, Y. Weng, E. Farantatos, and M. Patel, "An Unsupervised Learning Framework for Event Detection, Type Identification and Localization Using PMUs Without Any Historical Labels," *IEEE Power Energy Soc. Gen. Meet.*, vol. 2019-Augus, 2019, doi: 10.1109/PESGM40551.2019.8973580.

[157] J. Cordova, C. Soto, M. Gilanifar, Y. Zhou, A. Srivastava, and R. Arghandeh, "Shape Preserving Incremental Learning for Power Systems Fault Detection," *IEEE Control Syst. Lett.*, vol. 3, no. 1, pp. 85–90, Jan. 2019, doi: 10.1109/LCSYS.2018.2852064.

[158] M. Majidi, M. Etezadi-Amoli, and M. Sami Fadali, "A Novel Method for Single and

Simultaneous Fault Location in Distribution Networks," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3368–3376, Nov. 2015, doi: 10.1109/TPWRS.2014.2375816.

[159] Nan Zhang and M. Kezunovic, "Coordinating fuzzy ART neural networks to improve transmission line fault detection and classification," in *IEEE Power Engineering Society General Meeting, 2005*, 2005, vol. 1, pp. 1427–1433. doi: 10.1109/PES.2005.1489373.

[160] Y. Xu, Z. Y. Dong, K. Meng, R. Zhang, and K. P. Wong, "Real-time transient stability assessment model using extreme learning machine," *IET Gener. Transm. Distrib.*, vol. 5, no. 3, pp. 314–322, 2011, doi: 10.1049/iet-gtd.2010.0355.

[161] Y. Cai, S. Member, M. Chow, W. Lu, and L. Li, "Statistical Feature Selection From Massive Data in Distribution Fault Diagnosis," *IEEE Trans. Power Syst.*, vol. 25, no. 2, pp. 642–648, 2010.

[162] Le Xu, Mo-Yuen Chow, X. Z. Gao, L. Xu, and M. Y. Chow, "Comparisons of logistic regression and artificial neural network on power distribution systems fault cause identification," *Proc. 2005 IEEE Midnight-Summer Work. Soft Comput. Ind. Appl. 2005. SMCia/05.*, vol. 2005, pp. 128–131, 2005, doi: 10.1109/SMCIA.2005.1466960.

[163] Y. Cai and M.-Y. Chow, "Exploratory analysis of massive data for distribution fault diagnosis in smart grids," in *2009 IEEE Power & Energy Society General Meeting*, Jul. 2009, pp. 1–6. doi: 10.1109/PES.2009.5275689.

[164] M.-J. J. Chen, S. Lan, and D.-Y. Y. Chen, "Machine Learning Based One-Terminal Fault Areas Detection in HVDC Transmission System," *2018 8th Int. Conf. Power Energy Syst. ICPES 2018*, pp. 278–282, Dec. 2019, doi: 10.1109/ICPESYS.2018.8626976.

[165] A. Recioui, B. Benseghier, and H. Khalfallah, "Power system fault detection, classification and location using the K-Nearest Neighbors," in *2015 4th International Conference on Electrical Engineering (ICEE)*, Dec. 2015, pp. 1–6. doi: 10.1109/INTEE.2015.7416832.

[166] Linan Li, Renfei Che, and Hongzhi Zang, "A fault cause identification methodology for transmission lines based on support vector machines," in *2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, Oct. 2016, vol. 2016-Decem, pp. 1430–1434. doi: 10.1109/APPEEC.2016.7779725.

[167] H. Livani and C. Y. Evrenosoglu, "A fault classification method in power systems using DWT and SVM classifier," *Proc. IEEE Power Eng. Soc. Transm. Distrib. Conf.*, pp. 1–5, 2012, doi: 10.1109/TDC.2012.6281686.

[168] B. Sreewirote and A. Ngaopitakkul, "Classification of Fault Type on Loop-Configuration Transmission System Using Support Vector Machine," *Proc. - 2017 6th IIAI Int. Congr. Adv. Appl. Informatics, IIAI-AAI 2017*, pp. 892–896, 2017, doi: 10.1109/IIAI-AAI.2017.203.

[169] Zufeng Wang and Pu Zhao, "Fault location recognition in transmission lines based

on Support Vector Machines," in *2009 2nd IEEE International Conference on Computer Science and Information Technology*, 2009, pp. 401–404. doi: 10.1109/ICCSIT.2009.5234528.

[170] X. Deng, R. Yuan, Z. Xiao, T. Li, and K. L. L. Wang, "Fault location in loop distribution network using SVM technology," *Int. J. Electr. Power Energy Syst.*, vol. 65, pp. 254–261, 2015, doi: 10.1016/j.ijepes.2014.10.010.

[171] B. Bhalja and R. P. Maheshwari, "Wavelet-based Fault Classification Scheme for a Transmission Line Using a Support Vector Machine," *Electr. Power Components Syst.*, vol. 36, no. 10, pp. 1017–1030, Sep. 2008, doi: 10.1080/15325000802046496.

[172] X. G. Magagula, Y. Hamam, J. A. Jordaan, and A. A. Yusuff, "Fault detection and classification method using DWT and SVM in a power distribution network," in *2017 IEEE PES PowerAfrica*, Jun. 2017, pp. 1–6. doi: 10.1109/PowerAfrica.2017.7991190.

[173] V. Malathi and N. S. Marimuthu, "Multi-class support vector machine approach for fault classification in power transmission line," *2008 IEEE Int. Conf. Sustain. Energy Technol. ICSET 2008*, pp. 67–71, 2008, doi: 10.1109/ICSET.2008.4746974.

[174] O. A. S. Youssef, "An optimised fault classification technique based on support-vector- machines," *2009 IEEE/PES Power Syst. Conf. Expo. PSCE 2009*, 2009, doi: 10.1109/PSCE.2009.4839949.

[175] V. Malathi, N. S. Marimuthu, and S. Baskar, "Intelligent approaches using support vector machine and extreme learning machine for transmission line protection," *Neurocomputing*, vol. 73, no. 10–12, pp. 2160–2167, Jun. 2010, doi: 10.1016/j.neucom.2010.02.001.

[176] A. A. Yusuff, A. A. Jimoh, and J. L. Munda, "Determinant-based feature extraction for fault detection and classification for power transmission lines," *IET Gener. Transm. Distrib.*, vol. 5, no. 12, pp. 1259–1267, 2011, doi: 10.1049/iet-gtd.2011.0110.

[177] M. Singh, B. . Panigrahi, and R. P. Maheshwari, "Transmission line fault detection and classification," in *2011 International Conference on Emerging Trends in Electrical and Computer Technology*, Mar. 2011, pp. 15–22. doi: 10.1109/ICETECT.2011.5760084.

[178] H. Fathabadi, "Novel filter based ANN approach for short-circuit faults detection, classification and location in power transmission lines," *Int. J. Electr. Power Energy Syst.*, vol. 74, pp. 374–383, Jan. 2016, doi: 10.1016/j.ijepes.2015.08.005.

[179] B. Bhattacharya and A. Sinha, "Intelligent Fault Analysis in Electrical Power Grids," in *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, Nov. 2017, vol. 2017-Novem, pp. 985–990. doi: 10.1109/ICTAI.2017.00151.

[180] H. Livani and C. Y. Evrenosoglu, "A Machine Learning and Wavelet-Based Fault Location Method for Hybrid Transmission Lines," *IEEE Trans. Smart Grid*, vol. 5,

no. 1, pp. 51–59, Jan. 2014, doi: 10.1109/TSG.2013.2260421.

[181] H. Livani and C. Y. Evrenosoglu, “A fault classification and localization method for three-terminal circuits using machine learning,” *IEEE Trans. Power Deliv.*, vol. 28, no. 4, pp. 2282–2290, 2013, doi: 10.1109/TPWRD.2013.2272936.

[182] H. R. Baghaee, D. Mlakic, S. Nikolovski, and T. Dragicevic, “Support Vector Machine-Based Islanding and Grid Fault Detection in Active Distribution Networks,” *IEEE J. Emerg. Sel. Top. Power Electron.*, vol. 8, no. 3, pp. 2385–2403, Sep. 2020, doi: 10.1109/JESTPE.2019.2916621.

[183] N. Markovic, T. Stoetzel, V. Staudt, and D. Kolossa, “Hybrid fault detection in power systems,” *2019 IEEE Int. Electr. Mach. Drives Conf.*, pp. 911–915, May 2019, doi: 10.1109/IEMDC.2019.8785191.

[184] Y. Wang, X. Wang, Y. Wu, and Y. Guo, “Power System Fault Classification and Prediction Based on a Three-Layer Data Mining Structure,” *IEEE Access*, vol. 8, pp. 200897–200914, 2020, doi: 10.1109/ACCESS.2020.3034365.

[185] I. Nikoofekr, M. Sarlak, and S. M. Shahrtash, “Detection and classification of high impedance faults in power distribution networks using ART neural networks,” in *2013 21st Iranian Conference on Electrical Engineering (ICEE)*, May 2013, pp. 1–6. doi: 10.1109/IranianCEE.2013.6599760.

[186] A. Jain, A. Thoke, and R. Patel, “Fault classification of double circuit transmission line using artificial neural network,” *Int. J. Electr. ...*, vol. 2, no. 10, pp. 1029–1034, 2008, [Online]. Available: https://waset.org/journals/waset/v22/v22-149.pdf

[187] N. Roy and K. Bhattacharya, “Detection, Classification, and Estimation of Fault Location on an Overhead Transmission Line Using S-transform and Neural Network,” *Electr. Power Components Syst.*, vol. 43, no. 4, pp. 461–472, Feb. 2015, doi: 10.1080/15325008.2014.986776.

[188] D. Tzelepis *et al.*, “Advanced fault location in MTDC networks utilising optically-multiplexed current measurements and machine learning approach,” *Int. J. Electr. Power Energy Syst.*, vol. 97, no. October 2017, pp. 319–333, 2018, doi: 10.1016/j.ijepes.2017.10.040.

[189] P. Ray, B. K. Panigrahi, and N. Senroy, “Extreme learning machine based fault classification in a series compensated transmission line,” *PEDES 2012 - IEEE Int. Conf. Power Electron. Drives Energy Syst.*, pp. 1–6, 2012, doi: 10.1109/PEDES.2012.6484297.

[190] F. Unal and S. Ekici, “A fault location technique for HVDC transmission lines using extreme learning machines,” *ICSG 2017 - 5th Int. Istanbul Smart Grids Cities Congr. Fair*, pp. 125–129, 2017, doi: 10.1109/SGCF.2017.7947616.

[191] V. Malathi, N. S. Marimuthu, S. Baskar, and K. Ramar, “Application of extreme learning machine for series compensated transmission line protection,” *Eng. Appl. Artif. Intell.*, vol. 24, no. 5, pp. 880–887, 2011, doi: 10.1016/j.engappai.2011.03.003.

[192] S. R. Fahim, Y. Sarker, S. K. Sarker, M. R. I. Sheikh, and S. K. Das, “Self attention convolutional neural network with time series imaging based feature extraction for transmission line fault detection and classification,” *Electr. Power Syst. Res.*, vol. 187, no. February, p. 106437, 2020, doi: 10.1016/j.epsr.2020.106437.

[193] W. Li, D. Deka, M. Chertkov, and M. Wang, “Real-Time Faulted Line Localization and PMU Placement in Power Systems Through Convolutional Neural Networks,” *IEEE Trans. Power Syst.*, vol. 34, no. 6, pp. 4640–4651, Nov. 2019, doi: 10.1109/TPWRS.2019.2917794.

[194] P. Xi, P. Feilai, L. Yongchao, L. Zhiping, and L. Long, “Fault detection algorithm for power distribution network based on sparse selfencoding neural network,” *Proc. - 2017 Int. Conf. Smart Grid Electr. Autom. ICSGEA 2017*, vol. 2017-Janua, pp. 9–12, 2017, doi: 10.1109/ICSGEA.2017.19.

[195] M. Mahdi and V. M. I. Genc, “Post-fault prediction of transient instabilities using stacked sparse autoencoder,” *Electr. Power Syst. Res.*, vol. 164, no. August, pp. 243–252, Nov. 2018, doi: 10.1016/j.epsr.2018.08.009.

[196] A. Ajagekar and F. You, “Quantum computing based hybrid deep learning for fault diagnosis in electrical power systems,” *Appl. Energy*, vol. 303, p. 117628, Dec. 2021, doi: 10.1016/J.APENERGY.2021.117628.

[197] A. Yadav and A. Swetapadma, “Combined DWT and Naive Bayes based fault classifier for protection of double circuit transmission line,” in *International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014)*, May 2014, pp. 1–6. doi: 10.1109/ICRAIE.2014.6909179.

[198] S. Xiong, Y. Liu, J. Fang, J. Dai, L. Luo, and X. Jiang, “Incipient Fault Identification in Power Distribution Systems via Human-Level Concept Learning,” *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5239–5248, Nov. 2020, doi: 10.1109/TSG.2020.2994637.

[199] J. Upendar, C. P. Gupta, and G. K. Singh, “Statistical decision-tree based fault classification scheme for protection of power transmission lines,” *Int. J. Electr. Power Energy Syst.*, vol. 36, no. 1, pp. 1–12, Mar. 2012, doi: 10.1016/j.ijepes.2011.08.005.

[200] R. A. Sowah *et al.*, “Design of power distribution network fault data collector for fault detection, location and classification using machine learning,” *IEEE Int. Conf. Adapt. Sci. Technol. ICAST*, vol. 2018-Augus, pp. 1–8, 2018, doi: 10.1109/ICASTECH.2018.8506774.

[201] Z. Hajirahimi and M. Khashei, “Hybrid structures in time series modeling and forecasting: A review,” *Eng. Appl. Artif. Intell.*, vol. 86, no. July, pp. 83–106, Nov. 2019, doi: 10.1016/j.engappai.2019.08.018.

[202] S. Alshareef, S. Talwar, and W. G. Morsi, “A New Approach Based on Wavelet Design and Machine Learning for Islanding Detection of Distributed Generation,” *IEEE Trans. Smart Grid*, vol. 5, no. 4, pp. 1575–1583, Jul. 2014, doi: 10.1109/TSG.2013.2296598.

[203] T. Guo, P. Papadopoulos, P. Mohammed, and J. V. Milanovic, “Comparison of ensemble decision tree methods for on-line identification of power system dynamic signature considering availability of PMU measurements,” *2015 IEEE Eindhoven PowerTech, PowerTech 2015*, 2015, doi: 10.1109/PTC.2015.7232364.

[204] D. Patil, O. Naidu, P. Yalla, and S. Hida, “An Ensemble Machine Learning Based Fault Classification Method for Faults During Power Swing,” in *2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)*, May 2019, pp. 4225–4230. doi: 10.1109/ISGT-Asia.2019.8881359.

[205] C. D. Sutton, “Classification and Regression Trees, Bagging, and Boosting,” *Handb. Stat.*, vol. 24, no. 04, pp. 303–329, 2005, doi: 10.1016/S0169-7161(04)24011-1.

[206] K. Fawagreh, M. M. Gaber, and E. Elyan, “Random forests: From early developments to recent advancements,” *Syst. Sci. Control Eng.*, vol. 2, no. 1, pp. 602–609, 2014, doi: 10.1080/21642583.2014.956265.

[207] I. K. Nti, A. F. Adekoya, and B. A. Weyori, “A comprehensive evaluation of ensemble learning for stock-market prediction,” *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00299-5.

[208] S. R. Samantaray, “Ensemble decision trees for high impedance fault detection in power distribution network,” *Int. J. Electr. Power Energy Syst.*, vol. 43, no. 1, pp. 1048–1055, Dec. 2012, doi: 10.1016/j.ijepes.2012.06.006.

[209] Z. Qi, T. Yingjie, S. Yun, Q. Haini, and G. Naiwang, “Ungrounded Fault Detection in Medium Voltage Distribution Network Based on Machine Learning,” *2nd IEEE Conf. Energy Internet Energy Syst. Integr. EI2 2018 - Proc.*, pp. 1–5, 2018, doi: 10.1109/EI2.2018.8582360.

[210] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[211] X. Zeng, M.-F. Guo, and D. Chen, “Machine-learning-based single-phase-to-ground fault detection in distribution systems,” in *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)*, Nov. 2017, pp. 1–6. doi: 10.1109/EI2.2017.8245233.

[212] M. Chen, Q. Liu, S. Chen, Y. Liu, C. H. Zhang, and R. Liu, “XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System,” *IEEE Access*, vol. 7, pp. 13149–13158, 2019, doi: 10.1109/ACCESS.2019.2893448.

[213] A. Ibrahem Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie, “Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia,” *Ain Shams Eng. J.*, vol. 12, no. 2, pp. 1545–1556, 2021, doi: 10.1016/j.asej.2020.11.011.

[214] K. Moloi and A. O. Akumu, “Power distribution fault diagnostic method based on machine learning technique,” in *2019 IEEE PES/IAS PowerAfrica*, Aug. 2019, pp. 238–242. doi: 10.1109/PowerAfrica.2019.8928633.

[215] C. Fei, G. Qi, and C. Li, “Fault location on high voltage transmission line by applying support vector regression with fault signal amplitudes,” *Electr. Power Syst. Res.*, vol. 160, pp. 173–179, 2018, doi: 10.1016/j.epsr.2018.02.005.

[216] Q. Shi, M. Abdel-Aty, and J. Lee, “A Bayesian ridge regression analysis of congestion’s impact on urban expressway safety,” *Accid. Anal. Prev.*, vol. 88, pp. 124–137, 2016, doi: 10.1016/j.aap.2015.12.001.

[217] F. A. da Silva *et al.*, “Bayesian ridge regression shows the best fit for SSR markers in Psidium guajava among Bayesian models,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1038/s41598-021-93120-z.

[218] C. Ren and Y. Xu, “Transfer Learning-Based Power System Online Dynamic Security Assessment: Using One Model to Assess Many Unlearned Faults,” *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 821–824, Jan. 2020, doi: 10.1109/TPWRS.2019.2947781.

[219] H.-H. Chang, T.-C. Lai, H.-J. Chang, and W.-J. Lee, “Fault Location Identifications in HV Transmission Networks and Different MV Wind Farms Using Nonintrusive Monitoring Techniques,” *IEEE Trans. Ind. Appl.*, vol. 58, no. 2, pp. 1822–1830, Mar. 2022, doi: 10.1109/TIA.2022.3146107.

[220] H. Teimourzadeh, A. Moradzadeh, M. Shoaran, B. Mohammadi-Ivatloo, and R. Razzaghi, “High Impedance Single-Phase Faults Diagnosis in Transmission Lines via Deep Reinforcement Learning of Transfer Functions,” *IEEE Access*, vol. 9, pp. 15796–15809, 2021, doi: 10.1109/ACCESS.2021.3051411.

[221] J. G. Slootweg, “Wind Power Modelling and Impact on Power System Dynamics,” Delft University of Technology, 2003. [Online]. Available: http://resolver.tudelft.nl/uuid:f1ce3eaa-f57d-4d37-b739-b109599a7d21

[222] R. Orenge, M. Christopher Maina, and G. N. Nyakoe, “Optimal Sizing and Placement of Solar Photovoltaic Based DGs in the IEEE 9 Bus System Using Particle Swarm Optimization Algorithm,” in *2018 IEEE PES/IAS PowerAfrica*, Jun. 2018, pp. 1–6. doi: 10.1109/PowerAfrica.2018.8521006.

[223] T. I. Rahman, “Impact of Solar Generation on IEEE 9-bus System,” University of Maine, 2023. doi: https://digitalcommons.library.umaine.edu/etd/3818.

[224] D. L. Popa, M. S. Nicolae, P. M. Nicolae, and M. Popescu, “Design and simulation of a 10 MW photovoltaic power plant using MATLAB and Simulink,” *Proc. - 2016 IEEE Int. Power Electron. Motion Control Conf. PEMC 2016*, pp. 378–383, 2016, doi: 10.1109/EPEPEMC.2016.7752027.

[225] M. K. Behera, I. Majumder, and N. Nayak, “Solar photovoltaic power forecasting using optimized modified extreme learning machine technique,” *Eng. Sci. Technol. an Int. J.*, vol. 21, no. 3, pp. 428–438, Jun. 2018, doi: 10.1016/j.jestch.2018.04.013.

[226] M. Pazoki, “A New DC-Offset Removal Method for Distance-Relaying Application Using Intrinsic Time-Scale Decomposition,” *IEEE Trans. Power Deliv.*, vol. 33, no. 2, pp. 971–980, Apr. 2018, doi: 10.1109/TPWRD.2017.2728188.

[227] H. Karbouj, Z. H. Rather, and B. C. Pal, “Adaptive Voltage Control for Large Scale Solar PV Power Plant Considering Real Life Factors,” *IEEE Trans. Sustain. Energy*, vol. 12, no. 2, pp. 990–998, Apr. 2021, doi: 10.1109/TSTE.2020.3029102.

[228] M. Irwanto, I. Daut, M. Sembiring, R. Bin Ali, S. Champakeow, and S. Shema, “Effect of Solar Irradiance and Temperature on Photovoltaic Module Electrical Characteristics,” no. October, pp. 16–17, 2010.

[229] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, 2011. https://scikit-learn.org/stable/modules/classes.html

[230] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends,” *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/J.NEUCOM.2019.10.118.

[231] S. Duman, H. T. Kahraman, and M. Kati, “Economical operation of modern power grids incorporating uncertainties of renewable energy sources and load demand using the adaptive fitness-distance balance-based stochastic fractal search algorithm,” *Eng. Appl. Artif. Intell.*, vol. 117, p. 105501, Jan. 2023, doi: 10.1016/j.engappai.2022.105501.

[232] R. Aljarrah, H. Marzooghi, and V. Terzija, “Mitigating the impact of fault level shortfall in future power systems with high penetration of converter-interfaced renewable energy sources,” *Int. J. Electr. Power Energy Syst.*, vol. 149, no. July, p. 109058, Jul. 2023, doi: 10.1016/j.ijepes.2023.109058.

[233] H. Hassani, E. Hallaji, R. Razavi-Far, and M. Saif, “Unsupervised concrete feature selection based on mutual information for diagnosing faults and cyber-attacks in power systems,” *Eng. Appl. Artif. Intell.*, vol. 100, p. 104150, Apr. 2021, doi: 10.1016/J.ENGAPPAI.2020.104150.

[234] Y. M. Nsaif, M. S. H. Lipu, A. Ayob, Y. Yusof, and A. Hussain, “Fault Detection and Protection Schemes for Distributed Generation Integrated to Distribution Network: Challenges and Suggestions,” *IEEE Access*, vol. 9, pp. 142693–142717, 2021, doi: 10.1109/ACCESS.2021.3121087.

[235] A. C. Malti Bansal, Apoorva Goyal, “A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning,” *Decis. Anal. J.*, vol. 3, 2022, doi: https://doi.org/10.1016/j.dajour.2022.100071.

[236] S. Pongswatd and K. Smerpitak, “Applying radar chart for process control behavior,” *2018 3rd Int. Conf. Control Robot. Eng. ICCRE 2018*, pp. 90–93, 2018, doi: 10.1109/ICCRE.2018.8376440.

[237] S. Yin and M. G. Alfarizi, “Intelligent Fault Diagnosis of Manufacturing Processes Using Extra Tree Classification,” *IEEE Open J. Ind. Electron. Soc.*, vol. 4, no. December, pp. 618–628, 2023, doi: https://doi.org/10.1109/OJIES.2023.3334429.

[238] C. Yao, L. Edu, and D. C. Maceren, “extreme gradient boosting,” no. December

2023, 2024, doi: 10.1049/esi2.12144.

[239] J. Montiel, R. Mitchell, E. Frank, B. Pfahringer, T. Abdessalem, and A. Bifet, "Adaptive XGBoost for Evolving Data Streams," no. 1, 2020.

[240] Y. Yang, "Hybrid Prediction Method for Wind Speed Combining Ensemble Empirical Mode Decomposition and Bayesian Ridge Regression," *IEEE Access*, vol. 8, pp. 71206–71218, 2020, doi: 10.1109/ACCESS.2020.2984020.

[241] V. Le and X. Yao, "Ensemble machine learning based adaptive arc fault detection for DC distribution systems," *Conf. Proc. - IEEE Appl. Power Electron. Conf. Expo. - APEC*, vol. 2019-March, pp. 1984–1989, 2019, doi: 10.1109/APEC.2019.8721922.

[242] S. Saha, M. I. Saleem, and T. K. Roy, "Impact of high penetration of renewable energy sources on grid frequency behaviour," *Int. J. Electr. Power Energy Syst.*, vol. 145, no. February, p. 108701, Feb. 2023, doi: 10.1016/j.ijepes.2022.108701.

[243] H. S. Jang, K. Y. Bae, H.-S. Park, and D. K. Sung, "Solar Power Prediction Based on Satellite Images and Support Vector Machine," *IEEE Trans. Sustain. Energy*, vol. 7, no. 3, pp. 1255–1263, Jul. 2016, doi: 10.1109/TSTE.2016.2535466.

[244] S. Impram, S. Varbak Nese, and B. Oral, "Challenges of renewable energy penetration on power system flexibility: A survey," *Energy Strateg. Rev.*, vol. 31, no. August, p. 100539, 2020, doi: 10.1016/j.esr.2020.100539.

[245] B. Tamimi, C. Canizares, and K. Bhattacharya, "System Stability Impact of Large-Scale and Distributed Solar Photovoltaic Generation: The Case of Ontario, Canada," *IEEE Trans. Sustain. Energy*, vol. 4, no. 3, pp. 680–688, Jul. 2013, doi: 10.1109/TSTE.2012.2235151.

[246] S. Eftekharnejad, V. Vittal, Heydt, B. Keel, and J. Loehr, "Impact of increased penetration of photovoltaic generation on power systems," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 893–901, May 2013, doi: 10.1109/TPWRS.2012.2216294.

[247] G. M. Ven, T. Tuytelaars, and A. S. Tolias, "Three types of incremental learning," vol. 4, no. October 2021, 2022, doi: 10.1038/s42256-022-00568-3.

[248] Haibo He, Sheng Chen, Kang Li, and Xin Xu, "Incremental Learning From Stream Data," *IEEE Trans. Neural Networks*, vol. 22, no. 12, pp. 1901–1914, Dec. 2011, doi: 10.1109/TNN.2011.2171713.

LIST OF PUBLICATIONS

Journal

1. **Rachna Vaish**, U.D. Dwivedi, S. Tewari and S.M. Tripathi, "Machine Learning Applications in Power System Fault Diagnosis: Research Advancements and Perspectives", *Elsevier – Engineering Applications of Artificial Intelligence.* Vol. 106, November (2021).

2. **Rachna Vaish** and U.D. Dwivedi, "Performance and Adaptability Testing of Machine Learning Models for Transmission Network Fault Diagnosis with Renewable Energy Sources Integration", *IEEE Access – Paper Accepted.*

3. **Rachna Vaish**, and U.D. Dwivedi, "Fault Diagnosis in Evolving Power Systems: Incremental Learning of ML Models under Increasing Penetration of Renewable Energy Sources", (2024). (Manuscript submitted).

Book Chapter

1. **Rachna Vaish**, and U.D. Dwivedi, "Role of Machine Learning in Forecasting Solar and Wind Power Generation.", *Nova Science Publishers – Energy Conversion: Methods, Technology and Future Directions*, December 2022.

Conferences

1. **Rachna Vaish**, and U.D. Dwivedi, "Comparative Study of Machine Learning Models for Power System Fault Identification and Localization", 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), Jharkhand, India, 11-12 February 2022.

2. **Rachna Vaish**, and U.D. Dwivedi, "Bayesian Ridge Regression for Power System Fault Localization", 2024 5th International Conference for Emerging Technology (INCET 2024), Belgaum, Karnataka, India 24-26 May 2024.

3. **Rachna Vaish**, U.D. Dwivedi, and B. Chikondra, "Challenges Posed by Renewable Energy Source Integration to Machine Learning based Power System Fault Diagnosis", 51st IEEE International Communications Energy Conference (INTELEC 2024) Bengaluru, India, 4-7 August 2024. (Paper Accepted).

4. **Rachna Vaish**, B. Chikondra and U.D. Dwivedi, "Multi-switch Fault Diagnosis for the Voltage Source Inverter fed Multi-phase Drives based on Machine Learning", 51st IEEE International Communications Energy Conference (INTELEC 2024) Bengaluru, India, 4-7 August 2024. (Paper Accepted).

CURRICULUM VITAE Rachna Vaish

Project Scientist, JTRC, IIT Kanpur

✉: pee19001@rgipt.ac.in, rachnavaish91@gmail.com
✆: +91-8840329688, +91-9305315441
📍 : Prem Chandra Kesharwani, Town Area, Ward No. 9, Indira
Nagar, Sonaran Gali, Ajhuwa, Uttar Pradesh, India, Pin-212217

WORK EXPERIENCE

Project Scientist **(October 2024-To date)**
Just Transition Research Centre,
Indian Institute of Technology, Kanpur
Principal Investigator: Dr. Pradip Swarnakar

ACADEMIC OUTLINE

PhD in Electrical Engineering, 9.4 CGPA **(July 2019-To date)**
Department of Electrical and Electronics Engineering,
Rajiv Gandhi Institute of Petroleum Technology, Jais, Uttar Pradesh, India.
Thesis topic: Performance Assessment of Machine Learning Models for Transmission Network Fault Diagnosis Under Renewable Energy Source Integrations.
Thesis supervisor: Dr. Umakant Dhar Dwivedi

Master of Technology in Power Electronics and drives, 84.30% **2015-2018**
Department of Electrical Engineering,
Kamla Nehru Institute of Technology, Sultanpur, Uttar Pradesh, India.
Thesis topic: Design of PI Regulator Parameters with Damping Ratio Specification for CSI fed SCIM Drive using D-Partition Technique.
Thesis supervisor: Dr. Saurabh Mani Tripathi

Bachelor of Technology in Electrical Engineering, 79.58% **2010-2014**
Department of Electrical Engineering,
Shambhunath Institute of Engineering & Technology, Prayagraj, U.P., India.

Intermediate with PCM, 79% **2007-2009**
Girl’s High School & College,
Prayagraj, Uttar Pradesh, India.

High School, 81.30% **2005-2007**

Girl’s High School & College,
Prayagraj, Uttar Pradesh, India.

TEACHING and LABORATORY EXPERIENCE

- Teaching Assistant- to Dr. Umakant Dhar Dwivedi (Undergraduate Level: Electrical Workshop, Fundamental of Electronics Engineering, Microprocessor)
- Teaching Assistant- to Dr. Bheemaiah Chikondra (Undergraduate Level: Fundamental of Power Electronics, Measurement Lab)
- Teaching Assistant- to Dr. Amit Ranjan (Undergraduate Level: MATLAB)
- Teaching Assistant- to Dr. Saurabh Mani Tripathi, (Undergraduate Level: Fundamentals of Electrical Engineering)

FELLOWSHIPS AND AWARDS

- Qualified GATE (2014, 2015 and 2016).
- Research Teaching Assistantship, JRF RGIPT, Amethi 2019-2021.
- Research Teaching Assistantship, SRF RGIPT, Amethi 2021-2023.
- MHRD Scholarship, MTech, KNIT, 2015-2017.
- Third Prize in BTech for ranking third in UPTU semester exams within campus.

PERSONAL INFORMATION

Parent Name	Mr. Rakesh Vaish & Late Smt. Saroj Vaish
Nationality	Indian
Date of Birth	26th May 1991
Gender	Female
Marital Status	Married
Husband Name	Mr. Chitranshu Kesharwani
Languages	English, Hindi,

DECLARATION

I hereby declare that all the information mentioned above is true and correct to the best of my knowledge.

Thank you,

Rachna Vaish

Place: Jais, Amethi **Rachna Vaish**